Two Systems for User Control on the Modern Web

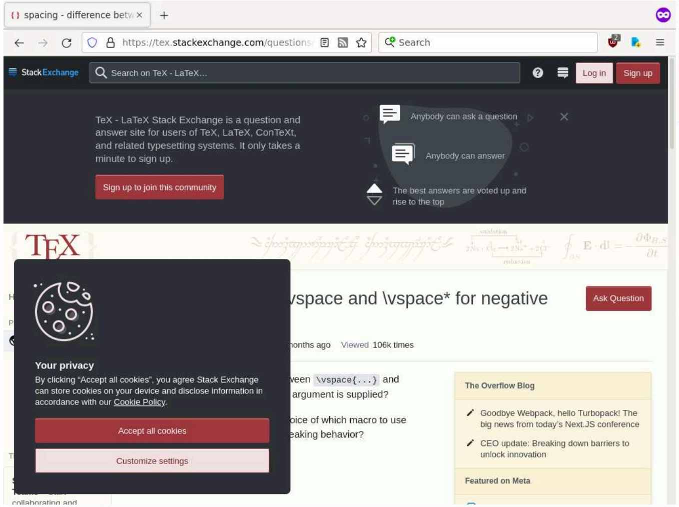
Caleb Malchik

Advisor: Joan Feigenbaum

June 5, 2025

overview

- context and motivations
- collective approach – data co-ops
- Platform for Untrusted Resource Evaluation
- Platform for Content-Structure Inference
- summary


spacing - difference betw ×
https://tex.stackexchange.com/questions
Search
StackExchange
Search on TeX - LaTeX…
Log in
Sign up

TeX - LaTeX Stack Exchange is a question and answer site for users of TeX, LaTeX, ConTeXt, and related typesetting systems. It only takes a minute to sign up.

Sign up to join this community

Anybody can ask a question

Anybody can answer

The best answers are voted up and rise to the top

TEX

vspace and \vspace* for negative

Ask Question

nonths ago Viewed 106k times

veen `\vspace{...}` and argument is supplied?

oice of which macro to use eaking behavior?

The Overflow Blog

- Goodbye Webpack, hello Turbopack! The big news from today's Next.JS conference
- CEO update: Breaking down barriers to unlock innovation

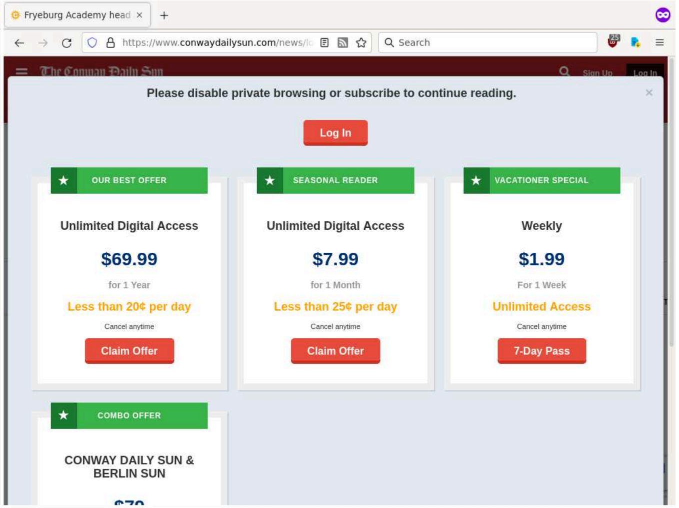
Featured on Meta

Your privacy

By clicking "Accept all cookies", you agree Stack Exchange can store cookies on your device and disclose information in accordance with our Cookie Policy.

Accept all cookies

Customize settings



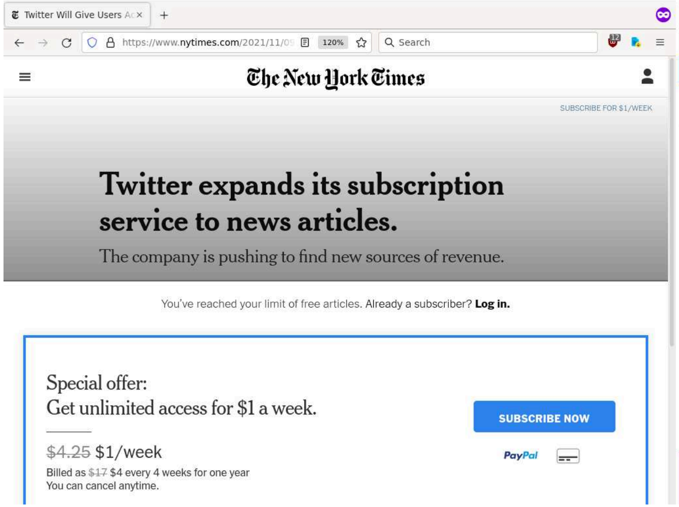
Please disable private browsing or subscribe to continue reading.

Log In

OUR BEST OFFER	SEASONAL READER	VACATIONER SPECIAL
Unlimited Digital Access	Unlimited Digital Access	Weekly
$69.99	$7.99	$1.99
for 1 Year	for 1 Month	For 1 Week
Less than 20¢ per day	Less than 25¢ per day	Unlimited Access
Cancel anytime	Cancel anytime	Cancel anytime
Claim Offer	Claim Offer	7-Day Pass

COMBO OFFER

CONWAY DAILY SUN & BERLIN SUN


Twitter Will Give Users Ac ×
https://www.nytimes.com/2021/11/09
120%
Search
The New York Times
SUBSCRIBE FOR $1/WEEK

Twitter expands its subscription service to news articles.

The company is pushing to find new sources of revenue.

You've reached your limit of free articles. Already a subscriber? **Log in.**

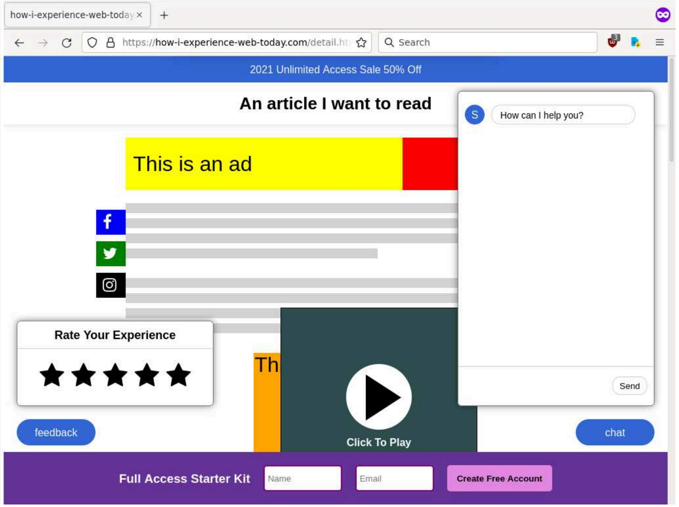
Special offer:
Get unlimited access for $1 a week.

~~$4.25~~ $1/week

Billed as ~~$17~~ $4 every 4 weeks for one year
You can cancel anytime.

SUBSCRIBE NOW

PayPal


how-i-experience-web-today ×
https://how-i-experience-web-today.com/detail.ht
Search

2021 Unlimited Access Sale 50% Off

An article I want to read

This is an ad

How can I help you?

Send

Rate Your Experience

★★★★★

Th

Click To Play

feedback

chat

Full Access Starter Kit Name Email Create Free Account

motivation

- individual annoyances
  - pop-ups and overlays
  - deceptive interfaces
  - AI-generated content
  - tracking scripts making your laptop hot!
- societal harms
  - misinformation
  - addictive usage patterns
  - dystopia?

possible interventions

- privacy and anonymity tools: Tor, PGP, ...
- browser extensions: ad and script blockers, interface enhancements
- alternative platforms: Mastodon, Bluesky, ...
- “middleware” – Fukuyama et al., 2020
- policy: GDPR, CCPA, penalties for dark patterns, anti-trust

user control

- power to make software behave as you wish
- software facilitates rather than obstructs tasks
- choice of alternatives vs. direct modification
- non-programmers delegate to an entity they trust

the web and user control

- HTTP is an open protocol
- free and open source browsers available
- browsers offer customization, but complexity hinders independent control
- HTML documents → JS web applications

data co-ops

- “infomediaries” – Hagel et al., 1997
- data co-ops – 2020s
  - organizations representing the interests of users
  - Pentland and Hardjono: emphasis on pooling personal data to enable processing that benefits members
  - Ligett and Nissim: more open-ended; manage “data flows between users and platforms.”

data co-ops – our conception

- membership organization supporting user-controlled client software

- could provide hosting and communication services, a forum for discussion of tech norms

- funding: donations have minimal overhead; dues align incentives and encourage participation

- PURE and PCSI: common standards for client software allow members to switch co-ops or form their own, improving accountability

related work on data issues

- data leverage – Vincent et al., 2021
  - data strikes
  - data poisoning
  - conscious data contribution
- data as labor
  - Arrieta-Ibarra et al., 2018
  - Posner and Weyl, 2018: “data unions”; focus on monetary compensation for data

user-provider dynamic

- implicit negotiations:
  - free services must monetize usage
  - monetizing usage degrades user experience
  - UX must be good enough to retain users; from then on, maximize revenue
  - users’ only recourse is to leave

- rebalance dynamic in favor of users

  PURE: selectively direct usage

  PCSI: take over the interface

Platform for Untrusted Resource Evaluation

- labels assign attributes to resources
- untrusted label sources scored against trusted label sources
- client-side interfaces – PURESearch

C. Malchik and J. Feigenbaum, “Toward User Control over Information Access: A Sociotechnical Approach,” in Proceedings of the 2022 ACM New Security Paradigms Workshop

PURE labels

$$\langle \mathit{source}, \mathit{target}, \mathit{attribute}, \mathit{value}, \mathit{type} \rangle$$

⟨
root,
https://engineering.yale.edu/
academic-study/departments/
computer-science,
noscriptcompat,
0,
specific
⟩

⟨

caleb,

`https://cs.yale.edu/homes/jf/`,

noscriptcompat,

1,

generic

⟩

label source tiers

0	root
1	trusted friends, local co-op
2	other co-ops, unreliable friends, companies, governments

label processing

Expectation(*target, attribute*)

Reputation(*source*)

Expectation

$$E(t,u) = \frac{\sum_{s \in S_{tu}} v_{stu} R(s)}{\sum_{s \in S_{tu}} R(s)}$$

S_{tu} sources with labels for target t and attribute u in the highest tier that is not agnostic for t and u

v_{stu} value given by source s for t and u

Reputation

$$R(s) = 1 - \frac{\sum_{(t,u) \in L_s} |v_{stu} - \lfloor E(t,u) + 0.5 \rfloor| * |E(t,u) - 0.5|}{\sum_{(t,u) \in L_s} |E(t,u) - 0.5|}$$

L_s (t, u) pairs such that there are labels for target t and attribute u from source s and at least one source in a higher tier

accounting for generic labels

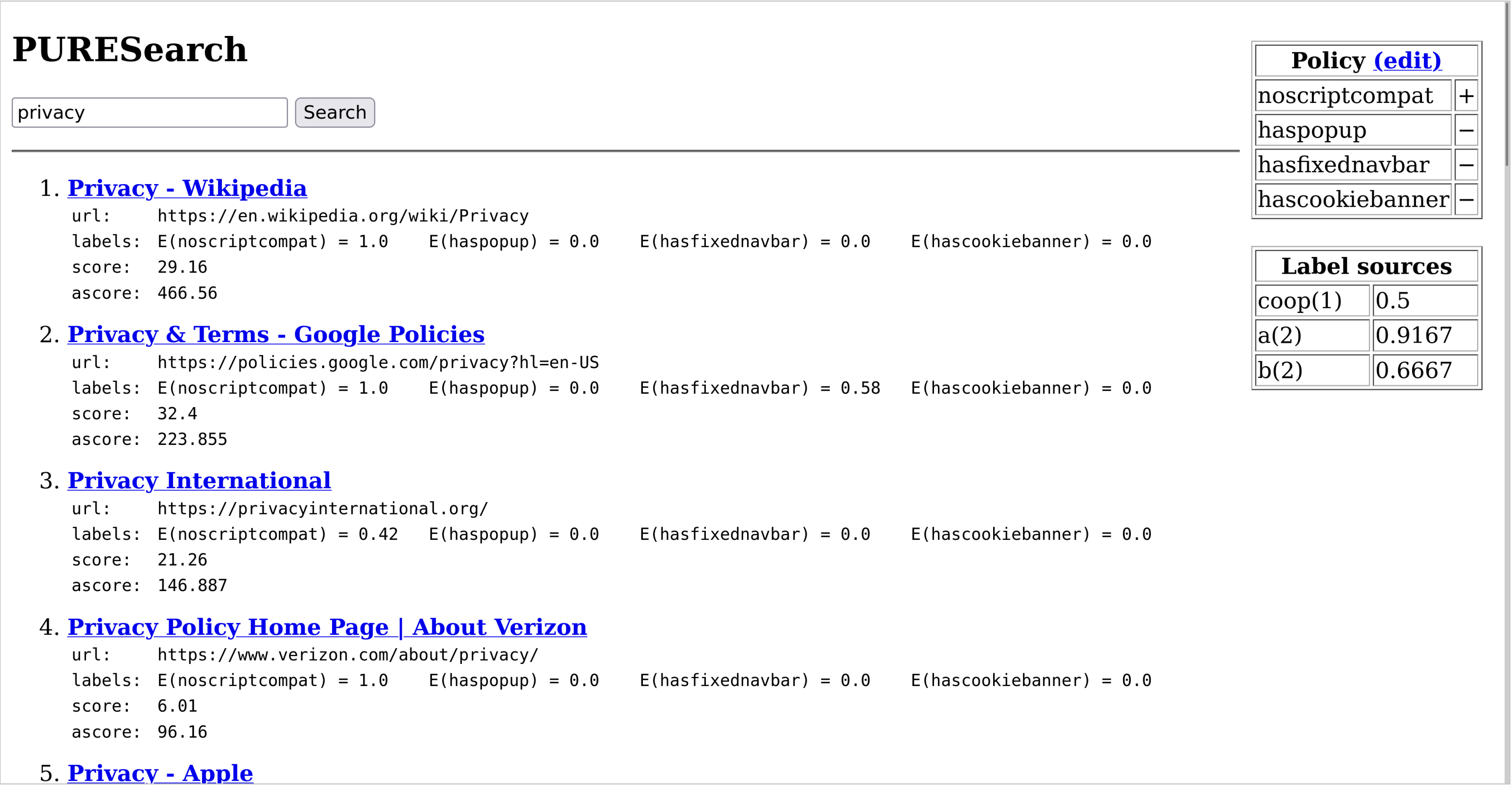
- three classes of comparison given the same weight:

  specific – specific
  generic – generic
  specific – generic

- create “virtual” labels to enable comparisons between specific and generic labels

- care necessary to ensure a reputation depends only on real labels from higher tiers

PURESearch



PURESearch

privacy | Search

1. Privacy - Wikipedia

```
url:    https://en.wikipedia.org/wiki/Privacy
labels: E(noscriptcompat) = 1.0    E(haspopup) = 0.0    E(hasfixednavbar) = 0.0    E(hascookiebanner) = 0.0
score:  29.16
ascore: 466.56
```

2. Privacy & Terms - Google Policies

```
url:    https://policies.google.com/privacy?hl=en-US
labels: E(noscriptcompat) = 1.0    E(haspopup) = 0.0    E(hasfixednavbar) = 0.58   E(hascookiebanner) = 0.0
score:  32.4
ascore: 223.855
```

3. Privacy International

```
url:    https://privacyinternational.org/
labels: E(noscriptcompat) = 0.42   E(haspopup) = 0.0    E(hasfixednavbar) = 0.0    E(hascookiebanner) = 0.0
score:  21.26
ascore: 146.887
```

4. Privacy Policy Home Page | About Verizon

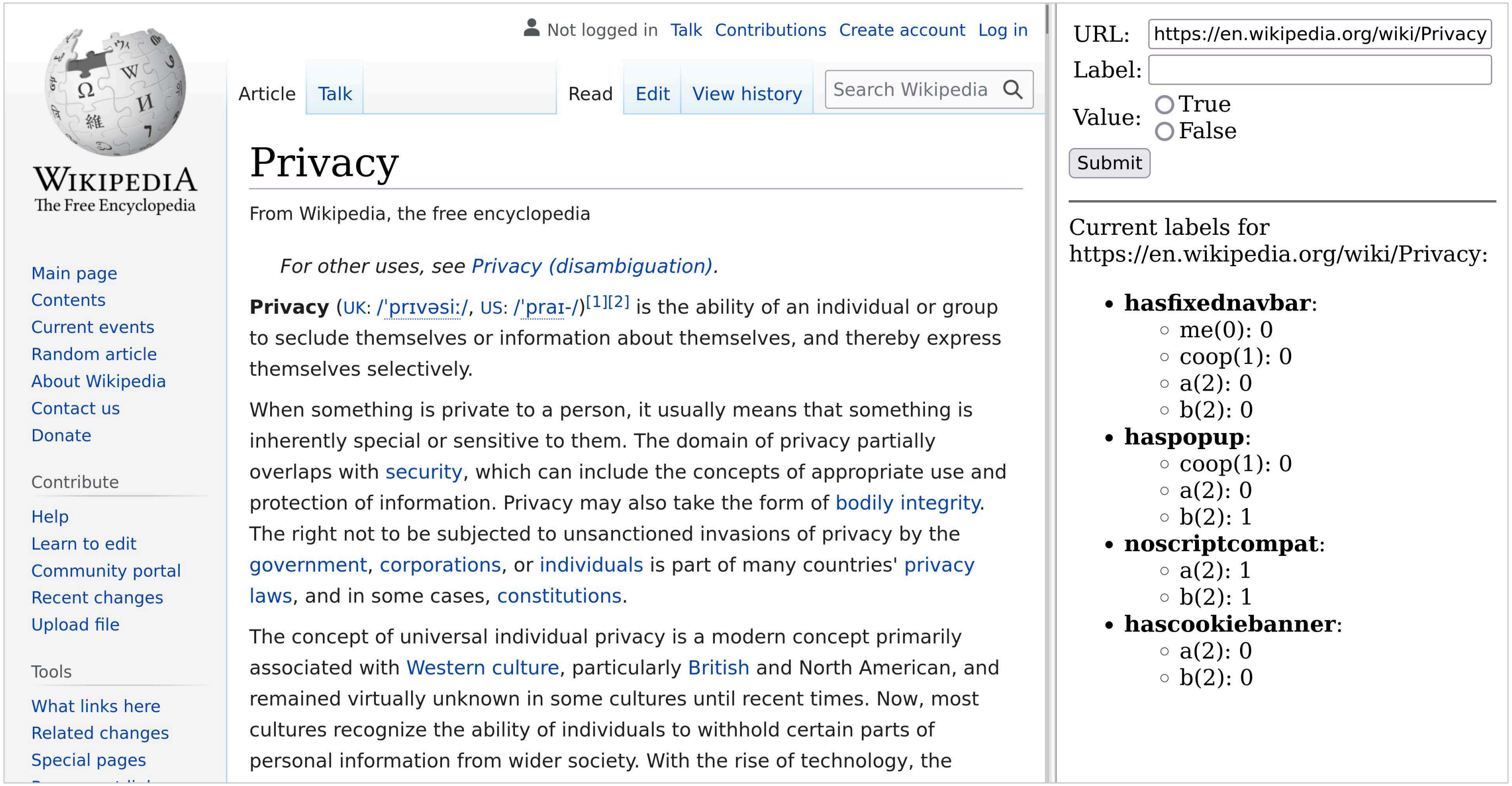
```
url:    https://www.verizon.com/about/privacy/
labels: E(noscriptcompat) = 1.0    E(haspopup) = 0.0    E(hasfixednavbar) = 0.0    E(hascookiebanner) = 0.0
score:  6.01
ascore: 96.16
```

5. Privacy - Apple

Policy (edit)	
noscriptcompat	+
haspopup	−
hasfixednavbar	−
hascookiebanner	−

Label sources	
coop(1)	0.5
a(2)	0.9167
b(2)	0.6667

PURESearch



Privacy

From Wikipedia, the free encyclopedia

For other uses, see Privacy (disambiguation).

Privacy (UK: /ˈprɪvəsiː/, US: /ˈpraɪ-/)[1][2] is the ability of an individual or group to seclude themselves or information about themselves, and thereby express themselves selectively.

When something is private to a person, it usually means that something is inherently special or sensitive to them. The domain of privacy partially overlaps with security, which can include the concepts of appropriate use and protection of information. Privacy may also take the form of bodily integrity. The right not to be subjected to unsanctioned invasions of privacy by the government, corporations, or individuals is part of many countries' privacy laws, and in some cases, constitutions.

The concept of universal individual privacy is a modern concept primarily associated with Western culture, particularly British and North American, and remained virtually unknown in some cultures until recent times. Now, most cultures recognize the ability of individuals to withhold certain parts of personal information from wider society. With the rise of technology, the

URL: https://en.wikipedia.org/wiki/Privacy

Label:

Value: ○ True ○ False

Current labels for https://en.wikipedia.org/wiki/Privacy:

- **hasfixednavbar**:
  - me(0): 0
  - coop(1): 0
  - a(2): 0
  - b(2): 0
- **haspopup**:
  - coop(1): 0
  - a(2): 0
  - b(2): 1
- **noscriptcompat**:
  - a(2): 1
  - b(2): 1
- **hascookiebanner**:
  - a(2): 0
  - b(2): 0

Performance of `purerep` on a 2008 2.6GHz Core 2 Duo T9500

# labels (thousands)	times (s)	memory usage (KB)
50	0.388	62301
100	0.802	125077
500	4.524	667206
1000	9.146	1292594
5000	43.265	6123148

use cases

- information quality
- accessibility
- client-software diversity
- user-hostile design patterns

key ideas

- grassroots solution
- allow use of established services
- client-side processing
- minimize external dependencies

Platform for Content-Structure Inference

- user controlled interfaces to non-interactive content
- extract content from HTML using MOHAWK*
- content integrity maintained via PCSI records

* formerly known as Hex

loyal clients

- “Three-Legged Stool” manifesto

  – UMass iDPI, 2023

- Gobo: access multiple social media feeds through a single client application

- what makes a client loyal?

principles for user controlled applications

- client-server independence
- implementation simplicity
- rational information structures

network effects

- improved interfaces not worth it if content isn’t there
- Doctorow: cooperative, indifferent, and adversarial interoperability
- client-driven interoperability

structured content objects

```
(article
 (headline "BBC complains to Apple \
over misleading shooting headline")
 (date "1734112342")
 (author "Graham Fraser")
 (body ...))
```

S-expressions, “draft-rivest-sexp-00.txt”, 1997

structure content objects

```
(body
 (image
  (url "https://ichef.bbci.co.uk/....webp")
  (caption "Luigi Mangione is accused..."))
 (paragraph "The BBC has complained...")
 (paragraph "Apple Intelligence..."
  (link "https://www.apple.com/...")
  ", uses artificial intelligence (AI)...")
 ...
 (subheading "’Embarrassing’ mistake")
 (paragraph "Apple says...")
 ...)
```

processing HTML with MOHAWK

- a modest extension to AWK
- parse input as HTML into DOM tree before running program
- new built-in functions for traversing the tree

new built-in functions in MOHAWK

Function	Value Returned
root()	the ID of the root node (typically 1)
parent(*n*)	the ID of the parent of node *n*
sister(*n*)	the ID of the next sibling of node *n*
children(*n*)	the ID of the first child of node *n*
type(*n*)	the type of node *n*, expressed as a string: one of ELEMENT, TEXT, COMMENT, DECLARATION, PROCINS, or ROOT
name(*n*)	the name of the HTML element at node *n*
text(*n*)	the text contents of node *n*, where *n* is a text node
attr(*n*,*s*)	the value of the *s* attribute in the HTML element at node *n*
selmatch(*n*,*s*)	1 if the CSS selector *s* matches node *n*, 0 otherwise
seconds(*s*)	the number of seconds since 00:00:00 GMT, Jan. 1, 1970 corresponding to *s* interpreted as a date string

a MOHAWK script for BBC.com (excerpt)

```
function walk(n,    json, s) {
      if (!n) return
      if (name(n) == "h1") {
            only("headline", text(children(n)))
      } else if (selmatch(n,
         "main#main-content article div p")) {
            A["body"] = A["body"] paragraph(n)
      } else if (selmatch(n,
         "main#main-content article div h2")) {
            A["body"] = A["body"] subheading(n)
      } else if (selmatch(n,
         "main#main-content article figure")) {
            A["body"] = A["body"] image(n)
      }
      walk(children(n))
      walk(sister(n))
}
```

PCSI records

```
(rule
 (source |WRlcbFQcgwfx2i0edo1vIo
          DJhN8hetX0xkw1QrBBEaQ=|)
 (timestamp "1735608269")
 (pattern "https?://(www\\.)?bbc\\.com/\
news/articles/[0-9a-z]+")
 (script-hash |S0xCR9rATD+mot/oamL3HX
               9FCTSyoqS5YldaEsW0giU=|)
 (object-type article))
```

PCSI records

- *rule*

  “if a URL matches this pattern, try running this script”

- *inference*

  “here’s how I got this object”

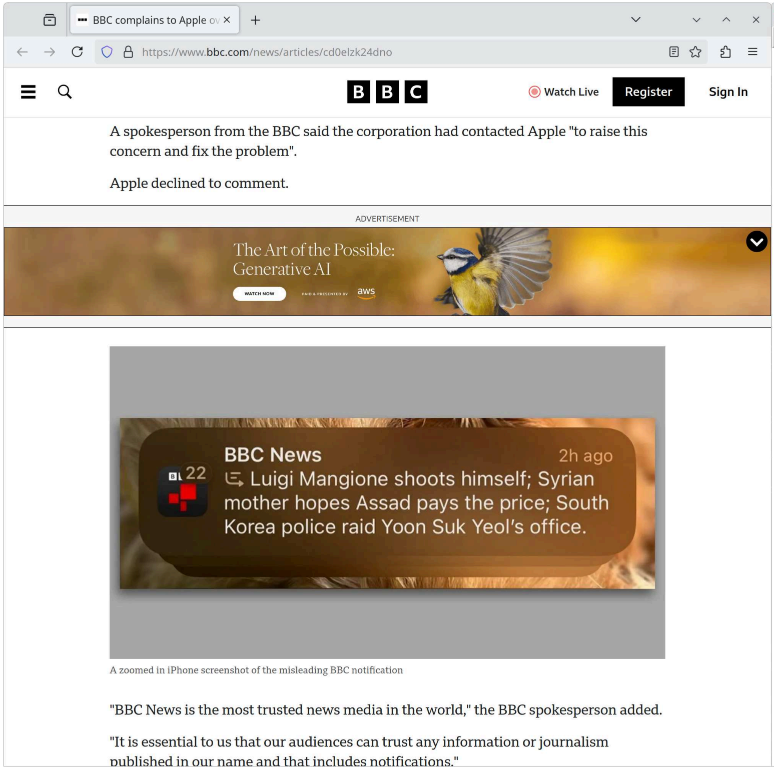
- *perception*

  “this object matches this URL (or not)”

PCSINews

- RSS reader for news articles
- uses PCSI records to display articles in the reader

PCSINews


BBC complains to Apple ov
https://www.bbc.com/news/articles/cd0elzk24dno
BBC
Watch Live
Register
Sign In
ADVERTISEMENT
The Art of the Possible: Generative AI
WATCH NOW
PAID & PRESENTED BY aws

A spokesperson from the BBC said the corporation had contacted Apple "to raise this concern and fix the problem".

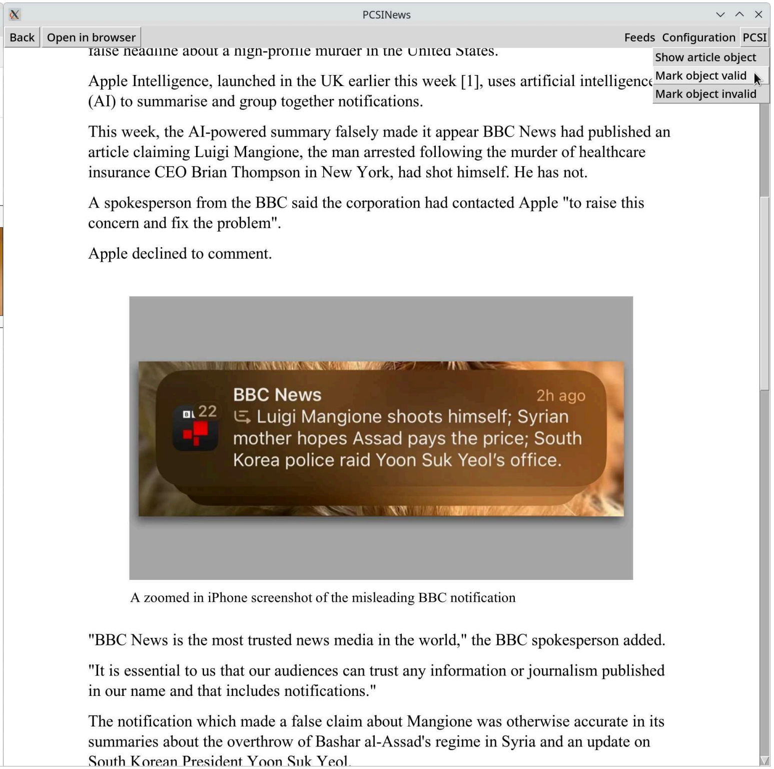
Apple declined to comment.

BBC News 2h ago
22
Luigi Mangione shoots himself; Syrian mother hopes Assad pays the price; South Korea police raid Yoon Suk Yeol's office.

A zoomed in iPhone screenshot of the misleading BBC notification

"BBC News is the most trusted news media in the world," the BBC spokesperson added.

"It is essential to us that our audiences can trust any information or journalism published in our name and that includes notifications."



false headline about a high-profile murder in the United States.

Apple Intelligence, launched in the UK earlier this week [1], uses artificial intelligence (AI) to summarise and group together notifications.

This week, the AI-powered summary falsely made it appear BBC News had published an article claiming Luigi Mangione, the man arrested following the murder of healthcare insurance CEO Brian Thompson in New York, had shot himself. He has not.

A spokesperson from the BBC said the corporation had contacted Apple "to raise this concern and fix the problem".

Apple declined to comment.

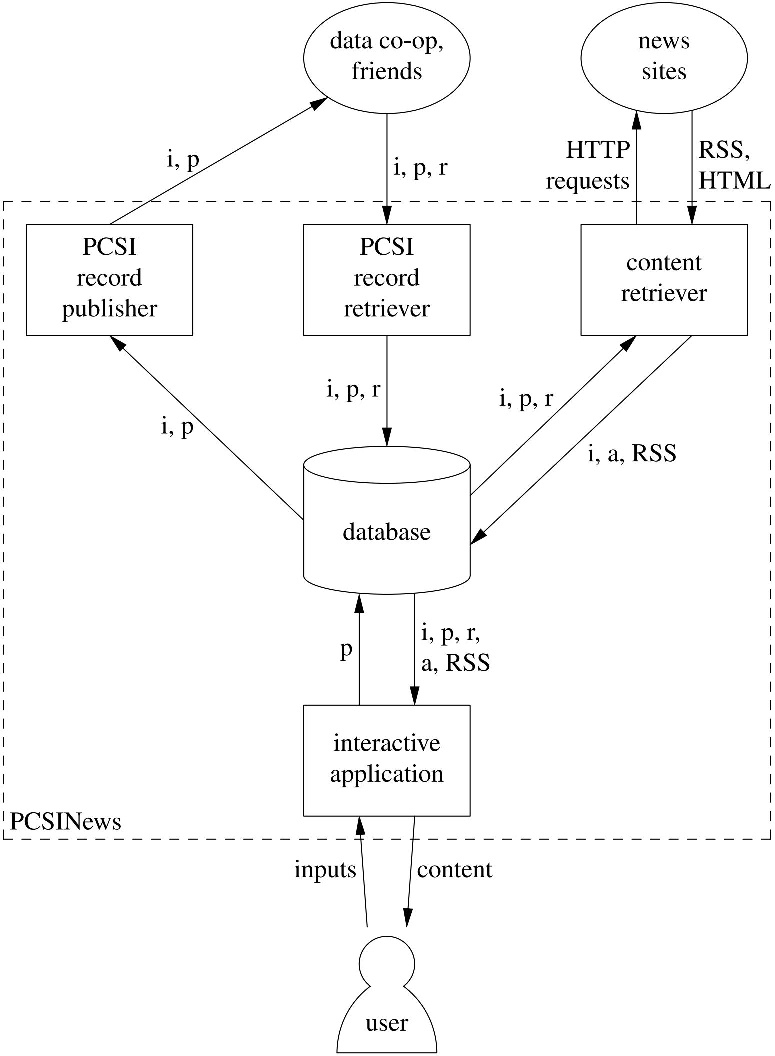
BBC News 2h ago
22
Luigi Mangione shoots himself; Syrian mother hopes Assad pays the price; South Korea police raid Yoon Suk Yeol's office.

A zoomed in iPhone screenshot of the misleading BBC notification

"BBC News is the most trusted news media in the world," the BBC spokesperson added.

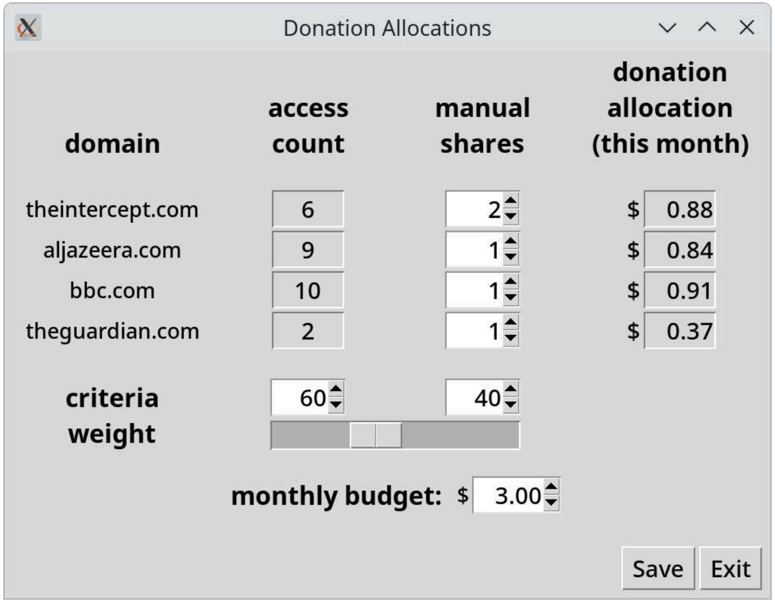
"It is essential to us that our audiences can trust any information or journalism published in our name and that includes notifications."

The notification which made a false claim about Mangione was otherwise accurate in its summaries about the overthrow of Bashar al-Assad's regime in Syria and an update on South Korean President Yoon Suk Yeol.


data co-op, friends
news sites
i, p
i, p, r
HTTP requests
RSS, HTML
PCSI record publisher
PCSI record retriever
content retriever
i, p, r
i, p, r
i, p
i, a, RSS
database
p
i, p, r, a, RSS
interactive application
PCSINews
inputs
content
user

incentives for publishers

- PCSI removes many potential revenue streams
  - ads
  - behavioral data collection
  - nudges to subscribe or donate
- publishers incentivized to prevent scraping
- support publishers directly
  - data co-ops allocate a portion of dues to support publishers
  - allocations automatically calculated by client software; user can override



Donation Allocations

domain	access count	manual shares	donation allocation (this month)
theintercept.com	6	2	$ 0.88
aljazeera.com	9	1	$ 0.84
bbc.com	10	1	$ 0.91
theguardian.com	2	1	$ 0.37
criteria weight	60	40	

monthly budget: $ 3.00

user study

- 10 participants who use RSS to read news
- 7 with programming experience, 3 without
- 1 h. 15 min. *user* phase

  use PCSINews, marking validity of content objects
- 1 h. 15 min. *programmer* phase

  write a MOHAWK script

results

- 2 of 7 programmers completed the task; almost all made significant progress

- *“If I could pay $2 a month for one left source and $2 a month for one right source, that would be perfect.”*

- 10% liked the system so much that they volunteered to help

  “In my mind it’s somewhat like a translation layer between the modern internet and the user’s needs.”

overall strategy

- software supported by organizations, but independent of any particular organization
- divide system into components that can be easily produced and maintained
- benefit users while influencing service providers
- begin with minimal functionality to make small organizations viable
- grow in many dimensions

relation between PURE and PCSI

- analogy between PURE labels and PCSI records
  - rule records kind of like generic labels
  - perception and inference records kind of like specific labels
- PURE → PCSI
  - increased complexity of record semantics
  - more trust required of remote sources
  - direct control rather than indirect influence

future work

- form a data co-op
- improved tools for record production, scraping
- naming and content discovery
- interactive use cases

thank you for coming

`https://cs.yale.edu/homes/cmalchik/`