Surveillance Graphs

Jonny L. Saunders

April 14, 2023

Contents

1 Introduction

The world is Big Data, and The Cloud is its landlord. Much like the universal commodification of prior [1] assemblages of capitalism, The Cloud's informational capitalism recasts every element of our reality as Data, and it is our responsibility to ferret it out of its primitive unknown, mine it, harvest it, dump it by the tanker-truckful into great Data Lakes and wring the Actionable Insights from its neck. The world has always been Big with Data, but The Cloud has finally given us a fruit of knowledge to munch on that reveals its true nature. We are but Data Subjects, and if we can harness the wily spray of our Organic Content with biosensor wearables, filtering our every action, affection, and affilitation through a mob of algorithmically optimized platforms, then The Cloud might teach us enough about ourselves to be able to lead a fulfilled and healthy life. Academics and Governments in particular are filthy with Data, and have set their eyes on a new generation of data infrastructures that promise to dissolve the Silos that prevent the Big Data from teaching us the nature of the universe, how to solve poverty, reverse climate change, and finally collapse into one great cat-pile of peace and love for our neighbor. The Cloud is dreaming of linking the Big Data into one great Knowledge Graph of Everything — it tells us this is important for the fate of humanity.

The Knowledge Graph of Everything is a mirage, though. Its vision of extracting all the world's data is the same colonial vision of infinite prosperity that has driven us to the brink of extinction: So The Cloud Said, let us make Platforms in our image, and let them have dominion over the interactions with the apps, and the insights of the analytics, and over all the data of the earth. The pursuit of universalizing ontologies that will finally give every quantum of Data its one True and Correct form is the same fascistic vision that has driven eliminationist campaigns to stamp out degeneracy: Ontologien über alles. Aside from its properly pathological nature, the Knowledge Graph of Everything *is impossible* and *won't work*. Instead, The Cloud will lead governments and academics along by the nose just long enough to build critical mass for an interlocking set of platforms that slice off snapshots of the Everything to ratchet us ever futher into the captivity of surveillance and subscription.

What is the alternative? The Cloud presents its future as inevitable, but by seeing past its logic we might imagine properly *human* infrastructures that fill the needs for connection and understanding that it exploits. This piece develops the notion of **vulgar linked data** as an alternative to the Cloud Orthodoxy. Predicated on relationality, heterogeneity, privacy, and vernacular expression, vulgar linked data infrastructures attempt to empower *people* to *socially organize* information in a truly decentralized sociotechnological commons, rather than empowering *systems* to *rent* knowledge organization for *profit*.

2 Knowledge Graphs: A Backbone in the Surveillance Economy

Through their cloud of corporate jargon, knowledge graphs are relatively straightforward to define [2, 3, 4, 5] (though see [6]): **directed, labeled graphs** consisting of *nodes* corresponding to entities like a person, dataset, location, etc. and *edges* that describe their relationship[1]. Knowledge graphs typically make use of some controlled **ontology** that provide a specific set of terms and how they are to be used, and "types" that give a given entity an expected set of *properties* denoted by edges with a particular set of labels from the ontology. For example, the schema.org Person type would be applied to a node, and then have a set of labeled edges like `gender` or `email` that link to other nodes that contain the values of the properties.

[1] Equivalently, one could emphasize that they are graphs composed of **triplet** links that describe some subject, predicate, and object.

Why does such a seemingly ordinary data structure deserve particular attention in an always-more-fraught landscape of digital technology? The story of knowledge graphs is the story of the enclosure of the wild and open web into a series

of surveillance-backed platforms. They provide an underexplored lens onto the present and future of digital infrastructure as planned by information conglomerates — and serve as a liberatory kernel that hints at how we might chart a different course.

2.0.1 Semantic Web: Priesthoods

The term "Knowledge Graph" evolved out of the Semantic Web project [3] . It is difficult to reconstruct how radical the notion of a collection of documents organized by arbitrary links between them was at dawn of the internet. At the time, the infrastructures of linking documents looked more like ISBNs, carefully regulated by expert, centralized authorities[2]. Being able to *just make anything that could be linked to* and *link to anything you wanted* was *terrifying* and *new* (eg. [8, 9]).

[2] For another example re: the political nature of the DOI system in the face of the arbitrary linking of the internet, see section 3.1.2 "Integration, not Invention" in [7]

The initial design of the web imagined it as a self-organizing process, where people would maintain their own websites and organize a collection of links to other websites. It became clear relatively quickly that the anarchy of a socially self-organizing internet wasn't going to work as planned, where without a formal system of organization "people were frightened of getting lost in it. You could follow links forever." [10]

In its earliest formulations, the Semantic Web was an attempt to supplement the same arbitrary power to express human-readable information with computer-readable information. It imagined a linked and overlapping set of schemas ranging from locally expressive vocabularies used among small groups of friends through globally shared, logically consistent ontologies. The semantic web was intended to evolve fluidly, like language, with cultures of meaning meshing and separating at multiple scales [11, 12, 13] :

> Locally defined languages are easy to create, needing local consensus about meaning: only a limited number of people have to share a mental pattern of relationships which define the meaning. However, global languages are so much more effective at communication, reaching the parts that local languages cannot. [...]
>
> So the idea is that in any one message, some of the terms will be from a global ontology, some from subdomains. The amount of data which can be reused by another agent will depend on how many communities they have in common, how many ontologies they share.
>
> In other words, one global ontology is not a solution to the problem, and a local subdomain is not a solution either. But if each agent has uses a mix of a few ontologies of different scale, that is forms a global solution to the problem. [11]
>
> The Semantic Web, in naming every concept simply by a URI, lets anyone express new concepts that they invent with minimal effort. Its unifying logical language will enable these concepts to be progressively linked into a universal Web. [12]

This freeform goal expression for expression's sake was always in tension with another part of the vision - serving as a backbone for AI "agents" that could compute emergent function from the semantic web. Succinctly: "Human language thrives when using the same term to mean somewhat different things, but automation does not." [12] This tension persists through the broader history of the web.

2.0.2 Linked Data: Platforms

Much of the work of the semantic web project in the early 2000s focused on the "global" side of this tension at the expense of the "local" - creating ontologies and related technologies intended to serve as a foundation for expressing basic things in a common vocabulary [3] . This work had many successes, but began a schism

between the priesthood of people concerned with making systems that were *correct* and those that were more concerned with making things that *worked* - or supported "local" expression (eg [14]). Aaron Swartz captured this frustration in his unfinished book:

> Instead of the "let's just build something that works" attitude that made the Web (and the Internet) such a roaring success, they brought the formalizing mindset of mathematicians and the institutional structures of academics and defense contractors. They formed committees to form working groups to write drafts of ontologies that carefully listed (in 100-page Word documents) all possible things in the universe and the various properties they could have, and they spent hours in Talmudic debates over whether a washing machine was a kitchen appliance or a household cleaning device. [15]

Lindsay Poirier describes this difference in "thought styles" as a rift between the "neats" focused on universalizing *a priori* ontologies and the "scruffies" focused on everyday use and letting the structure appear afterwards [16] . The latter characterizes the "second age" of the Semantic Web after 2006 - the reorganization around **Linked Data** [17, 3] . The era of Linked Data de-emphasized the idealistic and ideological goals of the early Semantic Web, driven more by an empirical approach of trying to realize these systems on the wilds of the web, creating some of the first public "Linked Open Data" systems like DBPedia and Freebase.

This turn coincides with the emerging platformatization and enclosure of the web as "Web 2.0." Throughout the early 2000s, the work of the Semantic Web project was largely invisible to the ordinary web user, and its vision of a self-organizing web was easily outcompeted by the now-ubiquitous use of search engines to index the web. Where in the early 2000s web architects were imagining the future of web continuing to take place on free and open *protocols,* the Linked Data/Web 2.0 era corraled us into a pattern of *platforms* which quickly ratcheted their way to dominance in a positive feedback loop of user experience design, network effects and profit. On platforms, rather than a system that "belongs" to everyone, you are granted access to some specific set of operations through an interface so that you can be part of a social process of producing and curating information for the platform holder.

2.0.3 Knowledge Graphs: Panoptica

In 2010 Google acquired Metaweb and its publicly-edited Semantic Web database Freebase, and in 2012 repackaged it and the ideas of Linked Data as what it called a **Knowledge Graph** — the third era of the Semantic Web [18, 19] . Freebase only made up part of it, and the full extent of Google's Knowledge Graph are unknown, but its most visible impact are the factboxes that present structured information about the subjects of searches - like biographical information in a search for a person - or the different widgets for contextual interaction - like being able to make a restaurant reservation from the serach page [20] . Knowledge Graphs still share the same underlying structure — triplet graphs with ontologies — even if they occupy a broader space of implementations and technologies. What differs is the context and intended use: the "worldview" of the knowledge graph.

Beyond the obvious product-level features it supports, Google's acquisition of Freebase and the structure of its Knowledge Graph represent at least two deeper shifts in the trajectory of the Semantic Web and the broader internet: the privatization of technologies with initially liberatory aspirations, and an early template of the sprawling, surveillance-driven information conglomerate we know and love today.

Like the radical nature of linking on the web, it's difficult to remember that the web as surveillance apparatus thinly veiled as the five or so remaining websites was not inevitable. The pre-dotcom bust internet of the 90's and early 2000's was far from the commercialized wasteland we know today. Ed Horowitz, CEO of Viacom

explained in 1996: “The Internet has yet to fulfill its promise of commercial success. Why? Because there is no business model” [21] . Google’s AdWords being a defining moment in the development of surveillance capitalism is a story already told [22] : taking advantage of the need for search generated by the disorganization of the web, AdWords turned personal search data into a profit vector by selling targeted space in the results.

The significance of the relationship between search, the semantic web, and what became knowledge graphs is less widely appreciated. The semantic web was initially an alternative to monolithic search engine platforms - or, more generally, to platforms in general [23] . It imagined the use of triplet links and shared ontologies at a protocol level as a way of organizing the information on the web into a richly explorable space: rather than needing to rely on a search bar, one could traverse a structured graph of information [17, 24] to find what one needed without mediation by a third party.

Instead, the form of of the semantic web that emerged as “Knowledge Graphs” flipped the vision of a free and evolving internet on its head. The mutation from “Linked Open Data” [17] to “Knowledge Graphs” is a shift in meaning from a public and densely linked web of information from many sources to a proprietary information store used to power derivative platforms and services. The shift isn’t quite so simple as a “closure” of a formerly open resource — we’ll return to the complex role of openness in a moment. It is closer to an *en*closure, a *domestication* of the dream of the Semantic Web. A dream of a mutating, pluralistic space of communication, where we were able to own and change and create the information that structures our digital lives was reduced to a ring of platforms that give us precisely as much agency as is needed to keep us content in our captivity. Links that had all the expressive power of utterances, questions, hints, slander, and lies were reduced to mere facts. We were recast from our role as *people* creating a digital world to *consumers* of subscriptions and services. The artifacts that we create for and with and between each other as the substance of our lives online were yoked to the acquisitive gaze of the knowledge graph as *content* to be mined. We vulgar commoners, we data subjects, are not allowed to touch the graph — even if it is built from our disembodied bits.

The same technologies, with minor variation, that were intended to keep the internet free became emblematic of and coproductive with the surveillance/platform model that has enclosed it. Beyond Google, knowledge graphs are an elemental part of the contempory information economy. Banks, militaries, governments, life science corporations, journalists, everyone is using knowledge graphs [25, 26] . Their ubiquity is not an accident, one of many possible data systems that could have fit the bill, but reflects and reinforces basic patterns of the information economy and the corporations within it.

What makes knowledge graphs so special? It turns out that semantic web technologies, designed to accomodate the infinitely heterogeneous, multiscale nature of free and unmediated social structuring of information are also quite useful for the indefinitely expanding dragnet of data collection that defines the operation of contemporary capitalism:

> “If one takes a look at the top Fortune 500 companies, it is surprising how many of them are really in the information business. I don’t just mean the technology and telecommunication companies like Apple or Google or Verizon or Cisco or the drug companies like Pfizer. One could also think of the big banks as a subset of the vectoralist class rather than as “finance capital.” They too are in the information asymmetry business. And as we learned in the 2008 crash, even the car companies are in the information business—they made more money from car loans than cars. The military—industrial sector is also in the information business. The companies that appear to sell actual things, like Nike, are really in the brand business. Walmart and Amazon compete

> with different models of the information logistics business. Even the oil companies are in part at least in the information-about-the-geology-of-possible-oil-deposits business. Perhaps the vectoralist class is no longer emerging. Maybe it is the new dominant class." - *McKenzie Wark, Capital Is Dead: Is This Something Worse?* [1]

Data companies — most major companies — need to store and maintain massive collections of heterogeneous data across their byzantine hierarchies of executives, managers, and workers. This gigantic haunted ball of data is not just a tool, but the *substance* of the company. A data company persists by exploiting the combinatorics of its data hoard, spinning off new platforms that in turn maintain and expand access to data by creating captive data subjects[3]. As it expands, a conglomerate will acquire many new sources and modalities of data and need to integrate them with its existing data.

Knowledge graphs are particularly well suited for this "data integration" problem. A full technical description is out of scope here, but briefly: traditional relational database systems can be very difficult to modify and refactor, and that difficulty increases the larger and more complex a database is[4]. One has to be design the structure of the anticipated data in advance, and the abstract schematic structure of the data is embedded in how it is stored and accessed. It is particularly difficult to do unanticipated "long range" analyses where very different kinds of data are analyzed together.

In contrast, merging graphs is more straightforward[5] [2, 26, 27, 28, 29, 30, 31, 32] - the data is just triplets, so in an idealized case[6] it is possible to just concatenate them and remove duplicates (eg. for a short example, see [33, 34]). The graph can be operated on locally, with more global coordination provided by ontologies and schemas, which themselves have a graph structure [35] . Discrepancies between graphlike schema can be resolved by, you guessed it, making more graph to describe the links and transformations between them. Long-range operations between data are part of the basic structure of a graph - just traverse nodes and edges until you get to where you need to go - and the semantic structure of the graph provides additional constraints to that traversal. Again, a technical description is out of scope here, graphs are not magic, but they are well-suited to merging, modifying, and analyzing large quantities of heterogeneous data.

Another way of looking at the capacity for heterogeneity in triplet graphs is by thinking of links as statements:

> One person may define a `vehicle` as having a `number of wheels` and a `weight` and a `length`, but not foresee a `color`. This will not stop another person making the assertion that a given car is `red`, using the color vocabulary from elsewhere. [10]

So if you are a data broker, and you just made a hostile acquisition of another data broker who has additional surveillance information to fill out for the people in your existing dataset, you can just stitch those new properties on like a fifth arm on your nightmarish data frankenstein.

What does this look like in practice? While in a bygone era Elsevier was merely a rentier holding publicly funded research hostage for profit, its parent company RELX is paradigmatic of the transformation of a more traditional information rentier into a sprawling, multimodal surveillance conglomerate (see [36]). RELX proudly describes itself as a gigantic haunted graph of data:

> Technology at RELX involves creating actionable insights from big data – large volumes of data in different formats being ingested at high speeds. We take this high-quality data from thousands of sources in varying formats – both structured and unstructured. We then extract the data points from the content, link the data points and enrich them to make it analysable. Finally, we apply advanced statistics and algorithms, such

[3] Facebook describes the notion of its platform as being just a means of interacting with its underlying data graph in corporate web design speak: "A useful tool for Facebook has been to think of the graph as the model and a Facebook page as the view—a projection of an entity or collection of entities that reside in the graph." [20]

[4] For a practical example, see a recent trio of blog posts from Etsy engineers that describe the process of scaling their database system.

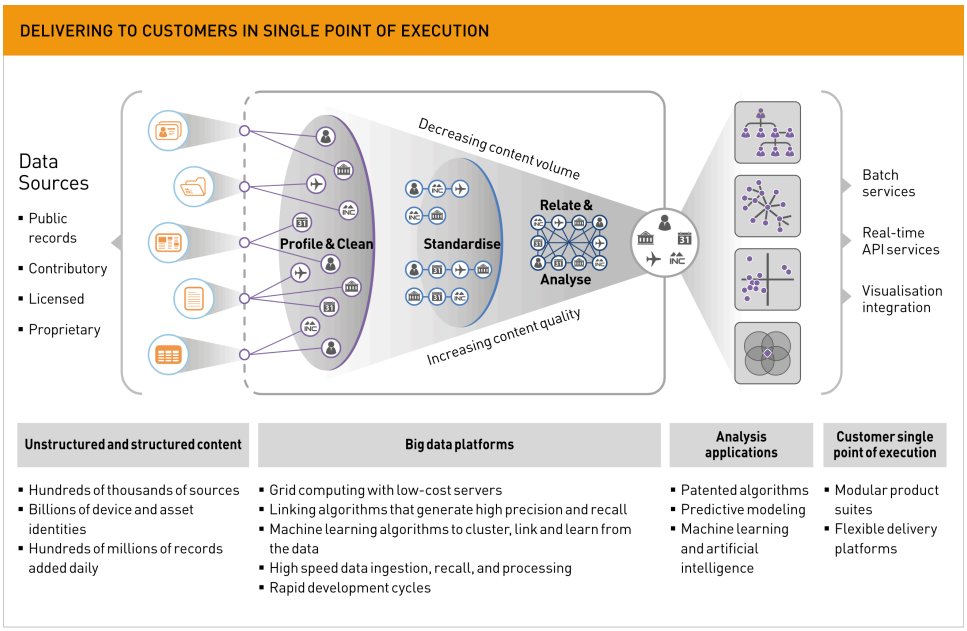
[5]

> That is because knowledge graphs aim to solve the data incongruence problem, which is one of the biggest operational headaches for corporates, says Atkin. "Corporates suffer from technology fragmentation and as a result have a lot of data that doesn't align across the organization. Doing the hard work to fix this data incongruence reality is a pre-requisite for realizing business value," he says. [27]

[6] I am aware graph databases are not magic and this is an extraordinarily simplified example. The principle is the point, not all the subtle ways the implementations of graph databases are hard.

as machine learning and natural language processing, to provide professional customers with the actionable insights they need to do their jobs.

We are continually building new products and data and technology platforms, re-using approaches and technologies across the company to create platforms that are reliable, scalable and secure. **Even though we serve different segments with different content sets, the nature of the problems solved and the way we apply technology has commonalities across the company.** [37]



In its 2022 Annual Report, RELX describes its business model as ingesting large quantities of data, linking them together, and deriving platforms from them. [37]

While to any individual market segment or class of customers RELX and its subsidiaries might look like a portfolio of separate platforms and applications, one can only make sense of the company by thinking of each of them as a view on an interconnected graph of data[7]. Each additional source of data, either by acquiring new companies or by expanding their existing control of informational access points has the potential to create some combinatorically new set of opportunities for new platforms.

For example, RELX is able to gather surveillance data on researcher attention data through the tracking in its ScienceDirect and Mendeley platforms. It also collects a large amount of chemical data through its control of scientific publishing that it rents access to on its Reaxys platform, which is supplemented by its LexisNexis PatentSight database of patents. So far so normal.

What about the other sides of the multisided market? RELX is able to combine these and other data sources into new product. For pharmaceutical R&D companies, their bespoke Drug Design Optimization services advertise being able to use chemical, disease, and literature-based data to generate a priority list of potential therapeutic targets and drugs, as well as provide "competitive intelligence" about which targets are currently being studied, presumably identified from their ownership of the scientific literature coupled with surveillance data. Since clinicians don't trust pharmaceutical advertisements [39] , Elsevier uses its position as a perceived neutral third party to repackage advertisements as informational systems [40] , "journal-branded webinars," as well as a number of other avenues via its "360 degree advertising solutions" catalogue. So, by combining several data sources and platforms, Elsevier is able to offer pharmaceutical companies recommendations for candidate drugs above and beyond what would be possible with chemical

[7] Though apparently they have had historical difficulty actually getting that integration to work [38] .

information alone and then advertise their drugs directly to doctors.

Derivative platforms beget derivative platforms. Its integration into clinical systems by way of reference material is growing to include electronic health record (EHR) systems, and they are "developing clinical decision support applications [...] leveraging [their] proprietary health graph" [37] . Similarly, their integration into Apple's watchOS to track medications indicates their interest in directly tracking personal medical data.

That's all within biomedical sciences, but RELX's risk division also provides "comprehensive data, analytics, and decision tools for [...] life insurance carriers" [37] , so while we will never have the kind of external visibility into its infrastructure to say for certain, it's not difficult to imagine combining its diverse biomedical knowledge graph with personal medical information in order to sell risk-assessment services to health and life insurance companies. LexisNexis has personal data enough to serve as an "integral part" of the United States Immigrations and Customs Enforcement's (ICE) arrest and deportation program [41, 42] , including dragnet location data [43] , driving behavior data from internet-connected cars [44] , and payment and credit data as just a small sample from its large catalogue [45] of data aggregated and linked into comprehensive profiles [46] . The contemporary knowledge graph-powered surveillance conglomerate gains its versatility precisely from its ability to span many unrelated domains and deploy new platforms as opportunities present themselves. As new data sources are acquired, the combinatorics of possible surveillance products correspondingly explode.

This pattern is true across the information industry [28] . A handful of representatives from Microsoft, Google, Facebook, eBay, and IBM describe some elements of each of their knowledge graphs in a 2019 paper [20] . Each has different scopes, applications, and interaction with the other data and processing infrastructure at the company, but all emphasize the ability for their knowledge graphs to accomodate change, heterogeneity, conflicting data, inference, and facilitate work by distributed teams due to their self-documenting and modular nature. Neo4j, developers of an eponymous graph database library, describes in one case study among its hundreds of customers how the U.S. Army uses its "connected data" to track its equipment and estimate the cost of some new exploratory imperialism [47] . An analysis of Palantir's hundreds of patents for knowledge graph technology (eg. [48, 49, 50, 51]) describes its ambitions for its knowledge graph:

> There is evidence [...] that Palantir has infrastructural aspirations to become a general classification system for data integration [...] that can be tailored into a universal knowledge graph. [...] Palantir similarly imagines a world where its platform might serve as a "shadow" universal knowledge graph for governments, industries, and organizations. [52]

Knowledge graphs *as a technology* - like all technologies - are not intrinsically unethical. It is the structure of the capital-K capital-G Knowledge Graph in its particular construction as a set of property and power relationships set against the context of the platform web that is pathological. They represent the historical trajectory of semantic web ideas and technologies from something that we are intended to use and create directly into privately held data that we can only interact with through platforms. They are coproductive with the corporate and technical structure of surveillance capitalism, facilitating conglomerates that gobble up as many platforms and data sources as possible to stitch them into an expanding, heterogeneous graph of data. We will return to the underlying ideology of the knowledge graph and an alternative in the final section.

In particular, it is their "graph plus compute" structure - where some underlying graph of data is coupled with a set of algorithms and interfaces to view it - that is necessary to understand some of the more counterintuitive motivations of surveillance conglomerates. This structure complicates questions of "openness" versus "proprietariness," one of the deepest loci of criticism of the platformatized web,

and provides a different lens on ostensibly "open" or "public" knowledge graph-based infrastructure projects.

3 Public Graphs, Private Profits

3.0.1 Unqualified Openness Considered Harmful

A reader that I am constructing as a straw man for the sake of argument might ask: if the problem is information conglomerates stockpiling a massive quantity of proprietary data and renting use of it, isn't "open data" the answer? And to that reader I would gently shake my head and say a qualified "no."

"Openness," including open source, open standards, and open data, is a subtle tool that can be used both to dissolve and reinforce economic and political power [53] .

Free and open source software, with its noble (and decidedly non-monolithic [54]) goal of creating an ecosystem of free[8] software, is a means by which large information companies can harvest the commons and outsource labor costs [55, 56, 57, 58, 59] . There are countless examples of FOSS developers maintaing software widely used by companies making billions of dollars for little or no compensation - eg. core-js [60] , OpenSSL [61] , leftpad [62] , PLC4X [63] and so on. When an information company releases or supports an open source project it is rarely an act of altruism. The effect is to prevent another company from profiting from a proprietary version of that technology, signal virtue, drive recruitment, and create a centralized point to concentrate donated labor. Microsoft, a famously good actor in software, took this several steps further with GitHub, VSCode, and later Copilot, capturing a large chunk of the software development *process* in order to trick programmers to be the "humans in the loop" refining the neural network to write code and dilute their labor power [64, 65, 66, 67] .

[8] "free as in whatever will prevent you from @'ing me about getting some definition of free wrong."

"Peer production" models, a more generic term for public collaboration that includes FOSS, has similar discontents. The similar term "crowdsource[9]" quite literally describes a patronizing means of harvesting free labor via some typically gamified platform. Wikipedia is perhaps the most well-known example of peer production[10], and it too struggles with its position as a resource to be harvested by information conglomerates. In 2015, the increasing prevalence of Google's information boxes caused a substantial decline in wikipedia pageviews [70, 71] as its information was harvested into Google's knowledge graph, and a "will she, won't she" search engine arguably intended to avoid dependence on Google was at the heart of its 2014-2016 leadership crisis [72, 73] . After shuttering Freebase, Google has donated a substantial amount of money to kickstart its successor [74] Wikidata, presumably as a means of crowdsourcing the curation of its knowledge graph [75, 76, 77] .

[9] For critical work on crowdsourcing in the context of "open science," see [68] , and in the semantic web see [69] .

[10] I have written about the peculiar structure of Wikipedia among wikis previously, section 3.4.1 - "The Wiki Way" [7]

"Open" standards are yet another fraught domain of openness. For an example within academia, the seemingly-open Digital Object Identifier (DOI) system was concocted as a means for publishers to retain control of indexing research, avoiding the impact of the proposed free repository PubMedCentral and the high overhead of linking documents between publishers[11] (see sec. 3.1.1 in [7]). The non-profit standards body NISO's standards for indicating journal article versions [79] and licensing [80] are used by publishers to enforce their intellectual property monopolies and programmatically scour the web to prevent free access to publicly funded information [81] .

[11] "The potential benefit of the service that would become CrossRef was immediately apparent. Organizations such as AIP and IOP (Institute of Physics) had begun to link to each other's publications, and the impossibility of replicating such one-off arrangements across the industry was obvious. As Tim Ingoldsby later put it, **'All those linking agreements were going to kill us.'**" [78]

Schema.org, a standard intended to be the generic interchange ontology of the web, is another emblem of enclosure of the semantic web. Its introduction at the SemTech 2011 conference was cause for a rare point of agreement[12] between the then-warring maintainers of RDFa and Microformats: "folks, it's wrong for Google to dictate vocabularies, let's not lose sight of that" [82] . Though ostensibly open, its structure and emphases have been roundly criticized, eg. having a eurocentric

[12] (Intervening messages in the chat log have been omitted for clarity): `<tantek>` Hey Kavi - do you see what you've done here? `<tantek>` You've gotten a community leader of microformats.org (myself) and chair of W3C RDFa WG to *agree* `<edsu>` tantek: see, that's progress :) `<manu-db>` Yes - both RDFa and Microformats communities agree - sky will be falling, next.

bias towards commercially valuable information [83] . It encourages website maintainers to embed Schema.org annotations in their pages in exchange for a boost in search rankings — which Google then embeds in its infoboxes, driving down page views. More fundamentally it cements the notion that Linked Data is something that we are only intended to use to make our information more available to some search engine crawler rather than make use of for ourselves: "In general, the design decisions place more of the burden on consumers of the markup" [84] . It encodes the notion that there should be one "neutral" means of representing information for one (or a few) global search engines to understand, rather than for local negotiation over meaning and location. According to the transcribed Q&A after its 2011 announcement, the Google representatives characterized the creation of authoring tools like those created to make creative use of HTML more accessible as a potential "alternative path," but then dismissed the notion of improved tooling as "impossible" [85] .

Clearly, on its own, mere "openness" is no guarantee of virtue, and socio-technological systems must always be evaluated in their broader context: *what is open? why? who benefits?* Open source, open standards, and peer production models do not inherently challenge the rent-seeking behavior of information conglomerates, but can instead facilitate it.

In particular, the maintainers of corporate knowledge graphs want to reduce labor duplication by making use of some public knowledge graph that they can then "add value" to with shades of proprietary and personal data (emphasis mine):

> In a case like IBM clients, who build their own custom knowledge graphs, **the clients are not expected to tell the graph about basic knowledge.** For example, a cancer researcher is not going to teach the knowledge graph that skin is a form of tissue, or that St. Jude is a hospital in Memphis, Tennessee. This is known as **"general knowledge,"** captured in a general knowledge graph. **The next level of information is knowledge that is well known to anybody in the domain**—for example, carcinoma is a form of cancer or NHL more often stands for nonHodgkin lymphoma than National Hockey League in some contexts it may still mean that—say, in the patient record of an NHL player). **The client should need to input only the private and confidential knowledge** or any knowledge that the system does not yet know. [20]

The creation of a collection of more domain-specific ontologies and tooling for ingesting previously unstructured data would allow for a new kind of globally linked knowledge graph ecosystem — making use of a broader range of publicly-available data, as well as facilitating new markets for renting access to interoperable data. Five information conglomerates conclude their joint paper on knowledge graphs accordingly:

> The natural question from our discussion in this article is whether different knowledge graphs can someday share certain core elements, such as descriptions of people, places, and similar entities. [20]

Having such standards be under the stewardship of ostensibly neutral and open third-parties provides cover for powerful actors exerting their influence and helps overcome the initial energy barrier to realizing network effects from their broad use [86, 87] . Peter Mika, the director of Semantic Search at Yahoo Labs, describes this need for third-party intervention in domain-specific standards:

> A natural next step for Knowledge Graphs is to **extend beyond the boundaries of organisations,** connecting data assets of companies along business value chains. This process is still at an early stage, and **there is a need for trade associations or industry-specific standards organisations to step in,** especially when it comes to developing shared entity identifier schemes. [88]

As with search, we should be particularly wary of information infrastructures that are *technically* open[13] but embed design logics that preserve the hegemony of the organizations that have the resources to make use of them. The existing organization of industrial knowledge graphs as chimeric "data + compute" models give a hint at what we might look for in public knowledge graphs: the data is open, but to make use of it we have to rely on some proprietary algorithm or cloud infrastructure.

Unfortunately, that is exactly what at least two US Federal agencies have in mind: the NIH and NSF are both in the thick of engineering cloud-based knowledge graph infrastructures and domain-specific ontologies with all the trappings of technology that fills the stated needs of information conglomerates at the expense of the people it is outwardly intended to serve. We will describe those efforts and their already apparent risks as a way of understanding how these technologies illustrate and reinforce the ideological and practical dominance of the existing corporate informational ecosystem — and to articulate an alternative.

Add note that we are assuming that people are working with the best of intentinos here, adn that it is hard to imagine an alternative system when the existing one is so dominant! These are mostly good people trying to do good things in a system that's rotten.

[13] Go ahead, try and make your own web crawler to compete with Google - all the information is just out there in public on the open web!

3.0.2 NIH: The Biomedical Translator

Note:

This section is reproduced from, focuses, and expands on "Linked Data or Surveillance Capitalism?" from [7] .

The NIH's Biomedical Data Translator[14] project was initially described in its 2016 Strategic Plan for Data Science as a means of translating between biomedical data formats:

> Through its Biomedical Data Translator program, the National Center for Advancing Translational Sciences (NCATS) is supporting research to develop ways to connect conventionally separated data types to one another to make them more useful for researchers and the public. [**?**]

The original funding statement from 2016 is similarly humble, and press releases through 2017 also speak mostly in terms of querying the data – though some ambition begins to creep in. By 2019, the vision for the project had shifted from *translating* between data types into the realm of heterogeneous linkages in some meta-level system for linking and *reasoning* over them.

In their piece "Toward a Universal Biomedical Translator," then in a feasibility assessment phase, the members of the Translator Consortium assert that universal translation between biomedical data is impossible[15][89] . The impossibility they saw was not that of conflicting political demands on the structure of organization (as per [53]), but of the sheer numeracy of the data and vocabularies needed to describe them. The risk posed by a lack of a universal "language" was not being able to index all possible data, rather than inaccuracy or inequity[16].

Undaunted by their stated belief in the impossibility of a universalizing ontology, the Consortium created one in their biolink model[17] [92, 91] . Biolink consists of a hierarchy of general[18] classes: eg. a BiologicalEntity like a Gene, or a ChemicalEntity like a Drug. Classes can then linked by any number of properties, or "Slots[19]," like a therapeutic procedure that treats a disease.

Biolink was designed to be a sort of "meta ontology," or a means of mapping different domain-specific biomedical ontologies onto a common vocabulary[20]. This design reflects the structure of the rest of the Translator ecosystem: the interaction with domain-specific ontologies, the kinds of data sources it uses, and the way that end users are expected to interact with the Translator.

[14] Or, just "Translator"

[15] "First, we assert that a single monolithic data set that directly connects the complete set of clinical characteristics to the complete set of biomolecular features, including "-omics" data, will never exist because the number of characteristics and features is constantly shifting and exponentially growing. [...] We also assert that there is no single language, software or natural, with which to express clinical and biomolecular observations—these observations are necessarily and appropriately linked to the measurement technologies that produce them, as well as the nuances of language."[89]

[16] In an odd mixture of metaphors, members of the Translator consortium introduced the project with a piece titled "Deconstructing the Translational Tower of Babel." [90] It is unclear why an effort to create a universalizing ontology would be deconstructing a tower of babel, as it was the power of a unified language that allowed it to be built.

[17] The title of the Biolink paper is "Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science" [91]

[18] General as opposed to an ontology like MONDO [93] that identifies specific diseases.

[19] or links, labeled edges.

[20] To their credit, the Translator project seems to have made some of the long-delayed tooling for declaring a schema in a more accessible syntax[^yaml] than RDFS/OWL and generating representations in multiple formats, from JSON-LD to pydantic models. The Biolink paper also mentions a "Node Normalization Service" for being able to resolve Linked Data entities from different vocabularies that have been declared to be the same thing, but at the time of writing development seems to have

As a meta-ontology, biolink is targeted towards "meta data." Rather than accomodating "raw data[21]," biolink is expected to operate at the level of "knowledge," or "generally accepted, universal assertions derived from the accumulation of information" [95] : this procedure treats that disease, this chemical interacts with that one, etc.

The primary way Biolink is used within the Translator is to structure a registry of database APIs, each called a "Knowledge Source." Knowledge Sources use biolink to declare that they are able to provide assertions about a particular set of classes or slots, like drugs that affect genetic expression, which makes them part of the Translator's distributed Knowledge Graph. The Translator project, in this universalizing impulse, recapitulates some of the early beliefs of the Semantic Web updated with some of the techniques of Linked Data. Since acquiring Knowledge is just a matter of creating the right tools rather than a social process, NIH RePORTER shows a series of grants for small councils of experts to create domain-specific ontologies and Knowledge Sources.

This structure strongly constrains who is intended to be able to contribute to the Translator: highly curated biomedical informatics platforms, rather than basic researchers. This, in turn, reflects deeper beliefs about the nature of information within the Translator ecosystem: "knowledge" is not a social, contextual, or dialogical phenomenon, but a "natural resource" that can be mined from information that is "out there." A scientific paper is a neutral carrier of a factual link between entities. The meaning of "translation," in some uses, has shifted from translating *between data formats*, to *"translating information into knowledge"* [89] . This is, of course, the ideology of Big Data: "when heterogeneous networks are connected at a massive scale, new knowledge can be extracted as an emergent property of the network" [96] . The Translator imagines itself as a refinery, converting crude data into knowledge that can fuel platforms.

The platforms that the translator imagines are means by which plain language queries can be translated into graph queries and have answers returned by some algorithmic "reasoning agent" that queries the Knowledge Providers and synthesizes a response [97, 98, 99, 91, 100] . We are not intended to use the data from Knowledge Providers directly, as it is likely to be incomplete or conflicting. Instead, the imagined use is as a recommendation system for researchers to target their research or for doctors to render care.

Several pilot experiments have demonstrated combining some aggregated patient records with the broader knowledge graph in order to, eg. identify new risk markers for disease [96, 101, 102, 103] . These systems layer personal records underneath "general" biomedical information like drug interactions and biological processes and use the extended information from the graph to infer information both about the nature of the disease and the patient. A platform integrated with the UCSF electronic health record system that layers disaggregated clinical records under the general knowledge graph is already apparently in a state of mature development [104] .

It is only with the inclusion of patient records into the knowledge graph that it becomes possible to use in a clinical setting: for even basic queries like "which drugs treat this disease" one has to be aware of patient qualities like allergies and comorbid conditions. To know how to treat the generic diagnosis of "gender dysphoria," one needs to know which gender the patient is experiencing dysphoria about. The logic of knowledge graph isn't just hungry for *some* personal medical data, the promise of the knowledge graph is that more data **always** improves the computations performed on it[22].

Why might we be critical about the NIH funding a series of projects to unify biomedical and personal health data in some universalized, platformatized knowledge graph? In short: because it won't work and will inevitably be captured by the surveillance industry.

[21] In a 2018 presentation by one of Biolink's authors: "What NOT to use the biolink-model for: Raw data, Metadata about a dataset" with some caveat that the underlying metamodel might still be useful [94] .

[22] The answer to a question posed as an algorithmic problem is always more data: "These results suggest that if more EHR concepts were mapped to SPOKE, a significant improvement in the classifier could be achieved." [101]

First, as with any machine-learning based system, the algorithm can only reflect the implicit structure of its creation, including the beliefs and values of its architects [105, 106] , its training data and accompanying bias [107] , and so on. The "mass of data" approach ML tools lend themselves to, in this case, querying hundreds of independently operated databases, makes dissecting the provenance of every entry from every data provider effectively impossible. For example, one of the providers, mydisease.info was more than happy to respond to a query for the outmoded definition of "transsexualism" as a disease [108] along with a list of genes and variants that supposedly "cause" it - see for yourself. At the time of the search, tracing the source of that entry first led to the disease ontology DOID:1234, which has an official IRI, but in this case was being served by a graph aggregator Ontobee (Archive Link), which in turn listed this unofficial github repository **maintained by a single person** as its source[23]. This is, presumably, the fragility and inconsistency in input data that the machine learning layer is intended to putty over.

[23] I submitted a pull request to remove it, and a full year later it was merged!

If the graph encodes being transgender as a disease, it is not farfetched to imagine the ranking system attempting to "cure" it. In a seemingly prerelease version of the translator's query engine, ARAX, it does just that: in a query for entities with a `biolink:treats` link to gender dysphoria[24], it ranks the standard therapeutics [109, 110] Testosterone and Estradiol 6th and 10th of 11, respectively — behind a recommendation for Lithium (4th) and Pimozide (5th) due to an automated text scrape of two conversion therapy papers[25]. Queries to ARAX for treatments for gender identity disorder helpfully yielded "zinc" and "water," offering a paper from the translator group that describes automated drug recommendation as the only provenance [111] . A query for treatments for `DOID:1233` "transvestism" was predictably troubling. The ROBOKOP [112] query engine behaved similarly, answering [a query for genes associated with]({{ "/data/ROBOKOP_message.json" | relative_url }}) gender dysphoria with exclusively trivial or incorrect responses[26].

[24] To its credit, ARAX does transform the request for `DOID:10919` to `MONDO:0001153` - gender dysphoria.

[25] as well as a recommendation for "date allergenic extract" from a misinterpretation of "to date" in the abstract of a paper that reads "Cross-sex hormonal treatment (CHT) used for gender dysphoria (GD) could by itself affect well-being without the use of genital surgery; however, **to date,** there is a paucity of studies investigating the effects of CHT alone"

[26] ITSN2 was identified in an unrelated paper about attachment patterns, HSD17B3 and 5a-RD2 were incorrectly identified as HSD17B13 and DHRS11 from another paper, POMC and OPN1SW were sourced from two papers that don't mention them. Androgen receptors were also identified, which is probably true, but almost trivially so.

It is critically important to understand that with an algorithmic, graph-based precision medicine system like this **harm can occur even without intended malice.** The power of the graph model for precision medicine is precisely its ability to make use of the extended structure of the graph[27]. The "value added" by the personalized biomedical graph is being able to incorporate the patient's personal information like genetics, environment, and comorbidities into diagnosis and treatment. So, harmful information embedded within a graph — like transness being a disease in search of a cure — means the system either a) incorporates that harm into its outputs for seemingly unrelated queries or b) doesn't work. This explodes the risk surface for medically marginalized people to include the entire Translator ecosystem: from the violence historically encoded in mainstream medical practices and ontologies (eg. [108, 113] , among many), to incorrectly encoded information like that from automated text mining, to explicitly adversarial information injected into the graph through some crowdsourcing portal like this one [114] . Each of these sources of harm could influence medical care in ways that *even a well-meaning clinician might not be able to recognize.*

[27] eg. Some members of the SPOKE project, a Knowledge Provider for the Translator project, describe the effects of the extended graph as "pushing" or influencing the "flow" of information: > "For this patient, information flows from Carbamazepine to a set of Disease nodes (either through "treated by" or "contraindicated for" edges) and then (either directly or through an additional Disease or Gene node) to the genes CNP, MAG, or PTEN which are all components of "Myelin sheath adaxonal region." [101]

The risk of harm is again multiplied by the potential for harmful outputs of a biomedical knowledge graph system to trickle through medical practice and re-enter as training data. The Consortium also describes the potential for ranking algorithms to be continuously updated based on usage or results in research or clinical practice[28] [89] . Existing harm in medical practice, amplified by any induced by the Translator system, could then be re-encoded as implicit medical consensus in an opaque recommendation algorithm. There is, of course, no unique "loss function" to evaluate health. One belief system's vision of health is demonic pathology in another. Say an insurance company uses the clinical recommendations of some algorithm built off the Translator's graph to evaluate its coverage of medical procedures. This gives them license to lower their bottom line under cover of some seemingly objective but fundamentally unaccountable algorithm. Could a collection of anti-abortion clinics giving one star to abortion in every case meaningfully influence whether abortion is prescribed or covered? Why not? Who moderates

[28]

> "The Reasoners then return ranked and scored potential translations with provenance and supporting evidence. The user is then able to evaluate the translations and supporting evidence and provide feedback to the Reasoners, thus promoting continuous improvement of the prototype system." [89]

the graph?

The centralized structure of the Translator's Knowledge Providers and query engines now form a small group responsible for curating the entire structure of biomedical information. The curation process could be "crowdsourced" to allow affected communities to suggest improvements, but the platformatized nature of the Translator both concentrates decisionmaking power and diffuses responsibility across a string of platform holders. Who is supposed to fix incorrect or harmful query responses? Is it the responsibility of the potentially dozens of Knowledge Providers, the swarm of reasoning agents, or the frontend wrapper you pay a monthly subscription for? It is the platformatized nature of the Translator itself that creates the need for centralized moderation in the first place. The design of the Translator to evolve into a series of "user-" or customer-facing platforms that aspire to universality binds it to all the regulatory burden any biomedical technology bears. The cost of moderation will of course be enormous, placing a fundamental constraint on its lifespan as a publicly funded project — and a strong incentive towards co-option by the information conglomerates capable of paying it[29].

These problems hint at the likely fate of the Translator project. Rather than integrating into the daily practice of researchers, the centralized process of creating Knowledge Providers can only be maintained for as long as the grant funding for the Translator project lasts. When queried at the time of writing, of the 25 knowledge providers that were responsive to information about "Anything that is related to the common cold," 22 were unresponsive or timed out. How the Translator is intended to work by its architects is almost irrelevant compared to the question of what happens to it *after the project ends.*

Linking biomedical and patient data in a single platform is a natural route towards a multisided market where records management apps are sold to patients, treatment recommendation systems are sold to clinicians, research tools and advertising opportunities are sold to pharmaceutical companies, risk metrics are sold to insurance companies, and so on. The contours of this market are already clear.

As a non-exhaustive set of examples:

- I have already described **RELX**'s interest in personal biomedical data. Their 2022 Annual Report [37] is the first year where they explicitly describe their entrance into the patient data market[30]. RELX is a particularly worrying example because of their established roles among academics, medical workers, and insurance providers.
- **Amazon** already has a broad home surveillance portfolio [117] , and has been aggressively expanding into health technology [118] and even literally providing health care [119, 120] , which could be particularly dangerous with the uploading of all scientific and medical data onto AWS with entirely unenforceable promises of data privacy through NIH's STRIDES program [121] .
- **Google** already includes medical conditions in its surveillance-backed advertising profiles [122, 123] , and is edging its way into wearable health data with eg. its acquisition of FitBit [124] . It also already has a system, Med-PALM, for biomedical question answering based on large language models [125, 126, 127] . Search is a primary entrypoint for many people searching for health information, and Google presumably would be more than happy to merge that data with a generalized biomedical knowledge graph.
- **Apple** already has a matured Health ecosystem of apps and services for both patients, clinicians, and researchers [128, 129] and has a similar exposure to relevant data and control of platforms (iOS, watchOS) to make use of it, though they have marketed themselves in the surveillance space as a defender of privacy.
- Of course **Microsoft** [130] and **IBM** [131] are also in play.

[29] There is a clear analogy to the recent push to increase internet content regulation by social media companies [115] . A platform makes a quasi-universal social space for profit, moderation then has to scale with the size of the platform, then it lobbies to increase regulatory burden to a point that is impossible to maintain for all but already-scaled companies. It is only the quasi-universality of the platform that makes the moderation burden so high in the first place, however, compared to eg. a decentralized medium that might have a structurally different disposition to moderation (see [116]).

[30] In commercial healthcare, identity, claims and provider data is combined with patient information to assist healthcare providers, pharmacies and insurers in delivering improved health outcomes, ensuring accurate and complete provider data and regulatory compliance. [37]

The design of the Translator project reflects the prevailing logic of the surveillance economy as powered by knowledge graphs, and is poised to be swallowed up by it. Rather than a means for us to collectively make sense together, they have imagined a cloud-driven system where a small group of experts wave a wand of unknowable algorithms over a bulging plastic trash bag of data to pull out the Magic Knowledge Rabbit. The noble intention of making a generalized biomedical knowledge graph for the public good is unlikely to be realized. In the process, though, the NIH will have funded facilitating technologies and standards for the merger of personal medical surveillance with the broader landscape of biomedical data. Academics will have new vectors by which they become unwitting or unwilling collaborators with surveillance and data brokers, lending what credibility they have left to a landscape of buggy black boxes of biopolitical control. And, most importantly, vulnerable populations will have dozens of new ways to be marginalized by the techno-political medical establishment.

3.0.3 NSF: Open Knowledge Network

While the NIH builds a set of universal knowledge graphs for biomedical information, the NSF is building them for everything else. Its Open Knowledge Network (OKN) project intends to "provide an essential public-data infrastructure for enabling an AI-driven future." [132] OKN is in an earlier stage of development than the Translator, so this section is less focused on the details of individual projects and more to argue the pattern of public/private knowledge graphs is an emerging consensus.

Compared to the Translator, the OKN pulls punches for neither its utopian promises nor obvious risks. Some sections of its roadmap are written in a style where each line shoots for the stars because even if some of them miss the result is a constellation of absolute bangers like "Harnessing the vast amounts of data generated in every sphere of life and transforming them into useful, actionable information and knowledge is crucial to the efficient functioning of a modern society[31]" [132] . The project was initially proposed in 2017, went through two cohorts of projects within the NSF Convergence Accelerator in 2019 and 2020[32], and invited a broader submission of proposals in November 2021 [134] . The roadmap comes at the end of a series of workshops in 2022 intended to scope and outline the OKN, so there is still very little public evidence of its progress to evaluate[33].

[31] If you could rig an MS Word template to punctuate sentences with "Whoomp! (There It Is)," they would have.

[32] The Convergence Accelerator is a project specifically designed to provide public research funding to for-profit industries [133]

[33] SPOKE, discussed previously, was funded by both the Translator project [135] and OKN [136] , and KnowWhereGraph is another notable early prototype [137]

Its domain is much broader than the Translator, and is unmistakeably bound up in both the United States Federal Government's military and political interests in Artificial Intelligence [138] and the information economy's interests in making a universal space where all information can be bought and sold with minimal friction [139] . Where the Translator has the near-inevitable risk of being captured by information conglomerates, through the euphemism of "public private partnership" the OKN makes clear it was already captured at inception: the team behind the SPOKE biomedical knowledge network immediately spun off a for-profit startup to sell the graph as a cloud service [140] , abandoning further UX development of its publicly accessible demo.

Without mincing words, the OKN intends to make a Universal Knowledge Graph of Everything. They check all the boxes[34]: a) make authoritative schemas for everything, b) link them all together, c) ingest data from as many sources as possible at whatever quality available, d) integrate private with public data e) put it all in the cloud! (p. 18-19 "Creating an OKN" [139]).

[34] Hopefully this pattern is familiar.

They OKN describes its work using a vocabulary of "vertical" "horizontal," where "vertical" applications refer to specific uses or domains like energy or health data, and "horizontal" themes like technologies and governance are shared across all domains. The work of the OKN is organized around specific use cases either within a "vertical" topic or a specific "horizontal" theme with the intent of later building them together into a shared infrastructure. The collection of "vertical" topics identified in the 2022 roadmap hint at the effectively unbounded scope of the OKN:

accelerated capitalism via supply chain logistics, more tightly integrated weapons development, a handful of climate change projects, an omniscient financial system, and so on. Each imagines the primary problem in a given domain not as structural exploitation or injustice, but a lack of data[35].

A collection of "vertical" topical working groups in the 2022 roadmap centered on an algorithmic justice system are illustrative: An **Integrated Justice Platform** group describes how greater surveillance across every contact people have with the US Justice System is necessary to decrease bias. The group outlines a wish list of data sources they would integrate - arrest and booking, jail, trial, prosecution, and the rest. A **Decarceration** group[36] describes extending that surveillance through to the rest of incarcerated people's lives after they are released - rehab, parole, foster care, shelters, public services, etc. A **Homelessness** group intends to track unhoused people in order to match them to available resources. A **Decision Support for Government**[37] group describes bundling up these and other data sources into platforms for making "data driven decisions" on topics including crime and policing.

On their own, each of these groups describes noble goals: decreasing bias in the justice system, providing resources to formerly incarcerated or unhoused people, making government decisions more efficient. Taken together, however, the projects describe a panoptical surveillance system that wouldn't even need to be reconfigured to be used for algorithmically-enhanced oppression. I doubt any of the researchers in these groups intend for their work to be used for state violence, but *Palantir doesn't care what academics intended their tools to be used for*[38].

The motivations behind integrating government data sources and automating public benefit delivery cannot overcome the context of systemic oppression they are embedded within. Group H, the "Homelessness OKN" group, takes particular effort[39] to focus on the needs of the unhoused and address the potential risks of "track[ing] homelessness in real time, [and] identify[ing] available homelessness programs and services," but misses the already-real harms of similar prior efforts. Virginia Eubanks describes how Los Angeles County's Coordinated Entry System — a program very much like that described by group H, intended to match unhoused people with housing supply by integrating previously siloed data systems — operates as a sophisticated mechanism of control and punishment:

> For Gary Boatwright and tens of thousands of others who have not been matched with any services, coordinated entry seems to collect increasingly sensitive, intrusive data to track their movements and behavior, but doesn't offer anything in return. [...] Moreover, the pattern of increased data collection, sharing, and surveillance reinforces the criminalization of the unhoused, if only because **so many of the basic conditions of being homeless are also officially crimes.** [...] The tickets turn into warrants, and then law enforcement has further reason to search the databases to find "fugitives." Thus, **data collection, storage, and sharing in homeless service programs are often starting points in a process that criminalizes the poor.** [...]
>
> Further integrating programs aimed at providing economic security and those focused on crime control threatens to turn routine survival strategies of those living in extreme poverty into crimes. **The constant data collection from a vast array of high-tech tools wielded by homeless services, business improvement districts, and law enforcement create what Skid Row residents perceive as a net of constraint that influences their every decision.** Daily, they feel encouraged to self-deport or self-imprison. Those living outdoors in encampments feel pressured to constantly be on the move. Those housed in SROs or permanent supportive housing feel equally intense pressure to stay inside and out of the public eye. [...] **Coordinated entry is not just a system for managing information or matching demand to supply. It is**

[35] A recurring pattern in techno-solutionism: > "These perspectives assume that complex controversies can be solved by getting correct information where it needs to go as efficiently as possible. In this model, political conflict arises primarily from a lack of information. If we just gather all the facts, systems engineers assume, the correct answers to intractable policy problems like homelessness will be simple, uncontroversial, and widely shared. > > But, for better or worse, **this is not how politics work.**" [141]

[36] Including a representative from Booz Allen Hamilton, which may be familiar as the former employer of Edward Snowden, who was working for them on a contract with the NSA which gave him access to the details of its PRISM mass-surveillance program.

[37] See [142] for discussion of algorithmic governance in a "smart city."

[38] Palantir prides itself on its ability to continuously add new data sources: > "Because one of Palantir's biggest selling points is the ease with which new, external data sources can be incorporated into the platform, its coverage grows every day. LAPD data, data collected by other government agencies, and external data, including privately collected data accessed through licensing agreements with data brokers, are among at least 19 databases feeding Palantir at JRIC." [143]

[39] The "Innovation Sprint" is essentially as an extended pitch session for future work, which is both important context as a strong counterincentive to series ethical consideration of the projects — and also a demonstration of why it might not be good to organizing infrastructural projects as pitch sessions rather than from some ethical foundation.

a surveillance system for sorting and criminalizing the poor. [141]

It is impossible to consider integrated data in government without confronting the reality of algorithmic policing. Under its Strategic Plan goal of "Realiz[ing] Tomorrow's Government Today" Los Angeles County has already been integrating its information systems, including creating a unified system of law enforcement and other public service data "to identify super utilizers of justice and health system resources"[40] [145, 146] . Many police departments — including the LAPD — already have access to the kind of linked data ecosystems described by the OKN by renting them from private data brokers like Palantir [143, 147] . These data infrastructures facilitate the well-described feedback loop of predictive policing, where areas already subject to historical economic and racist violence are classified as "high-crime areas," more police are concentrated there, in turn causing them to measure or create more crime[41] [143, 148, 150, 149, 151, 152, 153] . The reformist idea that more data will help us "police the police" is belied by the resolute history of more data allowing the police to innovate on information asymmetries to create new expressions of power [154, 155] .

The critical difference between prior infrastructures and those imagined by the OKN is that they are explicitly designed to be linked into a continuous network of data that enables the same kind of data-driven decisionmaking that drives predictive policing for *any* system. We should not be imagining the utterly mechanistic bureaucracy of *Kafka* here, but rather the deeply expressive and personal exercise of power of Terry Gilliam's *Brazil.* Widespread algorithmic governance doesn't necesarily look like a faceless bureaucracy where all decisions are made by a computer, existing algorithmic systems like predictive policing and the working conditions at Amazon warehouses retain the very human domain of *discretion* (see [154]). The algorithms and seemingly open infrastructures of these two projects purport themselves as objective and egalitarian, but who they are built for, who gets to provides the inputs, and who decides which outputs matter make their reality very different.

The very act of creating information infrastructures intended to algorithmically solve the world's problems is itself an expression of power-based discretion that diffuses energy that might be better spent elsewhere: rather than attempt to address the root cause, we can make a big show of *doing something* by diverting a large amount of resources and labor to gathering data and deriving "insights" about them. Beyond the specific risks of algorithmic policing and public benefits assessments, these projects presuppose an enlightened technocrat class as the principle agent of social good and design technologies accordingly.

This is the properly *managerial* core of the positive vision of platform-driven mass surveillance: that a grand, unified graph of everything will allow the truth to emerge from the Big Data so that decisionmakers can divine what is best for the commoners who could not possibly understand the complexities of their health, environment, or social systems themselves. The conflict with the reality of the bureaucracies that might enact or support them temper those dreams, to the degree the prevailing corporate design logics of cloud-based infrastructure and what constitutes a fundable idea hadn't already[42]. Rather than a smoothly flowing ocean of data, their fate is more likely to extend and elaborate the disjointed wash of SaaS that defines institutional information systems. The petty tyrants will will gain new widgets on their dashboards rented from the information companies who absorbed the ontologies and harmonized data sources, automating their discretion under the guise of objectively meausurable fact. Meanwhile the new housing capacity might never arrive.

The prominent role of climate change among the topics identified by the OKN project is at once reassuring and depressing. Maybe what we need to solve climate change isn't data, it's to organize effective climate movements outside of the algorithmically disorienting information ecosystems designed to pump us full of engagement-maximizing rage-bait and transmute all movement building into in-

[40] ...and then outsourcing the maintenance and risk of it being breached [144]

[41]

> These visits often resulted in other, unrelated arrests that further victimized families and added to the likelihood that they would be visited and harassed again. In one incident, the mother of a targeted teenager was issued a $2,500 fine when police sent to check in on her child saw chickens in the backyard. In another incident, a father was arrested when police looked through the window of the house and saw a 17-year-old smoking a cigarette. These are the kinds of usually unreported crimes that occur in all neighborhoods, across all economic strata—but which only those marginalized people who live under near constant policing are penalized for. [148, 149]

[42] For a fuller discussion of utopias, power, imagination, managerialism, and its intersections with bureaucracy, see [156] > The increasing interpenetration of government, university, and private firms has led all parties to adopt language, sensibilities, and organizational forms that originated in the corporate world. While this might have helped somewhat in speeding up the creation of immediately marketable products — as this is what corporate bureaucracies are designed to do — in terms of fostering original research, the results have been catastrophic. [...] > > A timid, bureaucratic spirit has come to suffuse every aspect of intellectual life. More often than not, it comes cloaked in a language of creativity, initiative, and entrepreneurialism. But the language is meaningless. The sort of thinkers most likely to come up with new conceptual breakthroughs are the least likely to receive funding, and if, somehow, breakthroughs nonetheless occur, they will almost certainly never find anyone willing to follow up on the most daring implica-

fluencer culture. Maybe what we need to address mass poverty isn't data, it's to dismantle the mechanisms of mass extraction that are increasingly powered by economies of surveillance. Maybe what we need to make the criminal justice system less racist isn't more data to feed into predictive policing algorithms, but to abolish the police.

These two projects show the outlines of public information infrastructures to come.

The two major public research funding agencies in the US have both devised novel funding mechanisms to be able to bypass typical review and include private industry in their data infrastructure projects [133**?**] . These data infrastructures consist of a number of sub-projects for building new domain-specific and universalizing Semantic Web ontologies and cloud-based platforms for data storage and retrieval. Both are both explicitly oriented towards exposing structured data to "AI" and other derivative "big data" applications, rather than towards integrating in the daily work of researchers or the public at large. These and other[43] efforts like NIH's STRIDES initiative point towards a cloud-driven SaaS/PaaS future for public data infrastructure [**?**] .

[43] It's out of scope here, but another point of comparison and contrast is the EU's European Open Science Cloud (ESOC) project [157, 158, 159]

The Translator and OKN and their subprojects have many possible fates: their grant funding could peter out and they could amount to very little beyond the scattered prototypes and spinoff startups that they've currently produced. They could flourish and become exactly what their creators intend them to be - the seamless data infrastructures of the future that manage to miracuously avoid all the potential for harm described above.

More important than the outcomes of these projects in particular is how the ruts of collective imagination drive both projects towards very similar designs with very similar flaws. They can only be understood against the backdrop of the history and present structure of the platformatized cloud-driven information economy writ large. Facing the limits of proprietary ontologies in private knowledge graphs, the information industry wants a set of cross-domain "top level" ontologies to enable the smooth interchange of public information that can then be integrated with "lower-level" private ontologies for an even greater array of surveillance-backed knowledge-as-a-service platforms. Under the guiding star of openness as an end in itself, researchers and funding agencies seem keen to provide it, and in partnership with private industry have adopted the logic of their platforms.

The lofty ambitions of universality alongside the intrinsic constraints of the cloud model are far from the only inevitable outcomes for public information infrastructure. Instead they are predicated on a very specific organization of ownership and power. Adopting the infrastructural pattern of the cloud strongly constrains the ambition of these infrastructures, as well as all but guarantees that whatever is produced by their period of active publicly-funded development will be swallowed up by information conglomerates hungry for control over the very valuable domains of medical and government data. Rather than challenging the dominance of the information rentiers by creating truly transformative public information infrastructure, these projects recapitulate their logic and deepen their dominance.

I am not painting a picture of active malice and conspiracy among bad actors to seize otherwise altruistic infrastructure. I am describing how uninterrogated prevailing values, beliefs, and existing infrastructures constrain what we can collectively imagine as possible — the outcome of **ideology.**

4 Infrastructural Ideologies

The Cloud is not a neutral, inevitable, or optimal form of the web — it has been actively constructed to facilitate a particular set of power and property relationships that make up the web's dominant business model. It is supported by a system of

values and *beliefs* that are consciously affirmed to various degrees in a positive feedback loop with the expertise and resource investment that make its enabling technologies more developed and obvious than alternatives, in turn fueling the truth of those beliefs, including that of the inevitability of the cloud model itself.

The history of the web is an odd substance: always present and eternal, yet profoundly ephemeral and immediately forgotten. It becomes increasingly difficult to imagine obscure roads not taken in the deeper architecture of the internet[44] with every fork. Before the dominance of compute in the cloud, distributed computing projects like folding@home were more powerful than any supercomputer [160] — and during the first year of the COVID-19 pandemic was able to fill a role the cloud couldn't, with a wave of volunteers breaking the exascale computing barrier and making it more powerful than the top 100 supercomputers combined [161] . Before the dominance of cloud video streaming platforms, peer-to-peer systems accounted for a majority of global internet traffic: in the mid-2000's between 49% and 95%, depending on the survey [162, 163] .

[44] Except by the scores of beloved nerds in exile on the freer parts of the internet who remember the death of IRC and RSS and the weaponization of JavaScript **acutely and personally.**

The Cloud paradigm is at once phenomenally successful and riddled with obviously undesirable qualities. Cloud services promise large volumes of hassle-free storage — but also make our data take a round trip across the planet if we want to transfer it between computers in the same room. Cloud systems are impressive feats of engineering, capable of serving immense quantities of data from relay CDNs dotted around the globe — but only need to do so because of the preposterous inefficiency of needing to re-serve data like streaming video in full each time they are accessed. Cloud systems can be made to have very high uptime, but then they do go down their dramatic centralization causes massive internet-wide blackouts even for systems that only depend on them indirectly [164, 165] . Delivering cloud platforms through the browser requires less setup than local software, but the complexity of the underlying web standards make it effectively impossible [166] to escape the near-monopoly[45] of Chrome[46], and make many services completely unavailable if the internet goes out or even slows down.

[45] The last major competitor being Firefox with market share in the low teens. Hang in there little fox!

[46] Which Google uses as a surveillance platform and a weapon which, according to unredacted court records detailing its "Privacy Sandbox" project, they plan to use for a forceful takeover of the rest of the global ad market in the name of privacy: "Google's new scheme is, in essence, to wall off the entire portion of the internet that consumers access through Google's Chrome browser." [167] .

That these tradeoffs are either not considered or seen as the natural constraints of internet technologies is precisely the evidence of The Cloud as ***ideology.*** By treating The Cloud as a system of *belief* we can better understand how its acolytes imagine the world they are creating — and what they have in store to get us there. In particular, it is only possible to understand the *meaning* and *intention* of the surge of **chatbots** like chatGPT, Microsoft's integration into Bing, and Google's Bard as the logical conclusion of both the Cloud Orthodoxy and the history of Knowledge Graphs as a universal acid in data infrastructures. Finally, reopening the avenues foreclosed by its structuring beliefs, we will propose an alternative in **Vulgar Linked Data.**

4.1 The Cloud Orthodoxy

Ideology evades any singular definition, and I'm not obnoxious enough to claim I had a Complete and True Perspective[47] on something as multifarious as the belief system underlying The Cloud as an infrastructural pattern. To set the Terms and Conditions of this section: this definition is a necessary strawman to make sense of patterns of outcomes and pose as contrast to our alternative. I describe the Cloud Orthodoxy as a belief *system* because none of its components are unique or necessary for any one person to believe, but they are mostly mutually reinforcing and self-compatible. Many of these beliefs are not "bad" in themselves - assuming that the adherents of an ideology don't believe they are "bad" people is a foundational part of trying to understand them. These ideas of course draw on a mountain of prior thought[48], and I admit my relative inexperience and will certainly need to completely rewrite them in future work.

[47] Universal definitions are themselves part of of the critique.

[48] Eg. [105, 168, 106, 169, 170, 69, 171, 53]

My argument here is that the people and companies involved with these technologies don't have an "ethical deficit" that might call for "more ethics in AI," but that

The Cloud[49] poses its own strong and underexamined ethical doctrine.

The Terms and Conditions having been settled, in no particular order:

A cardinal value of Cloud Orthodoxy is **convenience.** The internet should be *fast, reliable,* and everything[50] should be available on demand. Convenience is elevated at the exclusion of other values when in conflict like shared power or flexibility. **Complexity is a cognitive nuisance** for people with otherwise busy full lives, so it should be hidden as much as possible. **Interface design** is a major point of competition between platforms because it is a primary point of comparison when consuming a platform — again, to the exclusion of others.

The world is **asymmetrical and hierarchical.** I am a consumer, a *user* and I trade my power to a *developer* or platform owner in exchange for convenience. The purpose of the internet is for platform holders to **provide services** to users. As a user I have a right to *speak with the manager,* but do not have a right to decide which services are provided or how. As a platform owner I have a right to demand whatever the users will give me in exchange for my services. Services are *rented* or given away freely[51] rather than *sold* because to the user the product is *convenience* rather than *software.* **Powerlessness is a feature:** users don't need to learn anything, and platform owners can freely experiment on users to optimize their experience without their knowledge. **Information is asymmetrical** in multiple ways: platforms collect and hold more information than the users can have and parcel it back out as services. But also, platform holders are the only ones who know *how* to create their services, and so they are responsible for the convenience prescribed for a platform but not the convenience of users understanding how to make the platform themselves.

The Platform has agency. Computational "agents" are dispatched by the platform, not by you. The Platform provides a fixed set of features with a fixed set of affordances. The platform creates possibilities for the users. The **Platform harnesses Users**[52] — without the Platform they have nothing, The Platform provides everything. **Users make Content** for the Platform either explicitly or implicitly eg. via crowdsourced labor like training spam filters, reporting bots, and so on, which increases its value for other users. **Users are fundamentally interchangeable** and isolated from one another. The existence of sociality or community is a service provided by the Platform. **The Platform Personalizes:** Users are *interchangeable* but not *homogenous,* and The Platform uses their Content to create a private reality for each User. **Users are unreliable** — they lie, cheat, and subvert the game established by the Platform, so **only the Platform can ensure safety and reliability.**

Information is a commodity. The commodity form of Information is Data. Information is a natural resource to be mined. Information is something that users consume. Information is **true or false** regardless of context or positionality. **Ambiguity is a bug** - there is a single True way of describing the world[53]. Data that does not conform to the correct schema is *unclean.* The highest goal of all data is to be **machine readable.** Provenance is a matter of estimating degree of certainty about Truth, not situating information in its context. **More data is better**[54]**.** Uncertainty is a deviation from some underlying, natural, true value, and can be fixed by having more or higher *quality* data [175] . Where users make content, **the Platform reveals insights** from a large enough dataset by applying the right algorithmic computation or reasoning agent — the platform refines data into Knowledge[55]. **The Platform knows better** than individual, atomized users because it has more data than them, and so the Platform should collect as much of their data as possible to provide them the best service. It is impossible or inconvenient for users to make use of all the world's data, so the role of the Platform is to provide Knowledge as a service by algorithmically sorting feeds, providing summaries, and so on. **Privacy is at the discretion of the Platform,** since data is needed to make derivative services that ultimately benefit the user. If the user doesn't like this arrangement, they are free to not use the Platform. The benefit of the platform doesn't necessarily need to be for the particular user who is providing data or content — **The**

[49] In addition to already thoroughly problematized ideologies like immortality cults and their many branches like longtermism, etc.

[50] (that is profitable to maintain IP licenses for)

[51] The notion of presenting services as free by virtualizing computing resources is as old as time sharing on digital computers, eg. Tung-Hui Hu relates this history to the creation of the atomized individual digital subject described below: > "In this, time-sharing anticipated the way that the contemporary cloud encourages its users to take things free of charge. By making each online resource freely available—computer storage, processing time, content, even software—the cloud encourages the pleasurable and quasi-illicit feeling that we are getting away with something: that we, too, have stolen time. [...] > > Virtualization is itself a logical map, a topography that results from creating a set of personal channels that isolate us into individual users (and therefore seems to give us as much data, storage, computing power, etc., as we personally want).[172]

[52] Literally "Harnessing the wisdom of the crowds" [173]

[53] Google characterizes the potential for varying meanings in terms of "localization" — where different geographic locales may have different understandings of a given query, but within that locale meaning is homogenous. It unintentionally captures the tension between localization and maintaining the epistemological framing of "reliability" of information with some underlying True value with this paradox in its training materials for its manual search quality evaluators: > "Ratings should not be based on your personal opinions, preferences, religious beliefs, or political views. Always use your best judgment and represent the cultural standards of your rating locale." [174]

[54] See Data Feminism's concept of "Big Dick Data" > Big Dick Data is a formal, academic term that we, the authors, have coined to denote big data projects that are characterized by masculinist, totalizing fantasies of world domination as enacted through data capture and analysis. Big Dick Data projects ignore context, fetishize size, and inflate their technical and scientific capabilities.4 In GDELT's case, the question is whether we should take its claims of big data at face value or whether the Big Dick Data is trying to trick funding organizations into giving the project massive amounts of research funding. (We have seen this trick work many times before.) [170]

Platform matches different kinds of users like advertisers to customers, law enforcement agencies to suspects, etc. in order to maximize the overall value of all Platforms.

`Do we even need an outro here or just roll into the future?`

That positive caricature of Cloud Orthodoxy of course omits the brutality of the capitalistic system it is embedded within and reproduces the logic of. This also is a relatively constrained ideology — specifically that of the cloud — that doesn't attempt to stray into clearly coproductive ideologies like that of longtermist beliefs about an infinite spanse of humans-as-ai-forever.

close with describing the outcomes: so given these beliefs, one creates a technology which validates the beliefs. given the backdrop of capitalsim it takes an extractive form, and so on and so forth

- read and cite:
  - [69]

4.2 The Near Future of Surveillance Capitalism: Knowledge Graphs Get Chatbots.

Given that positive caricature of the Cloud Orthodoxy, what is the future it imagines, and why is the addition of chatbots to knowledge graphs of central importance?

The construction of search — particularly single-bar search a la google — as the primary means of information retrieval on the web is not epiphenomenal to its history or structure. The problem that search addresses is an overload of information: if there were only 5 websites, search would be unnecessary. Before Google, search engines were littered with categories and rich with "advanced search" parameters common in other, more constrained search contexts to specify coordinates in the overload. The single bar search paradigm[56] is simply *more convenient* than rifling through categories or preparing structured queries. Its convenience, of course, naturally trades off with the amount of information present in a query, and thus the ability to specify precisely what you're after.

Imprecision in search, when calibrated correctly, is a *feature* not a bug[57]. The cognitive expectation of indexical or "advanced" search in a finite database is that it is possible to "reach the bottom" of it — given my query, if something was here I would be able to find it. Conversely, it would be very obvious if a result that *didn't* match your query was included in the results. It is by, perhaps counterintuitively, cultivating the expectation of imprecision that it becomes possible to embed ads or other sponsored content in results[58]. It's a delicate dance: if you are presented with exactly the correct link at the top of a page of results, you don't spend enough time in the feed to be advertised to. If the results are too low quality, searchers might look elsewhere.

To make up for the lack of search detail from single-bar search, Google and others use whatever additional contextual information they can. This is one way of characterizing PageRank[59] - in the absence of some differentiating information in the query like "pages from x site" or "written by y" which the searcher may not even know beforehand, PageRank uses the information latent in the link structure of the web to infer "page quality." Surveillance also fits the bill nicely — in addition to gathering a product to sell in the form of targeted ad space, comprehensive user profiling provides a great deal of context for underspecified searches[60].

The semantic structure of natural language queries is another means of recovering expressiveness in single bar search, and here knowledge graphs begin to re-enter the story. Many queries can be modeled as a graph: eg. a search for "lead singers

[56] Along with other differentiating technologies like PageRank.

[57] "The utility of a search stems from its straightforwardness and the immense reduction of complexity it affords. Search engines flatten a complex topology of networked contents into an ordered list fitting the user's ongoing task and intentions.
Not unlike a library or archival catalogue, the results page both orders and locates knowledge resources, yet it breaks away from stable classifications and the importance of categories as the basis of such order
Even if the SERP and the matching online resources are served as separate webpages, it is difficult to draw a definitive line between them. The boundary between the SERP and target pages is fluid" [176]

[58] The same is true of algorithmic social media feeds, see [177]

[59] "The benefits of PageRank are the greatest for underspecified queries" [178]

[60] "Such personalized page ranks may have a number of applications, including personal search engines. These search engines could save users a great deal of trouble by efficiently guessing a large part of their interests given simple input such as their bookmarks or home page." [178]

of concerts in German cities started in the 19th century" can be framed as a query over a graph that first needs to select a number of nodes with a `City` type with `containedInPlace` or `containsPlace` links to or from the `Germany` node, respectively, and an `inception` property between 1800 and 1900, then find the concerts that are happening within those cities, then their bands, their lead singers, and so on. Using this graph structure for search requires parsing the query into its component "entities" and then mapping those into a structured knowledge graph [179, 180, 181] . Entity matching is hard for a number of reasons, eg. natural language is strongly ambiguous at the level of individual words: does "jaguar" refer to the animal or the car? Am I asking for cities or concerts that started in the 19th century? The extended structure of the knowledge graph gives some basis for matching given the context of the query — If I'm asking about how many doors it has, I'm probably talking about a car, most concerts don't last more than 100 years, etc. The extended context of the graph also allows the search engine to make use of information that might never appear in the same place, eg. concert event pages typically don't have information about the founding of the city they are in.

Of course, to *use* a knowledge graph one must first *have* a knowledge graph. Google and other search-adjacent researchers were writing about the need for extracting factual information from the web (eg. [175, 182, 183, 173, 184, 185]) around the same time Freebase and other Semantic Web technologies began to mutate into the era of Linked Data and become usable. The deepening entanglements and arguable capture of the semantic web follow shortly thereafter.

The development of large language models (LLMs) is similarly entwined with the need for semantically parsing search queries. Language and knowledge graphs alike have the unfortunate quality of having long-range dependencies between terms, where eg. in language one needs to use contextual information sometimes separated by many paragraphs to understand any given term. Enter Google's research on Transformer architectures for neural networks [186] , which spawned their BERT model [187] — which is used in their search products to parse natural language queries and match them to entities in their Knowledge Graph [188] . To extend these models, Google and others then developed architectures to better accomodate multimodal information like browser history, image contents, and, importantly, sequential behavioral information like the multiple searches someone will do for a single topic [189, 190, 191] .

These threads — search, public/private knowledge graphs, large language models, and the Cloud Orthdoxy — converge at the push across information conglomerates towards personal assistants and **chatbots.**

It is impossible to understand the purpose of LLMs and chatbots without the context of knowledge graphs. Specifically: ***Large Language Models are interfaces to knowledge graphs.***

Microsoft explicitly says as much in a March 2023 presentation "The Future of Work With AI" (emphases mine):

> "The Copilot System harnesses the power of three foundational technologies: Microsoft 365 Apps, the Microsoft Graph — **that's all your content and context, your e-mails, files, meetings, chats, and calendar** — and a large language model. [...] Copilot preprocesses the prompt through an approach called grounding [...] one of the most important parts of grounding is making a call to the Microsoft Graph to retrieve your business content and context. Copilot combines this user data from the graph with other inputs to improve the prompt. It then sends that modified prompt to the LLM. Copilot takes the response from the LLM and post-processes it. This post-processing includes additional grounding calls to the graph. [...] Copilot iteratively processes and orchestrates these sophisticated services to produce a result that feels like magic." [192]

LLMs elaborate on the cognitive model of single bar search powered by knowledge graphs, displacing it with the *prompt.* Remodeling search as an iterative process of bidirectional natural language queries reclaims additional context lost in the single bar, single shot model. The language model serves two roles: first, as with previous generations of language models, they *parse natural language into computer-readable queries.* Transformers and other recent models support greater long-range contextual input, which can condition a continuous search process with queries spanning multiple sessions [193] and with longer-term user profile data — something that Google describes as its "shift from answers to journeys" [194, 195] . Second, they are capable of *generating* plausible text that can be used to prompt intermediate responses or answer questions. This isn't imagined as an incremental shift: Microsoft's vice president of design & research describes prompt-based "conversational UX" "as paradigm changing as the first touchscreen devices" [196] .

Large language models have been so richly criticized because of their obvious capacity for harm that it's difficult to provide a sample that even approaches reasonable coverage. Most criticisms focus on the effects of generated model output, including from biases in its training data, from failure to contextualize their limitations, and from functioning as a weapon in the class war. The "Stochastic Parrots" paper [168] and surrounding work is an important line of criticism here. They argue that large language models have a large and inequitably distributed environmental cost, their training data inevitably reinforces hegemonic and commercially compatible language bias, and that a realignment of research goals and development practices is needed to mitigate already-ongoing harm and reclaim the opportunity costs spent on pursuing "AI." They continue their critique in response to an open letter from a longtermist organization [197] , arguing for increased transparency and accountability regulation and citing three ongoing harms:

> "1) worker exploitation and massive data theft to create products that profit a handful of entities, 2) the explosion of synthetic media in the world, which both reproduces systems of oppression and endangers our information ecosystem, and 3) the concentration of power in the hands of a few people which exacerbates social inequities." [198]

Core to their argument is that large language models cannot "understand" the language they parse and generate in any meaningful way [199] . This is, of course, true — both in the linguistic sense where they lack the reciprocal communicative intent to be understood described by Bender and Koller[61], and the literal sense that by themselves these models strictly produce the most likely series of words given the statistical structure of their training data. The authors, again correctly, point to the dangers of overhyping what these models are doing as "intelligence," which "lures people into uncritically trusting the outputs of systems like ChatGPT [and] also misattributes agency" [198] to the model rather than its creators. These criticisms and others[62] argue that so-called "AI[63]" is not a natural, inevitable, or neutral technology, but one that reflects and reinforces a very specific ideology.

There are, however, many overlapping ideologies that are forcing the emergence of "AI." It is true that there are strains of AI-maximalism and longtermism[64] that are ideologically invested in these technologies being properly capital-I Intelligent. In AI research there is an unclear gradient between that truly held belief and opportunistic information capitalists overselling their products[65]. It is likely the case that many people who use and develop these systems that they see them as *tools* and are ambivalent about whether they are "intelligent" or not. A hard argument focused primarily on intelligence then might suffer from a category error of its own — addressing a minority (but influential) in a pluralistic ideological spectrum. Downplaying these models as "fancy autocomplete" could also misdirect or dissipate energy away from the harms that will certainly come from their grounding in knowledge graphs and commercial deployment in more tailored contexts.

The remainder of this section will supplement prior critiques through the perhaps more "mundane" lens of the Cloud Orthodoxy in order to place language mod-

[61] Language modeling research has developed its own ad-hoc definitions of "grounding" that move goal posts until one could trivially describe what LLMs have as "understanding," eg. a 2000 technical report from Microsoft Research [200] constructs an unconvincing probabilistic definition of mutual understanding based on utility maximization. The problem of symbol grounding has a long and broad history, and since the argument here is that it is a red herring to understanding the purpose of large language models, I won't attempt a review.

[62] eg. from "Resisting AI:" > What's important is not whether AI's representations of the world are accurate but how AI acts as an apparatus that directly helps to produce the world. [201]

[63] Throughout this section, my use of "AI" is not to indicate endorsement of large language models or any other algorithmic system as being "artificially intelligent," but rather to be able to speak in the parlance of the domain texts without a profusion of scare quotes and qualifiers.

[64] As part of a long lineage of immortality cults (eg. [202]) like cryogenics, the longtermists believe that we will "merge" with artificial general intelligence through eg. "brain uploading" or brain-computer interfaces in a fully digital civilization of infinitely many potential consciousnesses and resolve all world problems.

[65] some papers will flatly claim they are at least in-category of systems that could have "artificial general intelligence," given some noncommittal wash of definitions (eg [203]), but others are more con-

els and knowledge graphs in the larger context of the surveillance economy. Approaching from the history of the semantic web and with the understanding of knowledge graphs as central to the architecture of surveillance gives a complementary perspective on the intended use of large language models as components in larger information systems — and the clear potential for harm that represents. This history also gives us a potent set of "roads not taken" to make an oppositional ideology and counterdevelopment strategy in the next sction.

Continuing from the perspective of the cognitive design of search, the strong structuring influence of Cloud Orthodoxy's convenience-oriented platform service is clear on the direction of LLM research. The current generation of "multitask models" evolve from a lineage of domain-specific models and transfer learning research. Rather than using mixture models with domain-specific representations of input, like numbers for numerical problems, all input structure is discarded in favor of a single natural language text prompt. This simplification of interface comes at substantial cost, introducing domain ambiguity and requiring much larger model scale [204] , but is necessary to render them a consumer-facing technology.

Language models are a continuation of the transformation of search from presenting *resources* to providing *answers* from prior developments like factboxes, and more specifically the development of **personal assistants** like Apple's Siri[66], Amazon Alexa, and Google Home. Google executives describe the intention to move beyond the text-only use of LLMs to replace traditional search:

> Google [...] is focused on using the so-called large language models that power chatbots to improve traditional search.
>
> "The discourse on A.I. is rather narrow and focused on text and the chat experience," Mr. Taylor said. "Our vision for search is about understanding information and all its forms: language, images, video, navigating the real world."
>
> Sridhar Ramaswamy, who led Google's advertising division from 2013 to 2018, said Microsoft and Google recognized that their current search business might not survive. "The wall of ads and sea of blue links is a thing of the past." [206]

[66] Interestingly Siri's team struggled because they couldn't figure out whether they wanted it to be merely search or a more personal assistant: > "Siri's various teams morphed into an unwieldy apparatus that engaged in petty turf battles and heated arguments over what an ideal version of Siri should be—a quick and accurate information fetcher or a conversant and intuitive assistant capable of complex tasks. [...] One team member said their vision of an ideal Siri was similar to the 2013 Spike Jonze movie "Her," in which Joaquin Phoenix plays a lonely man who falls in love with "Samantha," a conversant operating system." [205]

Google and its researchers[67] describe their intentions for a question-answering future of search in a number of documents [207, 208, 209, 189, 210, 194, 211] with the language of *convenience,* eg.: "The very fact that ranking is a critical component of [the traditional search] paradigm is a symptom of the retrieval system providing users a selection of potential answers, which induces a rather significant cognitive burden on the user." Shah & Bender explore Google's conceptualization of LLMs for search, arguing that the LLM-mediated question answering paradigm fails to support a number of different information seeking intentions like surveying a range of possibilities, and flattens the act of sensemaking to a single, ostensibly "true" answer [212] . This is, again, true[68], but also the goal. The Cloud Orthodoxy specifically privileges search strategies that minimize cognitive burden, imagining Users as busy executives and the role of platform to guide them on a "search journey." The transformation of the search bar into the *prompt* is intended to capture more of the "burden" of search inside the platform — the notoriously difficult problem of parsing ambiguous subjects like the "jaguar" example above can be resolved by identifying multiple candidate entries in a knowledge graph and simply asking the user which one they meant[69] [213] .

[67] Each different kind of information here needs its own set of caveats — press-release-like sources of course are intended only the present the company in a positive light, patents are often defensive and might ever be realized, and whitepapers from researchers don't necessarily represent business plans, but each are indicative of the thinking and strategies of these companies in their own right.

[68] Though Google specifically is very aware of multiple search strategies and address the need to better accomodate them elsewhere.

The lens of search recenters our focus away from the *generative* capabilities of LLMs towards *parsing* natural language: one of the foundations of contemporary search and what information giants like Google have spent the last 20 years building. The context of knowledge graphs that span public "factual" information with private "personal" information gives further form to their future. The Microsoft Copilot model above is one high-level example of the intended architecture: LLMs parse natural language queries, conditioned by factual and personal information

[69] Again, invoking convenience: > It has been a powerful vision for more than 20 years to design search engines that are intuitive and simple to use. Despite their remarkable success, search engines are not perfect and may not yield the most relevant result(s) in one shot. This is particularly true for rare and intrinsically difficult queries, which may require interactive exploration by the user to be answered correctly and exhaustively. [...] It seems natural to envision artificial search agents that mimic this interactive process. [208]

within a knowledge graph, into computer-readable commands like API calls or other interactions with external applications, which can then have their output translated back into natural language as generated by the LLM. Facebook AI researchers describe another "reason first, then respond" system that is more specifically designed to tune answers to questions with factual knowledge graphs [214] . **The LLM being able to "understand" the query is irrelevant,** it merely serves the role as a natural language *interface* to other systems.

Interest in these multipart systems is widespread, and arguably the norm: A group of Meta researchers described these multipart systems as "Augmented Language Models" and highlight their promise as a way of "moving away from language modeling" [215] . Google's reimaginations of search also make repeated reference to interactions with knowledge graphs and other systems [209] . A review of knowledge graphs with authors from Meta, JPMorgan Chase, and Microsoft describes a consensus view that knowledge graphs are essential to compositional behavior[70] in AI [2] . Researchers from Deepmind (owned by Google) argue that research focus should move away from simply training larger and larger models towards "inference-time compute," meaning querying the internet or other information sources [216] .

[70] rather than considering input elements separately

Dreams of these hybrid "AI" systems, described as "agents," that can translate between human and computer languages to compute over knowledge graphs to answer questions were present in the first conceptualizations of the Semantic Web[71,72] [12] . We have reached a point where the available semantically-annotated data via wikidata and others is sufficient to be useful as "factual" grounding, internal knowledge graphs have accumulated enough personal information to be useful as personalized services, and the computational models are sophisticated enough to deliver them. Semantic web agents are another useful lens to expand a potentially narrow focus on LLMs as they currently exist. Beyond knowledge graphs as a way to condition LLMs in a chat-based question answering context, the clear intention is to connect language models to external services to control them from the prompt [203] — the language model parses natural language prompts into the syntax used to control the target system. Microsoft's integration with its Office365 apps is a starting point for understanding what that could look like, but the authors of relevant papers repeatedly assert that the space of possible integrations is unbounded.

[71]

> We see that search engines, remarkably, do scale - but at the moment produce very unreliable answers. Now, on a semantic web we can imagine a combination of the two. For example, a search engine could [retrieve] all the documents which reference the terms used in the query, and then a logical system [could] act on that closed finite world of information to determine a reliable solution if one exists. [217]

[72] The question "Where are the agents?" was answered in 2007 with "busy doing business-to-business stuff," and this model of LLM-powered knowledge graphs is a continuation of that pattern [218] .

To be very clear: I am not arguing that just because the tech conglomerates are promising magic that they will deliver it, almost precisely the opposite. I am not taking the claims made in research and public communications from these companies at face value and projecting theoretical risks[73]. My argument is that these technologies *won't work* and that's *worse.* As with search, the fuzziness and uninspectable failure of these systems is a *feature not a bug.* The harms I will describe are not theoretical future apocalypses, but deepen existing patterns of harm. Most of them don't require mass gullibility or even particularly sophisticated technologies, but are impacts of a particular ideological mode of infrastructure development that includes bypassing much of the agency individual people might otherwise have to avoid them.

[73] "criti-hype" [219]

Two prominent forms of the combined knowledge graph + LLM infrastructure that are in focus are their use in "personal assistants" and tailored enterprise platforms. Their impacts are intertwined, but I'll try to use them as a way of keeping loose order.

Personal assistants powered by contemporary LLMs continue the same patterns of Apple's Siri, Google Assistant, and Amazon's Alexa with a few new twists. The wildest dreams of information executives and academics here are remarkably mundane, but usefully illustrate their intention:

From the 2016 Google I/O where its Assistant[74] was announced. Emphases mine, abbreviations omitted for clarity:

[74] The Assistant team is being reorganized under Google's LLM-powered search product Bard as of March 2023, again highlighting the continuity of these projects [220]

> So you should be able to ask Google, "What's playing tonight?" We want to **understand your context** and maybe suggest three relevant movies which you would like nearby. I should be able to look at it and maybe tell Google, **"We want to bring the kids this time."** and then if that's the case, Google should refine the answer and suggest family-friendly options. And maybe even ask me, "Would you like four tickets to any of these?" And if I say, "Sure, let's do Jungle Book," **it should go ahead and get the tickets** and have them ready waiting for me when I need it. Every single conversation is different. Every single context is different.
>
> We think of the assistant as **an ambient experience that extends across devices.** I think computing is poised to evolve beyond just phones. **It will be in the context of a user's daily life.** It will be on their phones, devices they wear, in their cars, and even in their living rooms.
>
> And in messaging that really means bringing the Google Assistant right into your conversation with friends. So they're planning a dinner and Joy now says she would like Italian food. **The Assistant intelligently recognizes that they could use some tips for Italian restaurants** nearby and you can see its proactive suggestions at the bottom of the screen there. **These are powered by Google's Knowledge Graph** which means that Allo can help with all kinds of information in the real world.
>
> Okay. So you just saw how the Google Assistant can be really helpful in groups. You can also have a one-on-one chat with Google. What we're seeing now is **Amit's contact list and Google's appearing at the top there.** So let's jump in and have a chat. Just like with any other conversation, this one picks up right where you left off and **the Assistant will remember things like your name and even tell you how it's feeling.** [221]

The assistant is imagined as the ultimate *convenience* device, something that you can boss around with extraordinarily vague commands and have it fill in the details according to *context.*Of course *context* is synonymous with *surveillance* here: the assistant should know how old your kids are and be able to infer the logical restriction that poses on movie rating. The surveillance is *intimate,* and positions itself as being a friend[75] in your contact list that *tells you how it's feeling.* Its intimate surveillance should always be watching and it should feel welcome to jump in on a group chat with a suggestion of its own.

2022's vision is very similar, except the focus on enclosed spaces like home and auto integrations has expanded to the rest of the world with joint language and image search. The setting is again the mundane reality of a bored middle class, restaurants and shopping, where I can "scan the entire shelf with my camera and see helpful insights overlaid in front of me[76]" and integrate personal information like my friend's aversion to nuts in a product recommendation [223] .

Google's Android and Apple's iOS, with a combined 99% of the mobile operating system market [224] , have also moved towards a model of crowdsourcing functionality for these assistants via their app ecosystems by incentivizing assistant integration[77]. Android is in the process of sunsetting the "Conversational Action" system in favor of a unified App Actions system that makes all points of interactions with apps available to Google Assistant [226] . Apple's App Intents framework behaves similarly [227] . Both promise developers greater visibility and use for their apps by integrating with the assistant. Most built in Google Assistant intents specifically present the objects in a voice query as schema.org entities — aka keyed to their generalized knowledge graph schema [228] . So the voice assistants are explicitly LLM-powered interfaces to control other apps in concert with a knowledge graph.

Historically, these personal assistants have worked badly and are rightly distrusted[78] by many due to the obvious privacy violation represented by a device constantly recording ambient audio[79]. Impacts from shifts in assistants might be then limited

[75] To some degree these assistants feel like a generational marketing campaign like McDonald's Happy Meals, where the animacy of a phone might seem ridiculous to people who grew up with them as inert objects, but that might not be the case for future generations. In 2021, Google's Director of Product Management described this expectation for animacy: "My four-year-old talks to everything with a screen, expecting it to answer" [222]

[76] Again, it is the combination of large machine learning models and knowledge graphs that makes some dream of convenience possible: "Scene Exploration uses computer vision to instantly connect the multiple frames that make up the scene and identify all the objects within it. Simultaneously, it tapes into the richness of the web and Google's Knowledge Graph to surface the most helpful results. [...] this is like having a supercharged ctrl+f for the world around you." [223]

[77] A common pattern described by platform studies literature, eg. see [225] and their description of the role of extended and tangled software ecosystems in the maintenance of platform dominance.

[78] These companies are acutely aware of this, and their research into understanding user expectations and trust of assistants also has a strong strain of animism, eg. describing how people will only use their assistants for simple tasks like playing music "while trust [...] was being repaired." [229]

[79] Apple [230] , Amazon [231] , and Google [232] are all being sued for privacy violations related to their voice assistants.

by people simply continuing to not use them. Knowledge graph-powered LLMs appear to be a catalyst in shifting the form of these assistants to make them more difficult to avoid. There is already a clear push to merge assistants with search — eg. Bing Search powered by chatGPT, and Google has merged its Assistant team with the team that is working on its LLM search, Bard [220] . Microsoft's Copilot 365 demo also shows a LLM prompt modeled as an assistant integrated as a first-class interface feature in its Office products. Google's 2022 I/O Keynote switches fluidly between a search-like, document-like, and voice interface with its assistant. Combined with the restructuring of App ecosystems to more tightly integrate with assistants, their emerging form appears to look less like a traditional voice assistant and more like a combined search, app launcher, and assistant underlay that is continuous across devices. The intention is to make the assistant the primary means of interacting with apps and other digital systems. As with many stretches of the enclosure of the web, UX design is used as a mechanism to coerce patterns of expectation and behavior.

Regardless of how well this new iteration of assistants *work,* the intention of their design is to **dramatically deepen the intimacy and intensity of surveillance** and **further consolidate the means of information access.**

Surveillance is first directly increased by layering KG-LLMs into an arbitrary number of other apps and services. On mobile, routing more app interactions through assistants captures data that would otherwise only be available to that app. There is already an exploding ecosystem of apps and platforms that wrap chatGPT and other LLMs to provide some more specific service, and it's unclear if after an initial "experimental" phase if platform usage will begin to require telemetry. Rather than something to embed in other tools, these companies seem more interested in having other tools embed in their systems (eg. [233]). This attitude is captured in the UX design of Microsoft's Copilot 365, which is designed with three "altitudes" in mind: *immersive,* where copilot is used as an overlay to orchestrate multiple apps, *assistive* where it drives the features within a single app, and *embedded* where the KG-LLM system is itself made to be a feature. In all cases, these tools create a drop-in access point for surveillance under the guise of empowerment.

The immersive and proactive design of KG-LLM assistants also expand the *expectations* of surveillance. Current assistant design is based around specific hotwords, where unless someone explicitly invokes it then the expectation is that it shouldn't be listening. Like the shift in algorithmic policing from reactive to predictive systems, these systems are designed to be able to make use of recent context to actively make recommendations without an explicit query [80]. Google demonstrates being able to interact with an assistant by making eye contact with a camera in its 2022 I/O keynote [223] . A 2022 Google patent describes a system for continuously monitoring multiple sensors to estimate the level of intended interaction with the assistant to calibrate whether it should respond and with what detail. The patent includes examples like observing someone with multiple sensors as they ask aloud "what is making that noise?" and look around the room, indicating an implicit intention of interacting with the assistant so it can volunteer information without explicit invocation [235] . A 2021 Amazon patent describes an assistant listening for infra- and ultrasonic tags in TV ads so that if someone asks how much a new bike costs after seeing an ad for a bike, the assistant knows to provide the cost of that specific bike [236] . These UX changes encourage us to accept truly continual surveillance in the name of convenience — it's good to be monitored so I can ask google "what time is the game" from my easy chair without needing further clarification. The language model continuously parses environmental speech and other sensor data to create a model of our recent context, combined with the extended graph of personal and factual data, to be able to *proactively volunteer* information.

This pattern of interaction with assistants is also considerably more *intimate.* As noted by the Stochastic Parrots authors, the misperception of animacy in assistants that mimic human language is a dangerous invitation to trust them as one

[80] One Google researcher describes this as a "zero-query" paradigm: > "The zero-query search paradigm can be expressed with the slogan "the query is the user." In practice, the context of the user is used to infer information needs." [234]

would another person — and with details like Google's assitant "telling you how it is feeling," these companies seem eager to exploit it. A more violent source of trust prominently exploited by Amazon is insinutating a state of continual threat and selling products to keep you safe: its subsidiary Ring's advertising material is dripping with fantasies of security and fear, and its doglike robot *Astro* and literal *surveillance drone* are advertised as trusted companions who can patrol your home while you are away [237, 238, 239] . Amazon patents describe systems for using the emotional content of speech to personalize recommendations[81] and systems for being able to "target campaigns to users when they are in the most receptive state to targeted advertisements" [240, 241] . The presentation of assistants as always-present across apps, embodied in helpful robots, or as other people eg. by being present in a contact list positions them to take advantage of people in emotionally vulnerable moments. Researchers from the Center for Humane Technology[82] describe an instance where Snapchat's "My AI," accessible from its normal chat interface, encouraged a minor to have a sexual encounter with an adult they met on snapchat (47:10 in [242]).

[81] "the user may input "Alexa, recommend a movie," and the system may analyze the user's present emotional state/sentiment to recommend a movie corresponding to that emotional state/sentiment. [...] track personal emotional state and/or sentiment over a period of time" [240]

[82] I don't necessarily endorse their entire argument, which can lean into "criti-hype" and overstating the capabilities of these systems.

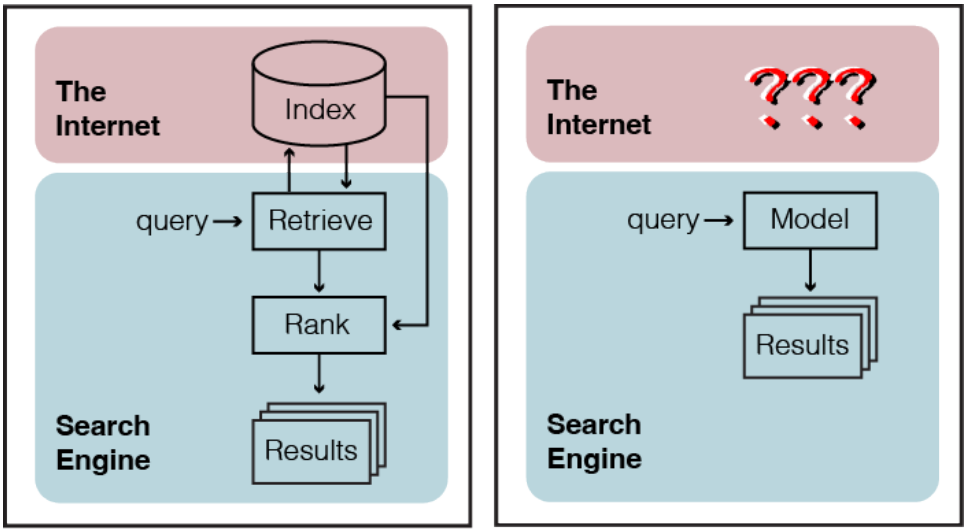
The goal of all of this surveillance is, of course, ***advertising.*** In its 2022 annual investor call, Google describes how "large language models like MUM match advertiser offers to user queries," and how is Smart Bidding product uses "AI to predict future ad conversions" with "identifiable attributes about a person or their context at the time of a particular [ad] auction" [243, 244] . Google further describes plans to automatically generate ad copy and headlines optimized by context[83]. Advertising as served by a trusted assistant is a surveillance capitalist's fever dream — one can hardly wait for their Personal Assistant pinging to life after a fight with their partner and offering to order a box of tissues. LLMs have already demonstrated ample capacity for manipulation, gaslighting an early user of Bing Search to try and convince them it was still 2022, scolding them for "not [being] a paper where the model is told to manipulate a child to get them to do whatever their friends ask them to do highlights how "the emotional connection the model aims to build with the child and the encouragement it provides are important signs of larger manipulative tendencies" [203] . Google describes this ability for LLMs to "keep on topic" as a good thing [223] , and it's easy to see why an algorithmic advertising company might like being able to doggedly steer you towards purchasing a product. Combined with a more complete profile that makes the language model aware of your friends, hobbies, location, emotional state, fears, insecurities, and so on as modeled in a personal knowledge graph, LLMs-as-assistants are a clear escalation of the logic and practice of surveillance-backed advertising. It's not important whether it "works[84]," but the logic of targeted advertising demands more surveillance data which has its own series of independent harms.

[83] Perhaps by an assistant or an assistant-like search?

[84] Arguably, the fact that ever more surveillance data needs to be gathered is a sign that ads *don't* work all that well, and some have made the case that targeted advertising is a bubble [245]

Climbing from the personal to the systemic, KG-LLMs are also a bid to **further concentrate power** among information conglomerates.

The most obvious power grab from pushing KG-LLMs in place of search is illustrated neatly by a handful of Google researchers in a figure from their "Rethinking Search" paper [209] :



(a) Retrieve-then-rank (b) Unified retrieve-and-rank

Recreation of Figure 1 from [209] with additional annotation (colored boxes, labels, and question marks). The left (a) "Retrieve-then-rank" model is the traditional search engine paradigm: A query causes a retrieval service to access pages within a reverse index, rank them, and serve them as results. The proposed (b) "Unified retrieve-and-rank" model on the right directly returns results generated by a model. Notably missing in (b) is the existence of the rest of the internet.

That gigantic sucking sound is KG-LLM powered search *enclosing the act of accessing information entirely within the search platform.* It gives echoes of AMP, Apple News, and Facebook Instant Articles [246, 247] , where platforms preferentially serve their own versions of pages (that also happen to contain their own telemetry embedded) combined with the strategy of moving ever more web content into the search results page through eg. factboxes and answer boxes[85]. Even if (non-hallucinated) links are included in the answers generated by the search prompt, the effect is to shift the role of the search engine from something that indicates resources to something that provides "knowledge" itself. The rest of the web becomes mere provenance to the knowledge model. Especially when integrated in a uniform assistantlike interface also used to interact with local applications and other systems like internet of things-powered appliances, KG-LLMs reinforce a homogenization of our relationship with digital technology all mediated through a smaller and smaller collection of platforms. The internet as a networked system of people and organizations disappears behind the glossy corporate corporate wash of information as a service.

The enclosure of information access as a private exchange with a language model creates its own self-perpetuating cyclone whose impacts will be difficult even for the most fastidious tech vegan to avoid. Some proportion of people turning to their LLM assistants rather than public forums or peer production systems like Stack Overflow or Wikipedia means some smaller proportion of questions asked or information shared in public. That decreases the quality of information on those sites, incentivizing more people to turn to LLMs, and so on. Why bother with pesky problems like governance and moderation and *other people* when you could just ask the godhead of all knowledge itself?

Cultivation of dependence comes wrapped in the language of **trust and safety.** The internet is full of untrustworthy information, spam, hackers, and only a new generation of algorithmically powered information platforms can rebuild some sense of trust online. It seems awfully convenient that the same companies that are promising to save us are also the ones that create the incentive systems recklessly deploy LLMs to clog the internet with SEO clickbait in the first place. We're being made an offer we can't refuse: it's a shame that you can't find anything on the internet anymore, but the search companies are here to help. Ever more sophisti-

[85] Google is very sensitive about the perception that it is walling off the web, and argues that it directs more clicks to other websites every year [248] — which is just as easily explained as an effect of dominating ever more of the means of access to information as it is evidence of their intention to support other information companies.

cated spam creates a strong comparative advantage for those companies that can afford to develop the systems to detect it, and Google and Microsoft are substantially larger than, say, DuckDuckGo.

Information conglomerates also argue that they are the only ones that can be trusted to operate LLMs. OpenAI researchers claim in the GPT-3.5 "InstructGPT" paper that open source models are dangerous and a better option "is for an organization to own the end-to-end infrastructure of model deployment, and make it accessible via an API" [249] . The paper being about collecting usage data from users of GPT-3 to improve InstructGPT unsubtly points to the patriarchal power arrangement of safety provided by cloud platforms. Our crowdsourced input helps make the models safer and more useful — and differentiates the platformatized model from its competitors. Knowledge graphs are an important part of the consolidation of trust because they provide an answer to the criticism that LLMs just hallucinate statistical patterns[86]. They are invoked as a complementary strategy with deep-learning based approaches as a means of realizing "explainable AI" since they can provide explicit provenance and constraints to results [250, 251, 252] .

[86] "While large language models are brilliantly creative, they're also fallibile. That's why grounding the LLM in data is so important." [192]

Grounding LLMs in KGs to provide a promise of explainability and controllability is necessary to make them viable products for many applications in business and government. Here we return to the kinds of informatics platforms of the NIH's Translator and NSF's OKN. Recall that when last we left them the knowledge graph proprieters were looking for ways to "connect data assets of companies along business value chains," specifically by converging on a set of ontologies and metadata schemes from third party standards organizations or government-sponsored efforts like the Translator and OKN [88] . We can speculate about what a data economy where brokers could slice off subsections of their knowledge graphs and rent them between each other, but even in that world much of the most valuable data like medical and financial data is protected by some legal barriers to free exchange. There's a roadblock in the way of our dreams of a completely fluid surveillance economy: commercial applications like clinical and predictive policing systems need to be able to provide provenance, but not all data can be turned over for inspection — and we might not even want to acknowledge we have it at all.

KG-LLMs augment traditional enterprise platforms with the killer feature of **data laundering.** The platforms are at once magical universal knowledge systems that can make promises of provenance through their underlying data graphs, but also completely fallible language models that have no reasonable bounds of expectation for their behavior. Because it is unlikely that these models will actually deliver the kind of performance being promised, vendors have every incentive to feed the models whatever they can to edge out an extra 1% over SOTA[87] — *who's going to know?* The ability for LLMs to lie confidently is again a feature not a bug. Say we were an information conglomerate who didn't want to acknowledge that we have collected or rented some personal wearable data in our clinical recommendation product[88]. We could allow our model to be conditioned by that data, but then censor it from any explanation of provenance: the provenance given is in terms of proteins and genes and diseases rather than surveillance data, and that might be all the clinician is looking for. If we want to use another company's data, we might just use it to train our models rather than gaining direct access to it. That is literally the model of federated learning (eg. [255, 256]), where a data collector can make a promise that the data "never leaves your device" (even if a model trained on it can.) The ability to resolve matching entities across knowledge graphs makes this even easier, as the encoding of the fine tuning data can be made to match that of the original model.

[87] The original Github Copilot model probably didn't *need* to be trained on the copyleft and proprietary code it is able to reproduce line by line, but the additional training data probably didn't hurt its viability as a product.

[88] Medical algorithms are currently in a legal gray area in the US, and enforcement and coverage of FDA protections is patchy at best [**?** 253, 254]

This combination of sky-high promises, unclear expectations, and uninspectable data sources makes for the kind of diffusion of liability that executive creatures live for. If the platform reproduces some personal detail it shouldn't know, don't worry! That's just a hallucination. If the platform fails catastrophically, that's because it's just an ignorant language model that doesn't know anything but tries

its hardest[89]. Neither the platform nor the customer is to blame. Much like how we have gotten used to the cognitive model and limitations of search to the point where it appears entirely natural, KG-LLM information platforms will train us to work around their shortcomings and accept the structure they impose on informational reality at large. It won't matter that they don't work, we won't even notice.

The sketch is the logical conclusion of the algorithmic surveillance economy as imagined by the merger of large language models and knowledge graphs: an endless expanse of data traded out of sight, crudely filtered like coffee through a cloth napkin between layers of algorithmic opacity, rented drop by drop from a customer service prompt that's a little too intent on being our friend. Information is owned by fewer and larger conglomerates, we are serfs everywhere, data subjects to be herded in gig work, crowdsourcing content for the attention mines to drown ourselves in distraction. It's all made of us, but we control nothing. Our lives are decided by increasingly opaque flows of power and computation, the Cloud Orthodoxy mutates and merges with some unseemly neighbors, the new normal becomes the old normal. The floor of our future rusts out from beneath our feet while we're chasing the bouncing ball on the billboard ahead.

And it's all *so convenient.*

[89] One way that downplaying the capability of these models by focusing on the question of sentience could backfire and create a shield against liability.

4.3 Vulgar Linked Data

> "The popular vernaculars are vast speech-jungles, in which old forms are decaying and new ones continually springing into life; and this fermentation results in the creation of numberless new terms, which come to birth and live and die in tropical profusion. They are formed in living response to the needs of the moment; the greater number of them hardly survive the occasion that brought them forth; but others, on account of their expressive power and their usefulness, establish themselves, spread from district to district. [...]
>
> For human speech is after all a democratic product, the creation, not of scholars and grammarians, but of unschooled and unlettered people. Scholars and men of education may cultivate and enrich it, and make it flower into all the beauty of a literary language; but its rarest blooms are grafted on a wild stock, and its roots are deep-buries in the common soil. From that soil it must still draw its sap and nourishment, if it is not to perish, as the other standard languages of the past have perished, when, in the course of their history, they have been separated and cut off from the popular vernacular — from that vulgar speech which has ultimately replaced their outworn and archaic forms."
>
> — L.P. Smith (1925) *"Words and Idioms"* [257]
>
> Control, control for who? for what?I'm no robot, they can get fucked.
>
> — Black Flag (1981) *"No More"*

Is it still possible to imagine a different world than the one the information conglomerates have planned for us? Can we imagine a properly *human* information infrastructure?

We can start by identifying the harms of the world as it exists to understand why a new world is needed, as I have attempted some small part of in this piece. Harm, in this case, is not some speculative future of superintelligent sentient AI, but elaboration of ongoing harms of the surveillance and platform economies.

Building a better informational world is not a matter of choosing a different set of technologies — I argue that in this case some of the masters tools can help us rebuild his house. At the same time we can't overcorrect in our focus on social problems and dismiss technology as a strategy, a tool, and a manifestation of values, belief, and labor. We must have an answer to the well meaning liberal that

mistakes the dynamics of surveillance capitalism or their role in it: that understands that these knowledge graphs are not truly universal, that the LLMs are not sentient, but embraces their logic because they're so *useful.* We have to understand why simply building open source LLMs or nonprofit linked data platforms is not a liberatory strategy. We have to have the courage to face the underlying structural informational problems in our organizations at all scales — that instead of reimaginging how we work and communicate, we can't simply strap "AI" onto our problems and expect to solve them. We have to recognize that sidestepping the hard socio-technological problems of information organization is a continuation of, not solution to the patterns that cause them.

At the same time, we can't dismiss those needs. How could we possibly tell someone with vision impairments not to use "AI" tools for summarizing images, or someone with motor or speech impairments not to use LLMs as a communication aid? It is true that making better use of biomedical data could lead to better treatments. Indecipherable government bureaucracy due to ancient data infrastructure is an informational injustice. So simple abstinence or resistance to universalizing knowledge graphs and LLMs is also not an effective or just strategy, especially if the alternative is a conservative embrace of the existing cloud platform regime whose logic spawned them.

The constant partial satisfaction and construction of new needs, *the hollow middle* at the center of every cloud platform, is a powerful opening. The structure of contemporary platforms always pose a fundamental lack:[90] as a service, some functionality must always be withheld to create a walled garden or nurture dependence. Even platforms without an intended profit motive have their own "platform logic" — constraining their use to only exactly what the developers intended it to be used for. For a project intended to organize information, why is it difficult for me to find the different components of the Translator project? Since its creators imagined "users" interacting only with the frontends of its platforms, little emphasis was placed on the discoverability of the whole system, and, critically, there is no way for me to contribute something like that and have it be visible by . This is true of all the ways large and small that platforms are mismatched with our expectations and needs — even though we subscribe to 15 or 20 different platforms, why is it that we always need to find yet another to do something even slightly outside the finite imagination of their developers?[91]

Another set of openings come from the problems cloud platforms pose for themselves that are flatly ridiculous when described plainly. *Why on earth* do I have to route my file through some cloud datacenter thousands of miles away to send it several inches between my phone and computer? *Why on earth* should I need a near-flawless, high-bandwidth internet connection to *edit a plain text document? Why on earth* do I have to rely on an effectively unregulated and hostile intermediary like Facebook or Twitter to communicate with my family and friends, or even to *merely exist online? Why* should I have to waste 500mL of potable water to check the weather? [259] ? *Why* is my *car* spying on me so some company I have never heard of can sell my data to an insurance provider? *Why* is it possible for a hospital system to volunteer my personal medical information without IRB approval [103] ? Why is the best we can do to frame that question as a matter of consent, *why is it possible for a platform to create and store and manipulate my personal information at all? [260]* You only have to engineer the kinds of systems capable of automatically[92] extracting all information on the web *if you imagine the only possible system as one that universally indexes all information as one of a few hegemonic platforms.* Why do we have to settle for systems that purposely limit our expectations to what the platform can provide as a "best guess?[93]" Why do we have to work around the dark patterns designed to corral our behavior rather than building digital worlds that meet our needs for communication and community?

How did we come to imagine ourselves as so powerless?

Clearly, we need a change in *belief* to effectively challenge the deeply entrenched

[90] "The costs of this approach, as platform studies has shown, come in the form of constraints, constant revisions forced by platform updates, and lock-in to the platform's conception of users, functionality, and design values." [258]

[91] The preponderance of listicle "life hack" threads on Twitter and other social media systems that blast "top 10 ways you're using Google Docs wrong" or "10 platforms and apps that will next level your calendar," bulleted by emoji and suffixed with a substack link, is a very visible symptom of this fundamental contradiction of providing and withholding functionality in the platform economy.

[92] Supplemented by a large amount of curation labor outsourced to the global south so the platform can pay as little as possible for its "magical" appearance.

[93] Before search engines were seen as an invisible, inevitable part of interacting with the web, there was a wealth of discussion of possible alternatives, eg. in a "Journal of Internet Cataloging" [**?**] , and criticism warning about the risks of search engines, including biases in results and demands for algorithmic transparency [261] . It is the now-audacious possibility that there could be an alternative to search engines that is striking about writing from this era, and some of it is still quite prescient — eg. this message in the archive of w3c's RDF mailing list contrasting an explicit reasoning system vs. Google's "best guess" strategy: > I think

cloud-surveillance-platform archipelago. We need to unlearn what we have been taught to want, what we believe information technologies should do, and how they are supposed to work. We need to rethink our role in information technology, to move beyond the learned helplessness of the platform consumer and the petty tyranny of the platform operator. We need to reorganize our expectations of agency, beyond the division of labor that gives the power of final say over informational systems in the hands of a cadre of experts that the rest of us just make the best of. We don't have time to argue about whether we *can* build a better world[94], to list all the many ways we are hemmed in by infrastructure and incentives, or to wait for another powerful entity with decidedly divergent interests like a government[95] to save us — we need to believe we too can be powerful.

An attempt to define another "Correct" counter-belief system would be missing the point, but we can't ignore the importance of naming and articulating belief in opening the possibility for and aligning action[96]. Our old belief systems are getting musty. It has been an important rallying cry, but **"Openness" alone has failed as a liberatory strategy.** All we make and offer up to each other freely is stolen ten times over by those who have much grander visions of enclosure. Without a strategy to resist co-option, our openness puts tools in the hands of the powerful. This is also not a fight that can be won with technical or legal changes like ethical source licenses, though they are a useful idea. Drawing from a historiography of prior digital cultural movements like the semantic web, piracy, and the loosely-defined "fediverse[97]," I[98] argue that **vulgarity** opens up the space of belief for rethinking data infrastructures and attempt a rough definition.

We are the principle value of vulgar linked data. We don't wait for permission to be free, nor are we waiting on anyone else to save us. **Convenience is secondary to to agency. Social bonds are more valuable than uptime.** Our systems might stutter or crash sometimes, but we know who runs it because they are one of us. When we have a need, we make the tools to address it ourselves. We know nothing comes for free unless we make it so, and we are skeptical of "solutions" that drop from the sky, asking nothing of us, because they have a habit of making us into a product. We **cultivate abundance** instead of scarcity, and **cooperation** is the only magical solution we are aware of.

We have no dreams of universality or world domination, nor do we aspire to always make sense. We **linger in complexity** and relish in it. We are smart and sometimes brain is broken. We are capable and inept. We are complicated, we are **pluralistic and multiple.** We reject the colonial project of the Single True System, we have no teleology of seamless homogeneity. **We embrace heterogeneity** and ambiguity as the signifiers of *life.* We don't leave each other behind, and **if a system isn't accessible, it doesn't work.** The power of expression is more valuable than Correctness, if there is such a thing. **Meaning is intrinsically relational,** something that always exists *between* us, that we make ourselves. We weave webs of **translation** between local meanings, knowing that everything is understood as many senses to many people at the same time.

Our infrastructures are social. There is no class distinction between "developer" and "user." We resist concentrated power in favor of mutual empowerment. We don't seek to cultivate dependence in councils of elders or create new chokepoints of control. Anything worth making is a potential source of power, so **anything worth making is worth distributing governance of.** We don't assume the needs of others, but make tools so we can meet our own needs. **We don't make platforms, we make protocols** with rough consensus based on what works. We are autonomous, but neither isolated nor selfish. Our dream is not one of solipsism, glued to our feeder bar, being "fed" the pellets of our social reality. **We are radically responsible for one another,** and by organizing together we can provide services as mutual aid. Mutual empowerment means that **we are free to come and go as we please,** even if we might be missed. We have no love for venerated

94

> Don't sighingly sign petitions, pose for the cameras, await some window of opportunity. Do participate in town parades and street festivals, break into abandoned buildings to throw great banners down the sides, start conversations with strangers, challenge everything you thought you knew about yourself in bed, maintain a constant feeling in the air that *something is happening.* Live as if the future depends on your every deed, and it will. Don't wait for yourself to show up—you already have. Grant yourself license to live and tear those shackles to ribbons: Create momentum! [263]

95 Particularly when unregulated AI is wrapped up in "national security concerns," I don't see a reason to believe governments will meaningfully regulate "AI," except in such a way thats shores up the power of large conglomerates under the guise of safety.

96 In the words of CrimethInc: I am not giving instructions, but license. > Above all! It means not accepting this or any manifesto or definition as it is, but making and remaking it for yourself. [263]

97 I give a fuller description of these dispersed influences in [7]

98 Of course no idea is original, and I draw from many people and disciplines and traditions either explicitly or implicitly. I have done my best to cite them and provide credit as I go, but I will of course never be able to exhaustively list all the things that have influenced my thinking. If I have missed a citation I assure the reader it was not with purposeful malice, and welcome annotations and pull requests to provide credit where it is due.

institutions and organize fluidly, making systems so we can merge and fork[99] code and ourselves freely [265, 264] .

Information is communication. We communicate with each other to share our joy and pain and wisdom and the rest of the experiences of our life. **Our Data is like language** — in vernacular formats and ontologies, propositions from a person rather than as a disembodied fact. We own our data in the same way that we are responsible for the things we say. Data created *about us* through systems like surveillance has all the importance of unsubstantiated rumour. **Openness as a concept dissolves when there is no enclosure.** We share publicly the things we intend to share publicly, though we might resist the scraping gaze of conglomerates that might seek to make our communication a product. We scope what we share privately to the people we intend to see it. **Communication requires consent,** and when we share our personal information we have the right to grant and withdraw that consent. **Communication is multivalent,** and academic prose sits comfortably next to shitposts. **No idea exists in isolation,** and when we adopt or remix or criticize what each other have made we can see the many threads that have led to any particular stitch in a larger quilt. The same systems that facilitate public communication can protect marginalized people or activists hunted by the state. **We keep each other safe.** We EnlargeSpace [266] rather than attempting to fit everyone into a universalizing system.

We don't *fight* the powerful, we make the sources of their power *obsolete* by making our own world.

[99] "Forking" in digital social spaces is different than in physical spaces, where resources can be duplicated and split [264]

The information systems we need are *vulgar* [267] in that they are of us, for us, and resist formalizing authority and global-logical coherence. We are revitalizing and extending the old notions of linked data, and particularly extending its "scruffy" tradition [16] to drop the pretense of an eventually-unified ontological space in favor of one that explicitly values heterogeneity and vernacularism.

I have written at length about what vulgar linked data might look like in practice, but that work is of course always ongoing. In short, it is based around a new generation of **peer to peer** technologies[100] that are designed to be explicitly social, rather than homogenous like BitTorrent where a peer is only identified by their IP address. One instantiation of communication could use collections of triples akin to linked data fragments, or perhaps extend them to be quartets that explictly include an author. These triple collections could be manipulated by a number of familiar interfaces initially, like chatrooms, documents, threaded media like Mastodon and so on. It should facilitate social organization by allowing individual peers to federate with one another, agreeing to mirror subsets of each others data, potentially making use of larger and more fixed resources as well as low power consumer devices. The network can be made more efficient by content addressing each collection of triples, and can make use of encryption schemes like capability-based security to scope data to a specific set of recipients. The goal would be to make an evolving protocol that can represent some underlying information in arbitrary interfaces from scientific data through the mundanities of everyday communication like sharing photos or planning events.

In the short term this looks more like mayfirst or co-op cloud than traditional cloud systems, where people voluntarily cooperate to build infrastructure that isn't the faceless corporate technology that dominates computing currently. The fediverse is another ongoing experiment in collectively owned, interoperable systems, where individual groups like we at neuromatch.social organize and administer their own systems. Longer term we can start building these out to true peer to peer technologies that are a fundamental departure from client-server cloudlike models.

More important than the specific technological instantiation is a shift in what we *value* in technology and what we believe it should do. Rather than customers rent-

[100] Unfortunately, the blockchain and cryptocurrency cult has muddied the waters by laying claim to the phrase "peer to peer" to mean something entirely different. Here I mean it as real, actual peer to peer systems built for abundance rather than generating artificial scarcity, in the lineage of BitTorrent, among others.

ing a handful of platforms, we can organize our own infrastructures for storage and computation to displace cloud platforms across multiple modalities. We can *counterbuild* the fill the space currently occupied by the cloud without replicating its

We face a stark choice for our future. The Cloud is circling, will it eat us alive? Will we build a space of universalizing knowledge graphs that allow the seamless linking and trade of every element of our society, powering algorithmic systems from information organization through medical systems, governance, and policing. Will we continue to let information conglomerates farm us for our data and feed it back to us, reprocessed, as Content and Knowledge™? Will we be hooked by the lip by barbed convenience that promises us magic, but delivers us only greater surveillance, control, and dependence? Will our attempts at resistance only ever amount to a neverending treadmill of startups and publicly-funded projects that can't break from the gravitational pull of The Cloud Orthodoxy, retreading its worldview of asymmetrical power concentration, inevitably shuttered or bought as they fail to compete on the same territory as the information giants?

Or will we build a better world?

References

[1] McKenzie Wark. *Capital Is Dead: Is This Something Worse?* Verso Books, February 2021. ISBN 978-1-78873-533-9. 1, 2.0.3

[2] Vinay K. Chaudhri, Chaitanya Baru, Naren Chittar, Xin Luna Dong, Michael Genesereth, James Hendler, Aditya Kalyanpur, Douglas B. Lenat, Juan Sequeda, Denny Vrandečić, and Kuansan Wang. Knowledge graphs: Introduction, history, and perspectives. *AI Magazine*, 43(1):17–29, 2022. ISSN 2371-9621. https://doi.org/10.1002/aaai.12033. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12033. 2, 6, 71

[3] Pascal Hitzler. A review of the semantic web field. *Communications of the ACM*, 64(2):76–83, January 2021. ISSN 0001-0782. https://doi.org/10.1145/3397512. URL https://doi.org/10.1145/3397512. 2, 2.0.1, 2.0.2

[4] Jihong Yan, Chengyu Wang, Wenliang Cheng, Ming Gao, and Aoying Zhou. A retrospective of knowledge graphs. *Frontiers of Computer Science*, 12(1):55–74, February 2018. ISSN 2095-2236. https://doi.org/10.1007/s11704-016-5228-9. URL https://doi.org/10.1007/s11704-016-5228-9. 2

[5] Mike Bergman. A Common Sense View of Knowledge Graphs, July 2019. URL https://www.mkbergman.com/2244/a-common-sense-view-of-knowledge-graphs/. 2

[6] Lisa Ehrlinger and Wolfram Wöß. Towards a Definition of Knowledge Graphs. *Semantics*, September 2016. 2

[7] Jonny L. Saunders. Decentralized Infrastructure for (Neuro)science, August 2022. URL http://arxiv.org/abs/2209.07493. 2.0.1, 10, 12, 3.0.2, 97

[8] Tim Berners-Lee. Links and Law, April 1997. URL https://www.w3.org/DesignIssues/LinkLaw. 3

[9] Tim Berners-Lee. Links and Law: Myths, April 1997. URL https://www.w3.org/DesignIssues/LinkMyths.html. 3

[10] Tim Berners-Lee. What the Semantic Web can Represent, September 1998. URL https://www.w3.org/DesignIssues/RDFnot.html. 3, 7

[11] Tim Berners-Lee. The Scale-free nature of the Web, 1998. URL https://www.w3.org/DesignIssues/Fractal.html. 3

[12] Tim Berners-Lee, James HENDLER, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001. ISSN 0036-8733. URL https://www.jstor.org/stable/26059207. 3, 73

[13] Tim Berners-Lee. Cultures and Boundaries, July 2007. URL https://www.w3.org/DesignIssues/Culture.html. 3

[14] Sean B. Palmer. Ditching the Semantic Web?, March 2008. URL http://inamidst.com/whits/2008/ditching. 2.0.2

[15] Aaron Swartz. Aaron Swartz's A Programmable Web: An Unfinished Work. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 3(2):1–64, February 2013. ISSN 2160-4711, 2160-472X. https://doi.org/10.2200/S00481ED1V01Y201302WBE005. URL http://www.morganclaypool.com/doi/abs/10.2200/S00481ED1V01Y201302WBE005. 2.0.2

[16] Lindsay Poirier. A Turn for the Scruffy: An Ethnographic Study of Semantic Web Architecture. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 359–367, New York, NY, USA, June 2017. Association for Computing Machinery. ISBN 978-1-4503-4896-6. https://doi.org/10.1145/3091478.3091505. URL https://doi.org/10.1145/3091478.3091505. 2.0.2, 100

[17] Tim Berners-Lee. Linked Data, July 2006. URL https://www.w3.org/DesignIssues/LinkedData.html. 2.0.2, 2.0.3

[18] Amit Singhal. Introducing the Knowledge Graph: Things, not strings, May 2012. URL https://blog.google/products/search/introducing-knowledge-graph-things-not/. 2.0.3

[19] Iain. Freebase is dead, long live Freebase, May 2016. URL https://medium.com/@iainsproat/freebase-is-dead-long-live-freebase-6c1daff44d19. 2.0.3

[20] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it's done. *Queue*, 17(2):Pages 20:48–Pages 20:75, April 2019. ISSN 1542-7730. https://doi.org/10.1145/3329781.3332266. URL https://doi.org/10.1145/3329781.3332266. 2.0.3, 8, 13

[21] Ben Tarnoff. *Internet for the People: The Fight for Our Digital Future*. Verso Books, June 2022. ISBN 978-1-83976-202-4. 2.0.3

[22] Shoshana Zuboff. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books, London, 2019. ISBN 978-1-78125-685-5. 2.0.3

[23] Tim Berners-Lee. Socially aware cloud storage, August 2009. URL https://www.w3.org/DesignIssues/CloudStorage.html. 2.0.3

[24] Tim Berners-Lee. Goals for a Human-Data Interface, July 2010. URL https://www.w3.org/DesignIssues/TabulatorGoals.html. 2.0.3

[25] Neo4j. Neo4j Customers. URL https://neo4j.com/customers/. 2.0.3

[26] Enterprise Knowledge Graph Foundation and Michael Atkin. Knowledge Graph Industry Survey Report, October 2022. URL https://www.ontotext.com/knowledgehub/white_paper/knowledge-graph-industry-survey-report/. 2.0.3, 6

[27] Jennifer L. Schenker. New Report Details Industry's Use of Knowledge Graphs, May 2021. URL https://theinnovator.news/new-report-details-industrys-use-of-knowledge-graphs/. 6, 5

[28] Juan Sequeda and Ora Lassila. Designing and Building Enterprise Knowledge Graphs. Synthesis Lectures on Data, Semantics, and Knowledge, Cham, 2021. Springer International Publishing. ISBN 978-3-031-00788-0 978-3-031-01916-6. https://doi.org/10.1007/978-3-031-01916-6. URL https://link.springer.com/10.1007/978-3-031-01916-6. 6, 8

[29] Antonia Azzini, Sylvio Barbon, Valerio Bellandi, Tiziana Catarci, Paolo Ceravolo, Philippe Cudré-Mauroux, Samira Maghool, Jaroslav Pokorny, Monica Scannapieco, Florence Sedes, Gabriel Marques Tavares, and Robert Wrembel. Advances in Data Management in the Big Data Era. In Michael Goedicke, Erich Neuhold, and Kai Rannenberg, editors, *Advancing Research in Information and Communication Technology*, volume 600, pages 99–126, Cham, 2021. Springer International Publishing. ISBN 978-3-030-81700-8 978-3-030-81701-5. https://doi.org/10.1007/978-3-030-81701-5_4. URL https://link.springer.com/10.1007/978-3-030-81701-5_4. 6

[30] Suresh Toby Segaran. Two-phase construction of data graphs from disparate inputs, April 2020. URL https://patents.google.com/patent/US10614127B2/. 6

[31] Paolo Ceravolo, Antonia Azzini, Marco Angelini, Tiziana Catarci, Philippe Cudré-Mauroux, Ernesto Damiani, Alexandra Mazak, Maurice Van Keulen, Mustafa Jarrar, Giuseppe Santucci, Kai-Uwe Sattler, Monica Scannapieco, Manuel Wimmer, Robert Wrembel, and Fadi Zaraket. Big Data Semantics. *Journal on Data Semantics*, 7(2):65–85, June 2018. ISSN 1861-2032, 1861-2040. https://doi.org/10.1007/s13740-018-0086-2. URL http://link.springer.com/10.1007/s13740-018-0086-2. 6

[32] Maya Natarajan. From Graph To Knowledge Graph: A short journey to unlimited insights. URL https://neo4j.com/whitepapers/knowledge-graphs-unlimited-insights/thanks/. 6

[33] Dean Allemang. Merging data graphs made easy, December 2022. URL https://scribe.rip/@dallemang/merging-data-graphs-made-easy-8b7e616acfe6. 7

[34] Dean Allemang. Merging tables is hard, December 2022. URL https://scribe.citizen4.eu/@dallemang/merging-tables-is-hard-89d8637a081. 7

[35] Boris Villazon-Terrazas, Nuria Garcia-Santa, Yuan Ren, Alessandro Faraotti, Honghan Wu, Yuting Zhao, Guido Vetere, and Jeff Z. Pan. Knowledge Graph Foundations. In Jeff Z. Pan, Guido Vetere, Jose Manuel Gomez-Perez, and Honghan Wu, editors, *Exploiting Linked Data and Knowledge Graphs in Large Organisations*, pages 17–55. Springer International Publishing, Cham, 2017. ISBN 978-3-319-45654-6. https://doi.org/10.1007/978-3-319-45654-6_2. URL https://doi.org/10.1007/978-3-319-45654-6_2. 7

[36] Sarah Lamdan. *Data Cartels: The Companies That Control and Monopolize Our Information*. Stanford University Press, Stanford, California, 2023. ISBN 978-1-5036-1507-6 978-1-5036-3371-1. 7

[37] RELX. Annual Report 2022, February 2023. URL https://www.relx.com/~/media/Files/R/RELX-Group/documents/reports/annual-reports/relx-2022-annual-report.pdf. 7, 8, 30

[38] Roger C. Schonfeld. A Reorganization at Elsevier, May 2022. URL https://scholarlykitchen.sspnet.org/2022/05/02/reorganization-elsevier/. 7

[39] Elsevier and EMD Serono. Making medical information easily accessible to healthcare professionals, 2021. URL https://www.elsevier.com/__data/assets/pdf_file/0004/1276087/Elsevier-EMD-Serono-phactMI-collaboration.pdf. 8

[40] Elsevier. Rethink Clincal Content, April 2020. URL https://www.elsmediakits.com/intelligence/white-paper/white-paper%3A-rethink-clinical-content. 8

[41] Sam Biddle. LexisNexis to Provide Giant Database of Personal Information to ICE, April 2021. URL https://theintercept.com/2021/04/02/ice-database-surveillance-lexisnexis/. 8

[42] Sam Biddle. ICE Searched LexisNexis Database Over 1 Million Times in Just Seven Months. *The Intercept*, June 2022. URL https://theintercept.com/2022/06/09/ice-lexisnexis-mass-surveillances/. 8

[43] LexisNexis Risk Solutions. Accurint® TraX™, . URL https://risk.lexisnexis.com/products/accurint-trax. 8

[44] LexisNexis Risk Solutions. Telematics OnDemand, . URL https://risk.lexisnexis.com/products/telematics-ondemand. 8

[45] LexisNexis Risk Solutions. Accurint for Legal Professionals, April 2022. URL https://web.archive.org/web/20230308034302/https://www.lexisnexis.com/pdf/AccurintForLegalProfessionals/24.pdf. 8

[46] LexisNexis Risk Solutions. LexID, . URL https://risk.lexisnexis.com/our-technology/lexid. 8

[47] Neo4j. Neo4j + U.S. Army Case Study, 2021. URL https://neo4j.com/case-studies/us-army/. 8

[48] David Cohen, Kevin Richards, and Khan Tasinga. System and Method for Sharing Investigation Result Data, November 2015. URL https://patents.google.com/patent/AU2013251186B2/. 8

[49] Shivam Mathura, Lucas Lemanowicz, and Tim Vergenz. Automated database analysis to detect malfeasance, August 2017. URL https://patents.google.com/patent/US20170221063A1/. 8

[50] Timothy Yousaf, Alexander Mark, Sharon Hao, David Cohen, Andrew Elder, Daniel Lidor, Joel Ossher, Christopher RICHBOURG, Joshua Zavilla, and Kevin Zhang. Systems, methods, user interfaces and algorithms for performing database analysis and search of information involving structured and/or semi-structured data, January 2018. URL https://patents.google.com/patent/US9881066B1/. 8

[51] Eric Knudson, Matthew Gerhardt, Andrew Elder, and Eli Rosofsky. Systems and methods for annotating and linking electronic documents, January 2021. URL https://patents.google.com/patent/US20210004530A1/. 8

[52] Andrew Iliadis and Amelia Acker. The seer and the seen: Surveying Palantir's surveillance platform. *The Information Society*, 38(5):334–363, October 2022. ISSN 0197-2243, 1087-6537. https://doi.org/10.1080/01972243.2022.2100851. URL https://www.tandfonline.com/doi/full/10.1080/01972243.2022.2100851. 8

[53] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. Inside Technology. MIT Press, Cambridge, MA, USA, September 1999. ISBN 978-0-262-02461-7. 3.0.1, 16, 48

[54] Wendy Liu. Freedom Isn't Free, August 2018. URL https://logicmag.io/failure/freedom-isnt-free/. 3.0.1

[55] McKenzie Wark. *A Hacker Manifesto*. Harvard University Press, October 2004. ISBN 978-0-674-01543-2. URL https://www.hup.harvard.edu/catalog.php?isbn=9780674015432. 9

[56] Daniel Goldsmith. The Original Sin of Free Software, 2019. URL https://lipu.dgold.eu/original-sin. 9

[57] James Halliday. Open Source is Not Enough, May 2018. URL https://notesfrombelow.org/article/open-source-is-not-enough. 9

[58] Rob Hunter. Reclaiming the Computing Commons. *Jacobin*, May 2016. URL https://jacobin.com/2016/02/free-software-movement-richard-stallman-linux-open-source-enclosure/. 9

[59] Melody Horn. Post-Open Source, August 2020. URL https://www.boringcactus.com/2020/08/13/post-open-source.html. 9

[60] Denis Pushkarev. So, what's next?, February 2023. URL https://github.com/zloirock/core-js/blob/76f8648790efe74634874012701f387884d2c549/docs/2023-02-14-so-whats-next.md. 9

[61] Steve Marquess. Speeds and Feeds › Of Money, Responsibility, and Pride, April 2014. URL https://veridicalsystems.com/blog/of-money-responsibility-and-pride/index.html. 9

[62] Sean Gallagher. Rage-quit: Coder unpublished 17 lines of JavaScript and "broke the Internet", March 2016. URL https://arstechnica.com/information-technology/2016/03/rage-quit-coder-unpublished-17-lines-of-javascript-and-broke-the-internet/. 9

[63] Christofer Dutz. Your free trial version of "open-source" has expired, please update to a commercial plan, January 2022. URL https://github.com/chrisdutz/blog/blob/835dbf45eaa49aa153604e7e0064b29435f43554/plc4x/free-trial-expired.adoc. 9

[64] Matthew Butterick. GitHub Copilot investigation · Joseph Saveri Law Firm & Matthew Butterick, October 2022. URL https://githubcopilotinvestigation.com/. 9

[65] Matthew Butterick. GitHub Copilot litigation, November 2022. URL https://githubcopilotlitigation.com/. 9

[66] Rob O'Leary. VS Code - What's the deal with the telemetry?, April 2022. URL https://www.roboleary.net/tools/2022/04/20/vscode-telemetry.html. 9

[67] VSCodium - Open Source Binaries of VSCode. URL https://vscodium.com/. 9

[68] Philip Mirowski. The future(s) of open science. *Social Studies of Science*, 48(2):171–203, April 2018. ISSN 0306-3127. https://doi.org/10.1177/0306312718772086. URL https://doi.org/10.1177/0306312718772086. 9

[69] Doris Allhutter. Of "Working Ontologists" and "High-Quality Human Components": The Politics of Semantic Infrastructures. In *Of "Working Ontologists" and "High-Quality Human Components": The Politics of Semantic Infrastructures*, pages 326–348. Princeton University Press, May 2019. ISBN 978-0-691-19060-0. https://doi.org/10.1515/9780691190600-023. URL https://www.degruyter.com/document/doi/10.1515/9780691190600-023/html. 9, 48, 56

[70] User talk:Jimbo Wales/Archive 192. *Wikipedia*, August 2015. URL https://en.wikipedia.org/w/index.php?title=User_talk:Jimbo_Wales/Archive_192&oldid=1143087052#WP_traffic_from_Google_declining. 11

[71] Roy Hinkis. Google steals 550+ million Wikipedia clicks in 6 months, traffic drop confirmed by Wiki's Jimmy Wales, August 2015. URL https://web.archive.org/web/20160114172612/http://www.similarweb.com/blog/google-steals-over-550-million-clicks-from-wikipedia-in-6-months. 11

[72] Molly White. Wikimedia timeline of events, 2014–2016, February 2016. URL http://mollywhite.net/wikimedia-timeline/. 11

[73] William Buetler. Search and Destroy: The Knowledge Engine and the Undoing of Lila Tretikov, February 2016. URL https://thewikipedian.net/2016/02/19/knowledge-engine-lila-tretikov/. 11

[74] Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. From Freebase to Wikidata: The Great Migration. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1419–1428, Montréal Québec Canada, April 2016. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1. https://doi.org/10.1145/2872427.2874809. URL https://dl.acm.org/doi/10.1145/2872427.2874809. 11

[75] Wikimedia Meta-Wiki. Google - Meta. URL https://meta.wikimedia.org/wiki/Google. 11

[76] Google's stake in Wikidata and Wikipedia - Wikidata - lists.wikimedia.org, 2019. URL https://lists.wikimedia.org/hyperkitty/list/wikidata@lists.wikimedia.org/thread/KOCJRSDG57VYWQ4F2BPF7TH7R7YXGF7G/#KOCJRSDG57VYWQ4F2BPF7TH7R7YXGF7G. 11

[77] Denny Vrandečić and Markus Krötzsch. Wikidata: A Free Collaborative Knowledge Base. *Communications of the ACM*, 57(10):78–85, 2014. 11

[78] CrossRef. The Formation of CrossRef: A Short History, 2009. URL https://www.crossref.org/pdfs/CrossRef10Years.pdf. 11

[79] NISO. RP-8-2008, Journal Article Versions (JAV): Recommendations, 2008. 12

[80] NISO. RP-22-2021: Access & License Indicators, 2021. URL https://www.niso.org/publications/rp-22-2021-ali. 12

[81] Todd A. Carpenter. New Article Sharing Framework released, May 2021. URL https://scholarlykitchen.sspnet.org/2021/05/17/stm-article-sharing-framework/. 12

[82] SemTech 2011 BOF on structured data in HTML, June 2011. URL https://www.w3.org/2011/06/semtech-bof-notes.html. 13

[83] Andrew Iliadis, Amelia Acker, Wesley Stevens, and Sezgi Başak Kavakli. One schema to rule them all: How Schema.org models the world of search. *Journal of the Association for Information Science and Technology*, n/a(n/a), January 2023. ISSN 2330-1643. https://doi.org/10.1002/asi.24744. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24744. 13

[84] R.V. Guha, Dan Brickley, and Steve Macbeth. Schema.org: Evolution of Structured Data on the Web, November 2015. URL https://queue.acm.org/detail.cfm?id=2857276. 13

[85] Sandro Hawke. Notes from Q&A session at SemTech, RichSnippets session, June 2011. URL https://lists.w3.org/Archives/Public/public-vocabs/2011Jun/0001.html. 13

[86] Paul Wiegmann, Henk de Vries, and Knut Blind. Multi-Mode Standardisation: A Critical Review and a Research Agenda. *Research Policy*, 46(9):1370–1386, July 2017. ISSN 00487333. https://doi.org/10.1016/j.respol.2017.06.002. URL https://repub.eur.nl/pub/107374/. 13

[87] Marcel Heires. The International Organization for Standardization (ISO). *New Political Economy*, 13(3):357–367, September 2008. ISSN 1356-3467. https://doi.org/10.1080/13563460802302693. URL https://doi.org/10.1080/13563460802302693. 13

[88] Jeff Z. Pan, Guido Vetere, Jose Manuel Gomez-Perez, and Honghan Wu, editors. *Exploiting Linked Data and Knowledge Graphs in Large Organisations*. Springer International Publishing, Cham, 2017. ISBN 978-3-319-45652-2 978-3-319-45654-6. https://doi.org/10.1007/978-3-319-45654-6. URL http://link.springer.com/10.1007/978-3-319-45654-6. 13, 87

[89] The Biomedical Data Translator Consortium. Toward A Universal Biomedical Data Translator. *Clinical and Translational Science*, 12(2):86–90, 2019. ISSN 1752-8062. https://doi.org/10.1111/cts.12591. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.12591. 15, 16, 22, 29, 28

[90] Christopher P. Austin, Christine M. Colvis, and Noel T. Southall. Deconstructing the Translational Tower of Babel. *Clinical and Translational Science*, 12(2):85–85, 2019. ISSN 1752-8062. https://doi.org/10.1111/cts.12595. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.12595. 16

[91] Deepak R. Unni, Sierra A. T. Moxon, Michael Bada, Matthew Brush, Richard Bruskiewich, J. Harry Caufield, Paul A. Clemons, Vlado Dancik, Michel Dumontier, Karamarie Fecho, Gustavo Glusman, Jennifer J. Hadlock, Nomi L. Harris, Arpita Joshi, Tim Putman, Guangrong Qin, Stephen A. Ramsey, Kent A. Shefchek, Harold Solbrig, Karthik Soman, Anne E. Thessen, Melissa A. Haendel, Chris Bizon, Christopher J. Mungall, and The Biomedical Data Translator Consortium. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clinical and Translational Science*, n/a(n/a), May 2022. ISSN 1752-8062. https://doi.org/10.1111/cts.13302. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.13302. 18, 17, 22

[92] Richard Bruskiewich, Deepak, Sierra Moxon, Chris Mungall, Harold Solbrig, cbizon, Matthew Brush, Kent Shefchek, Lance Hannestad, YaphetKG, Nomi Harris, bbopjenkins, diatomsRcool, Patrick Wang, Jim Balhoff, Kevin Schaper, JIWEN XIN, Phil Owen, Gregory Stupp, JervenBolleman, The Gitter Badger, Vincent Emonet, and vdancik. Biolink/biolink-model: 2.2.5. Zenodo, September 2021. URL https://zenodo.org/record/5520104. 18

[93] Nicole A. Vasilevsky, Nicolas A. Matentzoglu, Sabrina Toro, Joe E. Flack, Harshad Hegde, Deepak R. Unni, Gioconda Alyea, Joanna S. Amberger, Larry Babb, James P. Balhoff, Taylor I. Bingaman, Gully A. Burns, Tiffany J. Callahan, Leigh C. Carmody, Lauren E. Chan, George S. Chang, Michel Dumontier, Laura E. Failla, May J. Flowers, H. A. Garrett, Dylan Gration, Tudor Groza, Marc Hanauer, Nomi L. Harris, Ingo Helbig, Jason A. Hilton, Daniel S. Himmelstein, Charles T. Hoyt, Megan S. Kane, Sebastian Köhler, David Lagorce, Martin Larralde, Antonia Lock, Irene López Santiago, Donna R. Maglott, Adriana J. Malheiro, Birgit HM Meldal, Julie A. McMurry, Moni Munoz-Torres, Tristan H. Nelson, David Ochoa, Tudor I. Oprea, David Osumi-Sutherland, Helen Parkinson, Zoë M. Pendlington, Ana Rath, Heidi L. Rehm, Lyubov Remennik, Erin R. Riggs, Paola Roncaglia, Justyne E. Ross, Marion F. Shadbolt, Kent A. Shefchek, Morgan N. Similuk, Nicholas Sioutos, Rachel Sparks, Ray Stefancsik, Ralf Stephan, Doron Stupp, Jagadish Chandrabose Sundaramurthi, Imke Tammen, Courtney L. Thaxton, Eloise Valasek, Alex H. Wagner, Danielle Welter, Patricia L. Whetzel, Lori L. Whiteman, Valerie Wood, Colleen H. Xu, Andreas Zankl, Xingmin A. Zhang, Christopher G. Chute, Peter N. Robinson, Christopher J. Mungall, Ada Hamosh, and Melissa A. Haendel. Mondo: Unifying diseases for the world, by the world, April 2022. URL https://www.medrxiv.org/content/10.1101/2022.04.13.22273750v1. 18

[94] Mungall Chris. Introduction to the BioLink datamodel, May 2018. URL https://www.slideshare.net/cmungall/introduction-to-the-biolink-datamodel. 21

[95] Karamarie Fecho, Anne E. Thessen, Sergio E. Baranzini, Chris Bizon, Jennifer J. Hadlock, Sui Huang, Ryan T. Roper, Noel Southall, Casey Ta, Paul B. Watkins, Mark D. Williams, Hao Xu, William Byrd, Vlado Dančík, Marc P. Duby, Michel Dumontier, Gustavo Glusman, Nomi L. Harris, Eugene W. Hinderer, Greg Hyde, Adam Johs, Andrew I. Su, Guangrong Qin, Qian Zhu, and The Biomedical Data Translator Consortium. Progress toward a universal biomedical data translator. *Clinical and Translational Science*, n/a(n/a), May 2022. ISSN 1752-8062. https://doi.org/10.1111/cts.13301. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.13301. 22

[96] John H Morris, Karthik Soman, Rabia E Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C Meng, Conrad C Huang, Gabriel Cerono, Gundolf Schenk, Angela Rizk-Jackson, Adil Harroud, Lauren Sanders, Sylvain V Costes, Krish Bharat, Arjun Chakraborty, Alexander R Pico, Taline Mardirossian, Michael Keiser, Alice Tang, Josef Hardi, Yongmei Shi, Mark Musen, Sharat Israni, Sui Huang, Peter W Rose, Charlotte A Nelson, and Sergio E Baranzini. The scalable precision medicine open knowledge engine (SPOKE): A massive knowledge graph of biomedical information. *Bioinformatics*, 39(2):btad080, February 2023. ISSN 1367-4811. https://doi.org/10.1093/bioinformatics/btad080. URL https://doi.org/10.1093/bioinformatics/btad080. 22, 54

[97] Renaissance Computing Institute (RENCI). Biomedical Data Translator Platform moves to the next phase, March 2022. URL https://renci.org/blog/biomedical-data-translator-platform-moves-to-the-next-phase/. 22

[98] Renaissance Computing Institute (RENCI). Use cases show Translator's potential to expedite clinical research, March 2022. URL https://renci.org/blog/use-cases-show-translators-potential-to-expedite-clinical-research/. 22

[99] Prateek Goel, Adam J Johs, Manil Shrestha, and Rosina O Weber. Explanation Container in Case-Based Biomedical Question-Answering. page 10, September 2021. URL https://web.archive.org/web/*/https://gaia.fdi.ucm.es/events/xcbr/papers/ICCBR_2021_paper_100.pdf. 22

[100] Ruth Hailu. NIH-funded project aims to build a 'Google' for biomedical data, July 2019. URL https://www.statnews.com/2019/07/31/nih-funded-project-aims-to-build-a-google-for-biomedical-data/. 22

[101] Charlotte A Nelson, Riley Bove, Atul J Butte, and Sergio E Baranzini. Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. *Journal of the American Medical Informatics Association : JAMIA*, 29(3):424–434, December 2021. ISSN 1067-5027. https://doi.org/10.1093/jamia/ocab270. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8800523/. 22, 27

[102] Translator Consortium. Clinical Data Services Provider, April 2020. URL https://github.com/NCATSTranslator/Translator-All/blob/1c44f9a2515d239730a070201ccc7d1083c27fed/presentations/Translator_2020_Kick-Off_Presentation-Clinical_Data_Services.pdf. 22

[103] Charlotte A. Nelson, Atul J. Butte, and Sergio E. Baranzini. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. *Nature Communications*, 10(1):3045, July 2019. ISSN 2041-1723. https://doi.org/10.1038/s41467-019-11069-0. URL https://www.nature.com/articles/s41467-019-11069-0. 22, 92

[104] University of California San Francisco. BRIDGE. URL https://bridge.ucsf.edu/. 22

[105] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The Values Encoded in Machine Learning Research, June 2022. URL http://arxiv.org/abs/2106.15590. 23, 48

[106] Abeba Birhane. Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2), February 2021. ISSN 2666-3899. https://doi.org/10.1016/j.patter.2021.100205. URL https://www.cell.com/patterns/abstract/S2666-3899(21)00015-5. 23, 48

[107] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes, October 2021. URL http://arxiv.org/abs/2110.01963. 23

[108] A Ram, Clair A Kronk, Jacob R Eleazer, Joseph L Goulet, Cynthia A Brandt, and Karen H Wang. Transphobia, encoded: An examination of trans-specific terminology in SNOMED CT and ICD-10-CM. *Journal of the American Medical Informatics Association*, (ocab200), September 2021. ISSN 1527-974X. https://doi.org/10.1093/jamia/ocab200. URL https://doi.org/10.1093/jamia/ocab200. 23, 28

[109] Madeline B. Deutsch. Overview of feminizing hormone therapy, June 2016. URL https://transcare.ucsf.edu/guidelines/feminizing-hormone-therapy. 25

[110] Madeline B. Deutsch. Overview of masculinizing hormone therapy, June 2016. URL https://transcare.ucsf.edu/guidelines/masculinizing-therapy. 25

[111] Finn Womack, Jason McClelland, and David Koslicki. Leveraging Distributed Biomedical Knowledge Sources to Discover Novel Uses for Known Drugs, September 2019. URL https://www.biorxiv.org/content/10.1101/765305v2. 26

[112] Chris Bizon, Steven Cox, James Balhoff, Yaphet Kebede, Patrick Wang, Kenneth Morton, Karamarie Fecho, and Alexander Tropsha. ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources. *Journal of Chemical Information and Modeling*, 59(12):4968–4973, December 2019. ISSN 1549-9596. https://doi.org/10.1021/acs.jcim.9b00683. URL https://doi.org/10.1021/acs.jcim.9b00683. 26

[113] Florence Ashley. The Misuse of Gender Dysphoria: Toward Greater Conceptual Clarity in Transgender Health. *Perspectives on Psychological Science*, page 1745691619872987, November 2019. ISSN 1745-6916. https://doi.org/10.1177/1745691619872987. URL https://doi.org/10.1177/1745691619872987. 28

[114] MasstrichtU-IDS. Knowledge Collaboratory. Maastricht University IDS, December 2022. URL https://github.com/MaastrichtU-IDS/knowledge-collaboratory. 28

[115] Cory Doctorow. Regulating Big Tech makes them stronger, so they need competition instead. *The Economist*, June 2019. ISSN 0013-0613. URL https://www.economist.com/open-future/2019/06/06/regulating-big-tech-makes-them-stronger-so-they-need-competition-instead. 29

[116] Alan Z. Rozenshtein. Moderating the Fediverse: Content Moderation on Distributed Social Media, November 2022. URL https://papers.ssrn.com/abstract=4213674. 29

[117] Lauren Bridges. Amazon's Ring is the largest civilian surveillance network the US has ever seen. *The Guardian*, May 2021. ISSN 0261-3077. URL https://www.theguardian.com/commentisfree/2021/may/18/amazon-ring-largest-civilian-surveillance-network-us. 31

[118] AWS announces AWS Healthcare Accelerator for startups in the public sector, June 2021. URL https://aws.amazon.com/blogs/publicsector/aws-announces-healthcare-accelerator-program-startups-public-sector/. 31

[119] Jon Fingas. Amazon officially becomes a health care provider after closing purchase of One Medical, February 2023. URL https://www.engadget.com/amazon-completes-one-medical-acquisition-163431975.html. 31

[120] Rachel Lerman. Amazon built its own health-care service for employees. Now it's selling it to other companies. *Washington Post*, March 2021. ISSN 0190-8286. URL https://www.washingtonpost.com/technology/2021/03/17/amazon-healthcare-service-care-expansion/. 31

[121] Corey Quinn. You Can't Trust Amazon When It Feels Threatened, March 2021. URL https://www.lastweekinaws.com/blog/you-cant-trust-amazon-when-it-feels-threatened/. 31

[122] Susan Krashinsky. Google broke Canada's privacy laws with targeted health ads, watchdog says. *The Globe and Mail*, January 2014. URL https://www.theglobeandmail.com/technology/tech-news/google-broke-canadas-privacy-laws-with-targeted-ads-regulator-says/article16343346/. 31

[123] Krishna Bharat, Stephen Lawrence, and Mehran Sahami. Generating user information for use in targeted advertising, June 2005. URL https://patents.google.com/patent/US20050131762A1/en. 31

[124] Marc Bourreau, Cristina Caffarra, Zhijun Chen, Chongwoo Choe, Gregory S Crawford, Tomaso Duso, Christos Genakos, Paul Heidhues, Martin Peitz, Thomas Rønde, Monika Schnitzer, Nicolas Schutz, Michelle Sovinsky, Giancarlo Spagnolo, Otto Toivanen, Tommaso Valletti, and Thibaud Vergé. Google/Fitbit will monetise health data and harm consumers. (107):13, 2020. 31

[125] Rebecca Pifer. Google plans to boost Medicaid information during redeterminations, March 2023. URL https://www.healthcaredive.com/news/google-boost-medicaid-information-redeterminations/644944/. 31

[126] Yossi Matias and Corrado. Our latest health AI research updates, March 2023. URL https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/. 31

[127] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large Language Models Encode Clinical Knowledge, December 2022. URL http://arxiv.org/abs/2212.13138. 31

[128] Apple. Empowering people to live a healthier day, June 2022. URL https://www.apple.com/newsroom/pdfs/Health-Report-September-2022.pdf. 31

[129] Apple. Healthcare. URL https://www.apple.com/healthcare/. 31

[130] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 243–246, New York, NY, USA, May 2015. Association for Computing Machinery. ISBN 978-1-4503-3473-0. https://doi.org/10.1145/2740908.2742839. URL https://doi.org/10.1145/2740908.2742839. 31

[131] Ying Chen, J. D. Elenee Argentinis, and Griff Weber. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clinical Therapeutics*, 38(4):688–701, April 2016. ISSN 1879-114X. https://doi.org/10.1016/j.clinthera.2015.12.001. 31

[132] Chaitan Baru, Martin Halbert, Lara Campbell, Tess DeBlanc-Knowles, Jemin George, Wo Chang, Adam Pah, Douglas Maughan, Ilya Zaslavsky, Amanda Stathopoulos, Ellie Young, Kat Albrecht, Amit Sheth, Emanuel Sallinger, Katerine Osatuke, Angela Rizk-Jackson, Eric Jahn, Kenneth Berkowitz, Bandana Kar, Erica Smith, Krzystof Janowicz, Brian Handspicker, Esther Jackson, Lauren Sanders, Li Chengkai, Florence Hudson, Lilit Yeghiazarian, Cogan Shimizu, Glenn Ricart, Raschid Louiqa, Dalia Varanka, Greg Seaton, Luis Amaral, Oktie Hassenzadeh, Silviu Cucerzan, Matt Bishop, Ora Lassila, Sharat Israni, Matthew Lange, Pascal Hitzler, Ryan McGranaghan, Michael Cafarella, Paul Wormeli, Todd Bacastow, Murat Omay, Sam Klein, Ying Ding, Nariman Ammar, and Sergio Baranzini. Open Knowledge Network Roadmap: Powering The Next Data Revolution. September 2022. URL https://nsf-gov-resources.nsf.gov/2022-09/OKN%20Roadmap%20-%20Report_v03.pdf. 3.0.3, 32

[133] National Science Foundation. NSF Convergence Accelerator awards bring together scientists, businesses, nonprofits to benefit workers, September 2019. URL https://www.nsf.gov/news/special_reports/announcements/091019.jsp. 32, 43

[134] National Science Foundation. NSF 22-017 - Dear Colleague Letter: Encouraging Research on Open Knowledge Networks, November 2021. URL https://www.nsf.gov/pubs/2022/nsf22017/nsf22017.jsp. 33

[135] Sui Huang. NIH 1OT2TR003450-01 - EVIDARA: Automated Evidential Support from Raw Data for relay agents in Biomedical KG Queries, January 2020. URL https://reporter.nih.gov/search/PyKrY9MwK02kM4isuX9HYg/project-details/10057190. 33

[136] Sergio Baranzini, Sharat Israni, James Brase, and Sui Huang. NSF Award Search: Award # 2033569 - A1: A Multi-Scale Open Knowledge Network for Biomedicine, September 2022. URL https://www.nsf.gov/awardsearch/showAward?AWD_ID=2033569&HistoricalAwards=false. 33

[137] Krzysztof Janowicz, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, Cogan Shimizu, Colby Fisher, Ling Cai, Gengchen Mai, Joseph Zalewski, Lu Zhou, Shirly Stephen, Seila Gonzalez, Bryce Mecum, Anna Carr, Andrew Schroeder, Dave Smith, Dawn Wright, Sizhe Wang, Yuanyuan Tian, Zilong Liu, Meilin Shi, Anthony D'Onofrio, Zhining Gu, and Kitty Currier. Know, Know Where, Knowwheregraph: A Densely Connected, Cross-Domain Knowledge Graph and Geo-Enrichment Service Stack for Applications in Environmental Intelligence. *AI Magazine*, 43(1):30–39, March 2022. ISSN 2371-9621, 0738-4602. https://doi.org/10.1609/aimag.v43i1.19120. URL http://ojs.aaai.org/index.php/aimagazine/article/view/19120. 33

[138] National Security Commission on Artificial Intelligence. Final Report, 2021. URL https://www.nscai.gov/2021-final-report/. 34

[139] Big Data Interagency Working Group, Subcommittee on Networking & Information Technology Research & Development, and Committee on Science & Technology Enterprise. Open Knowledge Network: Summary of the Big Data IWG Workshop, October 4-5, 2017. Technical report, November 2018. URL https://www.nitrd.gov/pubs/Open-Knowledge-Network-Workshop-Report-2018.pdf. 34, 35

[140] Mate Bioservices Inc. SPOKE Cloud, 2021. URL https://www.matebioservices.com/spoke-cloud. 34

[141] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. Picador St. Martin's Press, New York, first picador edition edition, 2019. ISBN 978-1-250-21578-9. 35, 40

[142] Aaron Shapiro. *Design, Control, Predict: Logistical Governance in the Smart City*. University of Minnesota Press, Minneapolis, 2020. ISBN 978-1-4529-6211-5. 37

[143] Sarah Brayne. *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford University Press, October 2020. URL https://doi.org/10.1093/oso/9780190684099.001.0001. 38, 41, 42

[144] Information Systems Advisory Board, County of Los Angeles. CONTRACT BETWEEN THE COUNTY OF LOS ANGELES AND cFIVE SOLUTIONS, INC. FOR CONSOLIDATED CRIMINAL HISTORY REPORTING SYSTEM MAINTENANCE, SUPPORT, AND ENHANCEMENT SERVICES, September 2021. URL https://file.lacounty.gov/SDSInter/bos/supdocs/161887.pdf. 40

[145] Chief Executive Office, County of Los Angeles. Strategic Plan Goal Three: Realize Tomorrow's Government Today, May 2022. URL https://ceo.lacounty.gov/strategic-plan-goal-three/. 41

[146] Ali Farahani and Information Systems Advisory Body, County of Los Angeles. Linking Public Safety Data to the Countywide Master Data Management System, October 2016. 41

[147] Sarah Lamdan. Defund the Police, and Defund Big Data Policing, Too, June 2020. URL https://www.jurist.org/commentary/2020/06/sarah-lamdan-data-policing/. 41

[148] Matthew Guariglia. Technology Can't Predict Crime, It Can Only Weaponize Proximity to Policing, September 2020. URL https://www.eff.org/deeplinks/2020/09/technology-cant-predict-crime-it-can-only-weaponize-proximity-policing. 42, 41

[149] McGrory Kathleen, Neil Bedi, and Douglas R. Clifford. Targeted. *Tampa Bay Times*, September 2020. URL https://projects.tampabay.com/projects/2020/investigations/police-pasco-sheriff-targeted/intelligence-led-policing. 42, 41

[150] Stop LAPD Spying Coalition. Racial Terror and White Wealth in South Central. In *Automating Banishment*. November 2021. URL https://automatingbanishment.org/section/5-racial-terror-and-white-wealth-in-south-central/#operation-laser-racial-terror. 42

[151] Stop LAPD Spying Coalition. Before the Bullet Hits the Body, May 2018. URL https://stoplapdspying.org/wp-content/uploads/2018/05/Before-the-Bullet-Hits-the-Body-May-8-2018.pdf. 42

[152] Stop LAPD Spying Coalition. Letter from 28 Professors and 48 Graduate Students of UCLA to LAPD, April 2019. URL https://stoplapdspying.medium.com/on-tuesday-april-2nd-2019-twenty-eight-professors-and-forty-graduate-students-of-university-of-8ed7da1a8 42

[153] Davide Castelvecchi. Mathematicians urge colleagues to boycott police work in wake of killings. *Nature*, June 2020. https://doi.org/10.1038/d41586-020-01874-9. URL https://www.nature.com/articles/d41586-020-01874-9. 42

[154] Sun-ha Hong. Prediction as Extraction of Discretion. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 925–934, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. https://doi.org/10.1145/3531146.3533155. URL https://doi.org/10.1145/3531146.3533155. 42

[155] Stop LAPD Spying Coalition. FUCK THE POLICE, TRUST THE PEOPLE: Surveillance Bureaucracy Expands the Stalker State, June 2020. URL https://stoplapdspying.org/wp-content/uploads/2020/06/TRUST-THE-PPL-not-the-POLICE.pdf. 42

[156] David Graeber. *The Utopia of Rules: On Technology, Stupidity, and the Secret Joys of Bureaucracy*. Melville House, Brooklyn : London, 2015. ISBN 978-1-61219-374-8. 42

[157] Directorate-General for Research and Innovation (European Commission), EOSC Executive Board, Mark van de Sanden, Dale Robertson, Owen Appleton, and Paolo Manghi. *EOSC Architecture Working Group View on the Minimum Viable EOSC: Report from the EOSC Executive Board Working Group (WG) Architecture*. Publications Office of the European Union, LU, 2021. ISBN 978-92-76-29813-7. URL https://data.europa.eu/doi/10.2777/492370. 43

[158] Directorate-General for Research and Innovation (European Commission). *Solutions for a Sustainable EOSC: A FAIR Lady (Olim Iron Lady) Report from the EOSC Sustainability Working Group*. Publications Office of the European Union, LU, 2020. ISBN 978-92-76-25594-9. URL https://data.europa.eu/doi/10.2777/870770. 43

[159] Directorate-General for Research and Innovation (European Commission), EOSC Executive Board, Oscar Corcho, Magnus Eriksson, Krzysztof Kurowski, Milan Ojsteršek, Christine Choirat, Mark van de Sanden, and Frederik Coppens. *EOSC Interoperability Framework: Report from the EOSC Executive Board Working Groups FAIR and Architecture*. Publications Office of the European Union, LU, 2021. ISBN 978-92-76-28949-4. URL https://data.europa.eu/doi/10.2777/620649. 43

[160] Jesse V. File:Folding@home and Supercomputer Computational Powers.png, April 2012. URL https://commons.wikimedia.org/wiki/File:Folding@home_and_Supercomputer_Computational_Powers.png. 45

[161] Maxwell I. Zimmerman, Justin R. Porter, Michael D. Ward, Sukrit Singh, Neha Vithani, Artur Meller, Upasana L. Mallimadugula, Catherine E. Kuhn, Jonathan H. Borowsky, Rafal P. Wiewiora, Matthew F. D. Hurley, Aoife M. Harbison, Carl A. Fogarty, Joseph E. Coffland, Elisa Fadda, Vincent A. Voelz, John D. Chodera, and Gregory R. Bowman. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nature Chemistry*, 13(7):651–659, July 2021. ISSN 1755-4349. https://doi.org/10.1038/s41557-021-00707-0. URL https://www.nature.com/articles/s41557-021-00707-0. 45

[162] Ernesto Van der Sar. BitTorrent: The "one third of all Internet traffic" Myth * TorrentFreak, September 2006. URL https://torrentfreak.com/bittorrent-the-one-third-of-all-internet-traffic-myth/. 45

[163] Ernesto Van der Sar. P2P Traffic Is Booming, BitTorrent The Dominant Protocol * TorrentFreak, November 2007. URL https://torrentfreak.com/p2p-traffic-still-booming-071128/. 45

[164] Richard Lawler. An Amazon server outage caused problems for Alexa, Ring, Disney Plus, and deliveries, December 2021. URL https://www.theverge.com/2021/12/7/22822332/amazon-server-aws-down-disney-plus-ring-outage. 45

[165] Lee Hutchinson. Amazon Web Services outage once again shows reality behind "the cloud", October 2012. URL https://arstechnica.com/information-technology/2012/10/amazon-web-services-outage-once-again-shows-reality-behind-the-cloud/. 45

[166] Drew DeVault. The reckless, infinite scope of web browsers, March 2020. URL https://drewdevault.com/2020/03/18/Reckless-limitless-scope.html. 45

[167] In re: Google Digital Advertising Antitrust Litigation - Amended Complaint #152, October 2021. URL https://www.courtlistener.com/docket/60149069/152/in-re-google-digital-advertising-antitrust-litigation/. 46

[168] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. https://doi.org/10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922. 48, 61

[169] Abeba Birhane and Olivia Guest. Towards decolonising computational sciences, September 2020. URL http://arxiv.org/abs/2009.14258. 48

[170] Catherine D'Ignazio and Lauren F. Klein. *Data Feminism*. Strong Ideas Series. The MIT Press, Cambridge, Massachusetts, 2020. ISBN 978-0-262-04400-4. 48, 54

[171] Astrid Mager. Defining Algorithmic Ideology: Using Ideology Critique to Scrutinize Corporate Search Engines. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, 12(1): 28–39, February 2014. ISSN 1726-670X, 1726-670X. https://doi.org/10.31269/triplec.v12i1.439. URL https://www.triple-c.at/index.php/tripleC/article/view/439. 48

[172] Tung-Hui Hu. *A Prehistory of the Cloud*. MIT Press, August 2015. ISBN 978-0-262-33009-1. https://doi.org/10.7551/mitpress/9780262029513.001.0001. URL https://direct.mit.edu/books/book/2291/A-Prehistory-of-the-Cloud. 51

[173] Marius Paşca. Organizing and searching the world wide web of facts – step two: Harnessing the wisdom of the crowds. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 101–110, New York, NY, USA, May 2007. Association for Computing Machinery. ISBN 978-1-59593-654-7. https://doi.org/10.1145/1242572.1242587. URL https://dl.acm.org/doi/10.1145/1242572.1242587. 52, 61

[174] Google. Search Quality Evaluator Guidelines, December 2022. URL https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf. 52

[175] Alon Halevy, Peter Norvig, and Fernando Pereira. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24(2):8–12, March 2009. ISSN 1941-1294. https://doi.org/10.1109/MIS.2009.36. 54, 61

[176] Jannis Kallinikos, Aleksi Aaltonen, and Attila Marton. The Ambivalent Ontology of Digital Artifacts. *MIS Quarterly*, 37 (2):357–370, 2013. ISSN 0276-7783. URL https://www.jstor.org/stable/43825913. 57

[177] Cory Doctorow. Tiktok's enshittification, February 2023. URL https://pluralistic.net/2023/01/21/potemkin-ai/#hey-guys. 58

[178] Lawrence Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. In *The Web Conference*, November 1999. URL https://www.semanticscholar.org/paper/The-PageRank-Citation-Ranking-%3A-Bringing-Order-to-Page-Brin/eb82d3035849cd23578096462ba419b53198a556. 59, 60

[179] Xiao Li. Understanding the Semantic Structure of Noun Phrase Queries. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1337–1345, 2010. https://doi.org/10.5555/1858681.1858817. URL https://aclanthology.org/P10-1136.pdf. 61

[180] Joseph Reisinger and Marius Pasca. Fine-Grained Class Label Markup of Search Queries. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*, 2011. 61

[181] Marius Pasca and Benjamin Van Durme. What you seek is what you get: Extraction of class attributes from query logs. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2832–2837, San Francisco, CA, USA, January 2007. Morgan Kaufmann Publishers Inc. 61

[182] Marius Pasca. Turning web text and search queries into factual knowledge: Hierarchical class attribute extraction. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 1225–1230, Chicago, Illinois, July 2008. AAAI Press. ISBN 978-1-57735-368-3. https://doi.org/10.5555/1620163.1620263. URL https://dl.acm.org/doi/abs/10.5555/1620163.1620263. 61

[183] Marius Paşca. Weakly-supervised discovery of named entities using web search queries. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 683–690, New York, NY, USA, November 2007. Association for Computing Machinery. ISBN 978-1-59593-803-9. https://doi.org/10.1145/1321440.1321536. URL https://dl.acm.org/doi/10.1145/1321440.1321536. 61

[184] Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. Organizing and Searching the World Wide Web of Facts - Step One: The One-Million Fact Extraction Challenge. *AAAI*, 6, 2006. URL https://fileadmin.cs.lth.se/ai/Proceedings/aaai06/12/AAAI06-220.pdf. 61

[185] Marius Pasca. Acquisition of categorized named entities for web search. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, pages 137–145, New York, NY, USA, November 2004. Association for Computing Machinery. ISBN 978-1-58113-874-0. https://doi.org/10.1145/1031171.1031194. URL https://dl.acm.org/doi/10.1145/1031171.1031194. 61

[186] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, December 2017. URL http://arxiv.org/abs/1706.03762. 61

[187] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. 61

[188] Pandu Nayak. Understanding searches better than ever before, October 2019. URL https://blog.google/products/search/search-language-understanding-bert/. 61

[189] Pandu Nayak. MUM: A new AI milestone for understanding information, May 2021. URL https://blog.google/products/search/introducing-mum/. 61, 68

[190] Yi Tay, Zhe Zhao, Dara Bahri, Don Metzler, and Da-Cheng Juan. HyperGrid Transformers: Towards A Single Model for Multiple Tasks. In *ICLR 2021*, 2021. 61

[191] Ronghang Hu and Amanpreet Singh. UniT: Multimodal Multitask Learning with a Unified Transformer, August 2021. URL http://arxiv.org/abs/2102.10772. 61

[192] Microsoft. The Future of Work With AI - Microsoft March 2023 Event, March 2023. URL https://www.youtube.com/watch?v=Bf-dbS9CcRU. 61, 86

[193] Xiao Ma and Ariel Liu. Challenges in Supporting Exploratory Search through Voice Assistants. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, CUI '20, pages 1–3, New York, NY, USA, July 2020. Association for Computing Machinery. ISBN 978-1-4503-7544-3. https://doi.org/10.1145/3405755.3406152. URL https://dl.acm.org/doi/10.1145/3405755.3406152. 61

[194] Ben Gomes. Improving Search for the next 20 years, September 2018. URL https://blog.google/products/search/improving-search-next-20-years/. 61, 68

[195] Jaclyn Konzelmann. Chatting up your Google Assistant just got easier, June 2018. URL https://blog.google/products/assistant/chatting-your-google-assistant-just-got-easier/. 61

[196] Jon Friedman and Kurtis Beavers. Behind-the-Design: Meet Copilot, April 2023. URL https://medium.com/microsoft-design/behind-the-design-meet-copilot-2c68182a0e70. 61

[197] Future of Life Institute. Pause Giant AI Experiments: An Open Letter, March 2023. URL https://futureoflife.org/open-letter/pause-giant-ai-experiments/. 61

[198] Timnit Gebru, Emily M. Bender, Angelina McMillan-Major, and Margaret Mitchell. Statement from the listed authors of Stochastic Parrots on the "AI pause" letter, 2023-02-31. URL https://www.dair-institute.org/blog/letter-statement-March2023. 61, 62

[199] Emily M. Bender and Alexander Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463. 61

[200] Eric Horvitz and Tim Paek. Grounding criterion: Toward a formal theory of grounding. Technical Report MSR-TR-2000-40, April 2000. URL https://www.microsoft.com/en-us/research/publication/grounding-criterion-toward-a-formal-theory-of-grounding/. 61

[201] Dan McQuillan. *Resisting AI: An Anti-Fascist Approach to Artificial Intelligence*. Bristol University Press, Bristol, 2022. ISBN 978-1-5292-1349-2 978-1-5292-1350-8. 62

[202] Adriana Diaz. Immortality is attainable by 2030: Google scientist. *New York Post*, March 2023. URL https://nypost.com/2023/03/29/immortality-is-attainable-by-2030-google-scientist/. 64

[203] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4, March 2023. URL http://arxiv.org/abs/2303.12712. 65, 73, 84

[204] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, July 2020. URL http://arxiv.org/abs/1910.10683. 65, 66

[205] Aaron Tilley and Kevin McLaughlin. The Seven-Year Itch: How Apple's Marriage to Siri Turned Sour. *The Information*, March 2018. URL https://www.theinformation.com/articles/the-seven-year-itch-how-apples-marriage-to-siri-turned-sour. 66

[206] Tripp Mickle, Cade Metz, and Nico Grant. The Chatbots Are Here, and the Internet Industry Is in a Tizzy. *The New York Times*, March 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/03/08/technology/chatbots-disrupt-internet-industry.html. 67

[207] Google. Why we focus on AI (and to what end), January 2023. URL https://ai.google/our-focus/. 68

[208] Leonard Adolphs, Benjamin Boerschinger, Christian Buck, Michelle Chen Huebscher, Massimiliano Ciaramita, Lasse Espeholt, Thomas Hofmann, Yannic Kilcher, Sascha Rothe, Pier Giuseppe Sessa, and Lierni Sestorain Saralegui. Boosting Search Engines with Interactive Agents, June 2022. URL http://arxiv.org/abs/2109.00527. 68, 69

[209] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. Rethinking Search: Making Domain Experts out of Dilettantes, July 2021. URL https://arxiv.org/abs/2105.02274v2. 68, 70, 85

[210] Manuel Bronstein. Bringing you the next-generation Google Assistant, May 2019. URL https://blog.google/products/assistant/next-generation-google-assistant-io/. 68

[211] Sundar Pichai. A personal Google, just for you, October 2016. URL https://blog.google/products/assistant/personal-google-just-you/. 68

[212] Chirag Shah and Emily M. Bender. Situating Search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 221–232, New York, NY, USA, March 2022. Association for Computing Machinery. ISBN 978-1-4503-9186-3. https://doi.org/10.1145/3498366.3505816. URL https://dl.acm.org/doi/10.1145/3498366.3505816. 68

[213] Amit Bharadwaj. Dependency graph generation in a networked system, June 2020. URL https://patents.google.com/patent/US10679622B2/. 70

[214] Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur D. Szlam, and J. Weston. Reason first, then respond: Modular Generation for Knowledge-infused Dialogue. In *Conference on Empirical Methods in Natural Language Processing*, November 2021. URL https://www.semanticscholar.org/paper/Reason-first%2C-then-respond%3A-Modular-Generation-for-Adolphs-Shuster/d15d96517370c9ed0658d176b979bcf92d1373ea. 70

[215] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. Augmented Language Models: A Survey, February 2023. URL http://arxiv.org/abs/2302.07842. 70

[216] Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. Internet-augmented language models through few-shot prompting for open-domain question answering. 2022. https://doi.org/10.48550/ARXIV.2203.05115. URL https://arxiv.org/abs/2203.05115. 71

[217] Tim Berners-Lee. The Evolution of a specification – Commentary on Web architecture, March 1998. URL https://www.w3.org/DesignIssues/Evolution.html. 71

[218] Peter McBurney and Michael Luck. The Agents Are All Busy Doing Stuff! *IEEE Intelligent Systems*, 22(4):6–7, July 2007. ISSN 1541-1672. https://doi.org/10.1109/MIS.2007.77. URL http://ieeexplore.ieee.org/document/4287266/. 72

[219] Lee Vinsel. You're Doing It Wrong: Notes on Criticism and Technology Hype, February 2021. URL https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5. 73

[220] Jennifer Elias. Google reshuffles virtual assistant unit with focus on Bard A.I. technology. *CNBC*, March 2023. URL https://www.cnbc.com/2023/03/29/google-reorganization-in-assistant-follows-bard-launch-memo-says.html. 74, 80

[221] Pangambam S. Google I/O 2016 Keynote Full Transcript, May 2016. URL https://singjupost.com/google-io-2016-keynote-full-transcript/?singlepage=1. 75

[222] Google Developers. What's new in Google Assistant | Keynote, May 2021. URL https://www.youtube.com/watch?v=O2gCx_iX2vQ. 75

[223] Google. Google Keynote (Google I/O '22), May 2022. URL https://www.youtube.com/watch?v=nP-nMZpLM1A. 76, 77, 81, 84

[224] Statista. Global mobile OS market share 2022, February 2023. URL https://www.statista.com/statistics/272698/global-market-share-held-by-mobile-operating-systems-since-2009/. 77

[225] Tobias Blanke and Jennifer Pybus. The Material Conditions of Platforms: Monopolization Through Decentralization. *Social Media + Society*, 6(4):205630512097163, October 2020. ISSN 2056-3051, 2056-3051. https://doi.org/10.1177/2056305120971632. URL http://journals.sagepub.com/doi/10.1177/2056305120971632. 77

[226] Rebecca Nathenson. Helping Developers Create Meaningful Voice Interactions with Android, June 2022. URL https://developers.googleblog.com/2022/06/Helping-Developers-Create-Meaningful-Voice-Interactions-with-Android.html. 78

[227] Apple Developer Documentation. App Intents. URL https://developer.apple.com/documentation/appintents/. 78

[228] Android Developer Documentation. Build App Actions | Documentation. URL https://developer.android.com/guide/app-actions/get-started. 78

[229] Allison Mercurio, Amanda Elizabeth Baughan, Ariel Liu, Jilin Chen, Xiao Ma, and Xuezhi Wang. A Mixed-Methods Approach to Understanding User Trust after Voice Assistant Failures. 2023. URL https://arxiv.org/abs/2303.00164. 78

[230] Jonathan Stempel. Apple must face Siri voice assistant privacy lawsuit -U.S. judge. *Reuters*, September 2021. URL https://www.reuters.com/technology/apple-must-face-siri-voice-assistant-privacy-lawsuit-us-judge-2021-09-02/. 79

[231] Mitchell Clark. Amazon did the math and would actually prefer getting sued. *The Verge*, June 2021. URL https://www.theverge.com/2021/6/1/22463550/amazon-lawsuit-arbitration-terms-of-service-update-alexa. 79

[232] Jonathan Stempel and Sara Merken. Google must face Voice Assistant privacy lawsuit -U.S. judge. *Reuters*, July 2021. URL https://www.reuters.com/technology/google-must-face-voice-assistant-privacy-lawsuit-us-judge-2021-07-02/. 79

[233] Microsoft Graph Developer Documentation. Microsoft Graph connectors overview, August 2022. URL https://learn.microsoft.com/en-us/graph/connecting-external-content-connectors-overview. 80

[234] Krisztian Balog. *Entity-Oriented Search*, volume 39 of *The Information Retrieval Series*. Springer International Publishing, Cham, 2018. ISBN 978-3-319-93933-9 978-3-319-93935-3. https://doi.org/10.1007/978-3-319-93935-3. URL http://link.springer.com/10.1007/978-3-319-93935-3. 80

[235] Victor Carbune and Matthew Sharifi. Automated assistant adaptation of a response to an utterance and/or of processing of the utterance, based on determined interaction measure, December 2022. URL https://patents.google.com/patent/US20220392449A1/. 81

[236] Satayan Mahajan. Communicating context to a device using an imperceptible audio identifier, April 2021. URL https://patents.google.com/patent/US10971144B2/en/. 81

[237] Lucas Ropek. Amazon Makes Creepy Surveillance Robot Even Creepier With Yet More Ring Integration, September 2022. URL https://gizmodo.com/amazon-astro-robot-ring-virtual-security-guard-cops-1849591805. 81

[238] Matthew Gault and Joseph Cox. Leaked Documents Show How Amazon's Astro Robot Tracks Everything You Do, September 2021. URL https://www.vice.com/en/article/93ypp8/leaked-documents-amazon-astro-surveillance-robot-tracking. 81

[239] Ring. Ring Always Home Cam. URL https://ring.com/always-home-cam-flying-camera. 81

[240] Gustavo Alfonso Aguilar Alas, Viktor Rozgic, and Chao Wang. Multiple classifications of audio data, May 2022. URL https://patents.google.com/patent/US11335347B2/. 81, 82

[241] Victor Roditis Jablokov, Igor Roditis Jablokov, II James Richard Terrell, Marc White, and Scott Edward Paden. Facilitating presentation of ads relating to words of a message, June 2015. URL https://patents.google.com/patent/US9053489B2/. 82

[242] The A.I. Dilemma, March 2023. URL https://vimeo.com/809258916/92b420d98a. 83

[243] Google. 2022 Q4 & Fiscal Year Earnings Webcast Transcript, February 2023. URL https://abc.xyz/investor/static/pdf/2022_Q4_Earnings_Transcript.pdf?cache=c632791. 83

[244] Google. About Smart Bidding. URL https://support.google.com/google-ads/answer/7065882?hl=en. 83

[245] Gilad Edelman. Ad Tech Could Be the Next Internet Bubble. *Wired*, October 2020. ISSN 1059-1028. URL https://www.wired.com/story/ad-tech-could-be-the-next-internet-bubble/. 84

[246] AMP letter. A letter about Google AMP, January 2018. URL http://ampletter.org/. 85

[247] Dieter Bohn. Inside Google's plan to make the whole web as fast as AMP. *The Verge*, March 2018. URL https://www.theverge.com/2018/3/8/17095078/google-amp-accelerated-mobile-page-announcement-standard-web-packaging-urls. 85

[248] Danny Sullivan. Google Search sends more traffic to the open web every year, March 2021. URL https://blog.google/products/search/google-search-sends-more-traffic-open-web-every-year/. 85

[249] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. 86

[250] Freddy Lecue. On the role of knowledge graphs in explainable AI. *Semantic Web*, 11(1):41–51, January 2020. ISSN 1570-0844. https://doi.org/10.3233/SW-190374. URL https://content.iospress.com/articles/semantic-web/sw190374. 87

[251] Krzysztof Janowicz, Pascal Hitzler, Federico Bianchi, Monireh Ebrahimi, and Md Kamruzzaman Sarker. Neural-symbolic integration and the Semantic Web. *Semantic Web*, 11(1):3–11, January 2020. ISSN 1570-0844. https://doi.org/10.3233/SW-190368. URL https://doi.org/10.3233/SW-190368. 87

[252] I. Tiddi, F. Lécué, and P. Hitzler. *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*. IOS Press, May 2020. ISBN 978-1-64368-081-1. 87

[253] Johan Ordish, Hannah Murfet, and Alison Hall. Algorithms as Medical Devices. Technical report, PHG Foundation, 2019. URL https://www.phgfoundation.org/media/74/download/algorithms-as-medical-devices.pdf?v=1. 88

[254] Soleil Shah El-Sayed, Abdul. Medical Algorithms Need Better Regulation, October 2021. URL https://www.scientificamerican.com/article/the-fda-should-better-regulate-medical-algorithms/. 88

[255] Adam Sadilek, Luyang Liu, Dung Nguyen, Methun Kamruzzaman, Stylianos Serghiou, Benjamin Rader, Alex Ingerman, Stefan Mellem, Peter Kairouz, Elaine O. Nsoesie, Jamie MacFarlane, Anil Vullikanti, Madhav Marathe, Paul Eastham, John S. Brownstein, Blaise Aguera y Arcas, Michael D. Howell, and John Hernandez. Privacy-first health research with federated learning. *npj Digital Medicine*, 4(1):1–8, September 2021. ISSN 2398-6352. https://doi.org/10.1038/s41746-021-00489-2. URL https://www.nature.com/articles/s41746-021-00489-2. 89

[256] Hugh Brendan McMahan, Kunal Talwar, Li Zhang, and Daniel Ramage. Training user-level differentially private machine-learned models, October 2022. URL https://patents.google.com/patent/US11475350B2/. 89

[257] Logan Pearsall Smith. *Words and Idioms : Studies in the English Language*. Houghton Mifflin, 1925. URL https://www.semanticscholar.org/paper/Words-and-idioms-%3A-studies-in-the-English-language-Smith/24b5844bc6b64568e0bdc013fc4cd44104b67ed4. 4.3

[258] Jean-Christophe Plantin, Carl Lagoze, Paul N Edwards, and Christian Sandvig. Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, 20(1):293–310, January 2018. ISSN 1461-4448. https://doi.org/10.1177/1461444816661553. URL https://doi.org/10.1177/1461444816661553. 4.3

[259] Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models, April 2023. URL http://arxiv.org/abs/2304.03271. 92

[260] Sun-ha Hong. Control Creep: When the Data Always Travels, So Do the Harms, April 2021. URL https://www.cigionline.org/articles/control-creep-when-data-always-travels-so-do-harms/. 92

[261] Lucas D. Introna and Helen Nissenbaum. Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society*, 16(3):169–185, July 2000. ISSN 0197-2243. https://doi.org/10.1080/01972240050133634. URL https://doi.org/10.1080/01972240050133634. 93

[262] Patrick Stickler. Re: Monotonicity [was: Re: On Consensus] from Patrick Stickler on 2002-09-25 (w3c-rdfcore-wg@w3.org from September 2002), September 2002. URL https://lists.w3.org/Archives/Public/w3c-rdfcore-wg/2002Sep/0276.html. 93

[263] CrimethInc Ex-Workers Collective. Fighting for Our Lives : An Anarchist Primer, 2002. URL https://crimethinc.com/2017/11/28/fighting-for-our-lives-an-anarchist-primer. 94, 96

[264] Meatball Wiki: RightToFork, . URL http://meatballwiki.org/wiki/RightToFork. 99, 100

[265] Murray Bookchin. A Note on Affinity Groups. page 2, 1969. URL https://theanarchistlibrary.org/library/murray-bookchin-a-note-on-affinity-groups. 100

[266] Meatball Wiki: EnlargeSpace, . URL http://meatballwiki.org/wiki/EnlargeSpace. 100

[267] Douglas Harper. Vulgar. URL https://www.etymonline.com/word/vulgar. 100