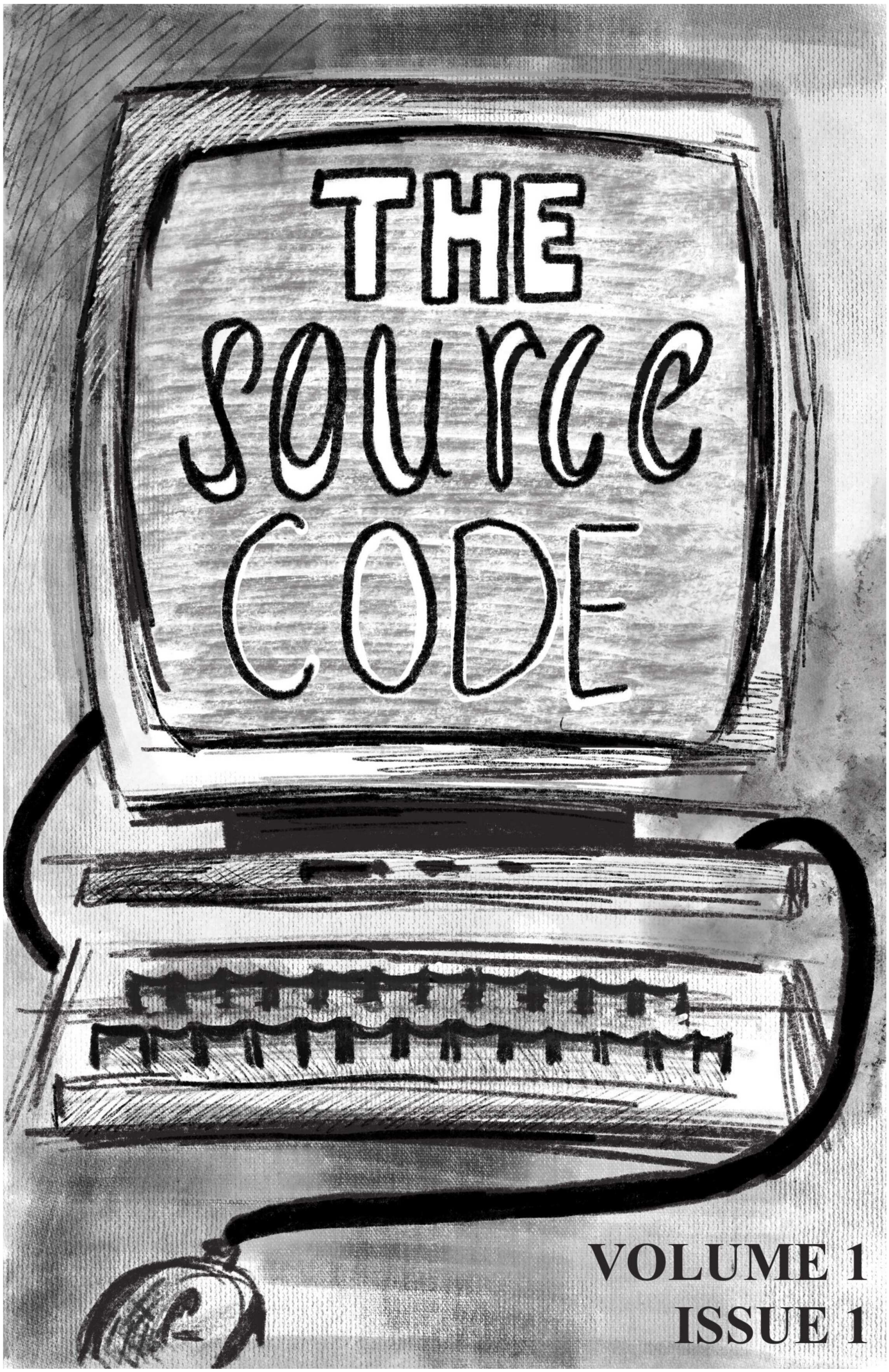



VOLUME 1
ISSUE 1

# Masthead

## Editor in Chief
Pooja Mangra

## Managing Editor
Shana Rosenberg

## Marketing Manager
Isla Parekh

## Editors
Seraphina Vaillancourt, Elise Corbin, Stefan Barna

# Not a Computer Scientist A Reflection

I was fifteen when I decided I would study computer science. I don't even remember why. I guess it had something to do with watching the Social Network and thinking Jesse Eisenberg was cute. Or it might have been because the billionaire techies all reminded me a bit of Iron Man. Whatever the reason was, I decided I would learn to code, get a job at Google or some other Silicon Valley powerhouse and live happily ever after.

But by the time I graduated high school, that plan had gotten a little foggy. Sure, I loved learning about computers, but I spent my spare time with oil paints and poetry. I dreamed of joining a rock band, writing literary fiction, designing haute couture. And when I looked around my CSC110 class, at the geniuses who spent their childhoods learning Java and winning hackathons, I felt so lost.

I would go to the registrar's office and ask, "Are there CS students who didn't like first year but grew to enjoy the program?" and she'd give me the half smile you give to a toddler checking if anyone saw them trip.

To be honest, a big reason I stayed in CS that first year was because I knew it was something a lot of people wanted. And surely it meant something that I had it.

But in the winter of my second year, it finally clicked for me. I was taking courses that reminded me why I liked computers. There is something so beautifully complicated about a computer that is always alluded to. And that winter, I was sitting with the enigma and finally seeing her — how she was made, how she worked. And like a giggling school girl, I was completely lovestruck.

I felt the same satisfaction learning about cache design as I did reading personal essays or viewing a self portrait. Almost like I could see the brushstrokes of the computer scientist in the design of the computer. My relationship with CS became an intimate one that burned slowly in lecture halls and quiet nights at the library. It pushed me to experiment and learn, not for my resume or validation from my peers, but for the romance I was brewing with the material.

Although I've grown into a Linux using, RSS feed curating, avid reader of How-to-Geek and Slashdot, I don't think I'll ever feel like a computer scientist in the way that my Tech Bro idols do. But my failure (or reluctance) to fit the CS mold is what gives me the freedom to love a little differently. There's less pressure when you call yourself an admirer of computers, a fan, just along for the ride. I want you to know that it's okay to be a little different, even when it feels isolating and horrific. You'll figure it all out in time.

With love,

Pooja Mangra ♡

# Where's My Water

## How Artificial Intelligence Continues to Put a Strain on Our Fresh Water Supply

Fresh water is a scarce resource. Of the more than one billion cubic kilometers of water on our planet, only 3% is fresh, and of that 3% only about one-sixth of it is readily available for consumption. Millions of people globally already lack sufficient access to safe drinking water. As the global demand for artificial intelligence (AI) continues to grow, so does AI's demand for the fresh drinking water we so desperately need.

Many people are probably surprised to hear that AI uses water, period. After all, AI is *artificial*, and doesn't have any biological processes that require water like we do. In fact, the AI models popularized today run out of massive data centers scattered across the globe, which *do* require water to function. The servers running the models use enormous amounts of electricity to perform. This electricity produces a lot of heat that–if left unchecked–would damage the computers. Data centers use technologies like cooling towers to cool themselves down. These cooling towers circulate cold fresh water into the rooms housing the servers, which absorbs the heat and is then evaporated off. Why *fresh* water? It's a maintenance and cost saving measure: water with chemical or microbial contaminants has the potential to damage plumbing and technical equipment.

Data centers are widely distributed globally, and are especially present in countries like the United States and Canada.

However, tech companies are increasingly turning their gaze to the Global South, specifically to countries in Latin America and Africa, attracted by the cheaper costs of labor and materials. The governments of these countries also see an opportunity to enter the AI market, expand high-speed internet access, and create jobs. It is important to note that many of these nations are located in water-stressed areas of South and Central America and the Middle East. Moreover, data centers are notorious for their high ratio of resource consumption to job creation. To cut costs, many companies also find ways to construct and operate data centers without paying taxes to these nations' governments. Thus, they use massive quantities of public resources while providing next to nothing of value to the public. In response, the citizens of these countries have been anything but ambivalent. Many organizations are actively taking legal and grassroots action to try and prevent the construction of data centers in their communities.

These issues, however, are not confined to the Global South. Bloomberg News found two-thirds of data centers built or developed in the U.S. since 2022 are in places dealing with high levels of water stress. This has very real consequences for the lives of people living near data centers. They are suffering at the hands of our thirst for AI. In 2016, Beverly Morris moved into what she thought would be her new forever home in rural Georgia. Soon after, Meta began construction of a supermassive data center a mere 400 meters away from her property. Morris' new neighbors were anything but gracious. Soon after Meta's arrival, Morris found her home's water pressure drop to next to nothing. Even worse, her tap water began running orange-brown due to an excess of dissolved sediments contaminating the private well that supplied water to her property. Despite this, she still uses the water to cook, clean, and brush her teeth, feeling powerless to do anything to ameliorate her situation. Meta conducted a survey of the groundwater near their data center, claiming the construction of it "did not adversely affect groundwater conditions in the area." Gordon Rogers, the executive director of Flint Riverkeeper in Georgia, is suspicious. He took BBC reporters to a creek downstream of another data center in Georgia. The water in that creek was a similar colour to Morris' and contained high concentrations of flocculants, which are used to bind soil and prevent erosion during construction, but can be easily leached into groundwater.

As we look into our collective future, AI will no doubt continue to be a part of our lives. What this means for its water consumption is unclear, though researchers have made some startling projections. Probably one of the most surprising predictions is that–by 2030–the average AI consumption of an

OECD-European individual is projected to consume roughly three liters of water per day, more than the average amount consumed by drinking. So the question then becomes, what can we do about it?

The first and most important step is increasing transparency. Unlike electricity use and greenhouse gas emissions, many companies are not required to report their precise water use. Without this information, not only are environmental impact studies nearly impossible to conduct, but it is also a futile endeavor to try and regulate AI companies without receipts to hold them responsible for. The European Union's relatively recent AI Act has been a huge step forward: not only regulating the ways in which AI companies are and are not allowed to operate, but also mandating them to report fresh water consumption. In Canada, however, water regulation is a provincial responsibility as opposed to a federal one, making unilateral action difficult. There has nevertheless been some progress. In areas of Saskatchewan and Alberta that have experienced water shortages, some lawmakers have begun to push for regulatory changes. Recently, Quebec has also passed legislation requiring major fresh water users to report consumption. However, it's limited to water drawn from a natural body, like a river or lake, instead of a utility.

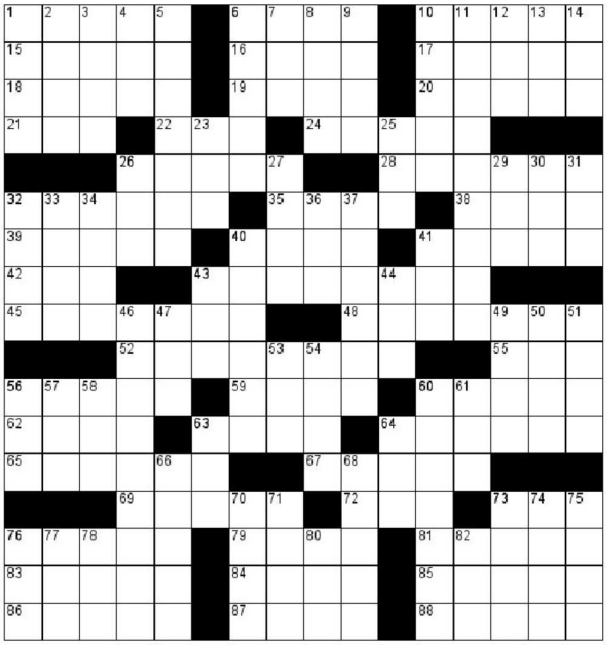
Ultimately, it is clear AI is here to stay. Other than increasing transparency, the best actions governments and tech companies can take is work towards more efficient cooling technologies. At the individual level, my message is this: vote with your dollar and your attention. The less willing we are to use AI models, the more the companies developing them will have to listen to us and our concerns.

Written By:
Jake Cohen

# I'm Doing It Again (and again and again) Baby!

**Elise Corbin**

*This puzzle originally appeared on my blog in 2024, but I've changed the fill and clues. It pays homage to a useful concept in computer science, but non-CS majors will be able to enjoy it too!*





## ACROSS

**1.** You might put a 39-Across on it
**6.** Organism made up of many 19-Acrosses
**10.** Person more developmentally mature than a 41-Across
**15.** ___ Doone (cookie brand)
**16.** Language with the place names "Manitoba" and "Saskatchewan"
**17.** Civil rights activist Desmond who appears on the $10 bill
**18.** Strongly advised
**19.** Organism made up of many 40-Acrosses
**20.** Actor Hawke
**21.** Queen ___ (nickname for a pop icon)
**22.** Emotion experienced by a comic book character with steam coming out of their ears
**24.** Not together
**26.** ___ B (rap icon)
**28.** Credit in a magazine
**32.** Conduits for elevators
**35.** Police procedural with spinoffs in many major U.S. cities
**38.** "Artemis Fowl" author Colfer
**39.** You might put a 56-Across on it
**40.** [BASE CASE] Fundamental unit of matter
**41.** Person more developmentally mature than a 60-Across
**42.** Gambling parlor letters
**43.** Shakespearean sorcerer who says "We are such stuff as dreams are made on"
**45.** Facial hair for a cat
**48.** Craft requiring special paper sometimes called "kami" (which just means "paper" in Japanese)
**52.** "You can count on me!"
**55.** Japanese honorific
**56.** You might put a 86-Across on it
**59.** What you might do after a long 79-Across
**60.** Person more developmentally mature than an 88-Across
**62.** Danceable tunes
**63.** 90° from norte and 270° from sur
**64.** Pleasantly warm
**65.** Horoscope app that "deciphers the mystery of human relations through NASA data and biting truth"
**67.** Attire on the Great British Bake Off
**69.** Like some online purchases
**72.** Feel ill
**73.** Acronym after "I'm a lobster diver who recently survived being inside of a whale" and "Hey! I was born with perfect polydactyly"
**76.** "Adios, ___!"
**79.** What you might do after a long 87-Across
**81.** Peripheral
**83.** Expressions of content or exasperation
**84.** Willie ___ (first Black NHL player, for whom the league named their Community Hero Award)
**85.** Barely noticeable
**86.** [BASE CASE] Vehicle for a meal
**87.** [BASE CASE] About 1/52 of a year
**88.** [BASE CASE] Unborn person

## DOWN

**1.** Totally blow it
**2.** ___ drop (interesting story about your life)
**3.** Not-so-conservative party?
**4.** Anything to the power of zero (except zero or infinity)
**5.** Emit
**6.** Secretly included on an email
**7.** Iron taken from a vein, e.g.
**8.** Hip-hop trio ___ Soul
**9.** Google Reviews alternative
**10.** First word in a musical mnemonic
**11.** Living accessory for many Yorkville residents
**12.** "___ Canada" (2019 Simpsons episode in which Lisa falls down Niagara Falls and is granted asylum in Canada)
**13.** Portuguese "she"
**14.** Didn't stick around
**23.** Monopoly properties that don't get hotels, for short
**25.** Muscles strengthened by planks
**26.** Sports org. with the Toronto Argonauts and the Montreal Alouettes
**27.** Sweet on
**29.** Promise to pay
**30.** Zero
**31.** Treebeard or Beechbone, in Middle-earth
**32.** Put in the overhead bin
**33.** "Hell ___ no fury ..."
**34.** Actress Jacobson who co-created and starred in "Broad City"
**36.** sin/tan
**37.** Be inconvenient
**40.** What a quiver holds
**41.** "When somebody says, ___, I think you've totally changed" ("Sympathy is a knife" remix lyric)
**43.** The "p" of 47-Down
**44.** Before, in poetry
**46.** Waits patiently
**47.** No. on a speedometer
**49.** Phrase on a yard sale tag
**50.** Relative of a milkshake
**51.** ___ 500 (car race held at a motor speedway in Speedway, Indiana)
**53.** Yoga necessity
**54.** Company whose contingent at Toronto Pride this year waved Blahajs attached to sticks
**56.** Radio-Canada's English counterpart
**57.** Water closet
**58.** Follower of black or special
**60.** Go jump in a lake, maybe
**61.** World's largest ethnic group, making up about 17.5% of the global population
**63.** Period
**64.** Prefix with angle or cycle
**66.** By ___ (just barely)
**68.** ___ Philippe (Swiss watchmaker)
**70.** Front of a ship
**71.** French father
**73.** According to Dictionary.com: "Vigorously pursuing an activity, especially a fight, but also sex or some other activity"
**74.** "Dropdown" thing on a website
**75.** Pottery and illustration, for example
**76.** ___ caterpillar (moth also called the puss caterpillar, woolly slug, and opossum bug)
**77.** 1,000 G's
**78.** Tennis star Swiatek who made pasta with strawberries famous this summer
**80.** "I told you!"
**82.** Nation whose currency is the dirham, for short

# Remember when Google tried building a smart city in Toronto?

At first glance, the piece of land known as Quayside seems like any other underdeveloped stretch of waterfront in Toronto. Situated at the foot of Parliament Street and named for its proximity to Queens Quay, Quayside was once home to fish and soybean processing plants, but it's been a desolate parking lot for decades. Around the late 2010s, though, Quayside almost became much more than a parking lot. When Sidewalk Labs, an urban-planning-focused division of Google, drew up a plan to transform Quayside into a "smart city", Toronto was suddenly at the forefront of debates about tech companies' role in civic life. The project was cancelled in 2020, but looking back, Quayside has a lot to teach us about the long history of the authoritarian tendencies present in Big Tech today.

The story of Quayside development begins with Sidewalk Labs, which started as a pet project of Google co-founder Larry Page, In the mid-2010s, Sidewalk teamed up with nonprofit Waterfront Toronto to build a futuristic city on the Quayside land. Years of pamphlets and press conferences followed, during which the public learned, bit by bit, of Sidewalk's plans. Their city would be carbon-neutral, made with sustainable building material and running on clean energy. They would eliminate extreme weather complications with retrofitted buildings and roads. And, perhaps most alarmingly, they would collect data on almost every aspect of residents' lives through a "ubiquitous community network" of sensors, and they explicitly stated that one of their main goals was to let "third parties" use this data to "build new services".

These plans rang every alarm bell for Toronto ethical tech advocate Bianca Wylie. To Wylie, the vicious cycle of using sensors to inform new tech that would then use even more sensors sounded like it could spiral out of control quickly, before the government had a chance to regulate it. She'd already seen this happen with the rise of smartphones, when Big Tech companies insisted they needed to collect data on citizens to improve apps people could no longer live without, and then the companies used this data to improve their predictions about users' lives. But you can always turn off social media or put down your smartphone. This time around, the data being collected would involve every aspect of your life, and there would be no way to totally opt out. In Quayside, even opting for a little more privacy could have negatively affected your quality of life: Sidewalk's original internal plan for Quayside, known as the "Yellow Book", proposed tiered access to neighbourhood amenities, including finances, based on how much data about themselves they were willing to share.

Other critics worried about the effect of the Sidewalk partnership on city governance: as a division of Google, a foreign corporation, Sidewalk was not beholden to the city of Toronto, so how could Torontonians hold Sidewalk accountable? As Bianca Wylie pointed out in a city council meeting early in the project's planning stages, the city was not only throwing away all the money that its residents' data was worth, it was also giving away control of that data to a third party that specifically stated it would sell that data to other third parties for profit.

By agreeing to the partnership with Sidewalk, the city of Toronto was even giving away its own citizens' consent to be experimented on. From the beginning of the project, Sidewalk was insistent that they eventually wanted to expand their smart city technology into the nearby Port Lands, which was beyond Waterfront's jurisdiction and considered part of the city proper. Using unsuspecting Torontonians as guinea pigs for their surveillance projects—people who didn't even choose to live at Quayside—was a

nonstarter for Waterfront executives, and it may have been the initial kernel of doubt that led them to eventually listen to critics' concerns and push Sidewalk harder on data ethics.

Bianca Wylie and her fellow data rights critics were often considered the leaders of the anti-Sidewalk contingent in Toronto, but the Association of Community Organizations for Reform Now (ACORN) was leading the movement against another of Sidewalk's regressive policies: the lack of affordable housing planned for Quayside. Sidewalk repeatedly bragged about how roughly 20 percent of its residential units would be below-market, but according to the fine print, only 5 percent would go towards what Sidewalk called "deep affordability needs". (The market price for waterfront properties is so high that "deep affordability needs" housing would likely

be the only affordable place to live in Quayside.) Another 20 percent of units would be middle-income housing, and the remaining 60 percent would be condos. ACORN organizers staged a protest inside the Waterfront offices, expecting to be thrown out. But to their surprise, the chairman of the board, Steve Diamond, listened to their concerns about affordable housing and thanked them for stopping by.

It would be dishonest to say the Quayside project fell through as the direct result of community action like this, but the discussions the community activists were having eventually made their way up to the Waterfront boardroom. Critic Jim Balsillie gained the ear of the chairman of the board, Steve Diamond, and explained his concerns about Sidewalk's plans for collecting data on citizens. Between the increasingly clear picture of an

unaccountable lakefront surveillance state and the disagreement over lands allocated to Sidewalk, the project was looking doomed. Multiple Waterfront board members quit in protest, and in fall 2019, the remaining board nearly voted to dissolve the partnership.

In the end, it was Sidewalk who pulled out of the project; they were annoyed at how difficult it was to get things done, which was partly because of how involved and engaged citizens were, and partly because Waterfront Toronto was listening to citizens' concerns by the end of the whole saga. Today, we look back on the late 2010s as a time when companies like Google seemed to have a utopian, progressive vision for the future; after all, their motto used to be "Don't Be Evil". But after all you've just read, do you really believe that not being evil was ever Google's first priority?

With the second Trump administration currying favour from the likes of Elon Musk and Mark Zuckerberg, it's become common knowledge that Big Tech has shifted to the political right. But the story of Sidewalk Labs, which happened at the height of Big Tech's popularity among liberal leaders, clearly shows that Big Tech was always on the side of profit and control over people. Even on top of their underwhelming affordable housing policies, Sidewalk's technological plans for Quayside show early tendencies toward authoritarian control that would begin showing up on the right more and more in the coming years. Sidewalk's desire to use Quayside residents' data to control their access to neighbourhood necessities and finances is classic authoritarianism; it sounds more like something coming out of the pages of a Margaret Atwood novel than the pages of

an innovative proposal for modern city life. Similarly, Sidewalk's insistence on expanding their Google-controlled "smart city" beyond the boundaries they were promised is reminiscent of Trump's recent desire to annex Canada; both thought they could buy and control their northern neighbours for economic gain. Big Tech has never worked for the people, and it was never going to save us. Now, as then, we need to resist authoritarianism with one voice. And in the case of Big Tech, no one understands that better than Torontonians.

WRITTEN BY ELISE CORBIN

# In the Stream

## Reflecting on growing up with music streaming

Ethically, there are numerous reasons to avoid music streaming services: extremely low compensation for musicians, profit-driven systems, parent companies being terrible, on and on. Lots of articles, videos, and more are out there describing this. I've recently cancelled my Spotify subscription and begun exploring different options because of reasons like these.

Now forget all that! I wanna talk about what I have more genuine, first-hand experience in: using them.

In middle school my parents bought an Apple Music subscription for the family. I began assembling large amounts of my own music in the library: creating playlists for different genres, exploring the playlists Apple created for songs I would want to keep, etc. At some point I wanted to keep my own separate library and switched to Spotify in early high school, quickly buying a subscription because the ads were annoying as hell. Spotify Premium then comprised the majority of my music discovery and listening, bar some stuff on Youtube, until today. All in all, that's nearly a decade of pure, unrestricted music streaming.


Over those years, my relationship with music developed considerably. I could spend forever detailing how my tastes evolved, wax about the different playlists I created ad infinitum, list my favourite albums across my library for perpetuity, or endlessly expound on how far I fell in love with music (and I will if I need to pad. That is a threat.) However, the hopefully more intriguing discussion I'll explore is this: how did I interact with music through streaming technology, and how did streaming technology influence how I interact with music?

One overarching thing to keep in mind is that my love and reverence for music grew over time. It became increasingly important in my life, and I cared more and more about how and what I *consumed* (I hate that word), especially as my love and skill for my instrument developed too.

The yapping shall cease. Finally. Onward with discussion.

Within the various pieces criticizing streaming platforms I read before writing would be claims that these services negatively impact what and how you listen to music. In other words, saying they make you a 'dumber' music listener, whatever that means to the author personally. For example, in the book *Mood Machine*, Liz Pelly claims that Spotify attempts to focus on context-based (e.g. time of day, activity) and mood-based (e.g. chill, focus, upbeat) music, tries to reduce user input and just feed things to listen to, and flattens art into 'vibes'. While I think it's safe to trust her research and judgment, and I can definitely see Spotify pushing for such things, I want to compare these and similar claims to my own experiences and habits. One isn't a sheep (hate that word too) for just using a technology.

I began listening to music mainly through playlists. I would periodically go through the genre playlists Apple would curate, saving the songs that I liked into the various genre playlists I was curating myself. Because the Apple library was shared with the rest of my family, the saved albums and songs were overflowing wastelands, so playlists were the easiest way to carve my little niche. As such, I only sporadically tried whole albums, and I almost never examined an artist's discography. However, even as a wee lass developing my taste in music, I would actively seek out and listen to music, rather than just throw on a radio station or a playlist generated by the streaming platform. Additionally, I primarily explored by genre, not mood or activity.

At some point in middle school, my lessons teacher told me to find a sax player I liked and just fully dive into their music, listen to them often, and try to really grasp and imitate how they play. I initially just saw this as another part of the jazz education, but through exploring Eddie Harris (one of the greatest tenors imo), my love for not only jazz but music in general skyrocketed. Music became more personal and I gained a greater awareness for the artists behind my music. As such, I created a playlist dedicated to Harris tracks, a precursor to how my habits would evolve into being more artist and album based.

When I first moved to Spotify Premium in early highschool, I started by liking all the music I wanted to keep from my old Apple playlists, then divvying them up into more specific playlists. For example, my broad jazz playlist turned into funk, fusion, traditional, etc. I primarily used these playlists and would add to them from Spotify's genre playlists, as I did with Apple. However, something I started to do was save albums. Within a year or two, albums were far outpacing playlists in my library. In the last 2 years I created 3-ish playlists, while in that time I saved upwards of 200 albums. I would very rarely use those playlists as well; liking songs and adding them to playlists was more for keeping track of singles and feeding my algorithm.

I had deep and personal connections with the playlists I created and added to them based on genre and enjoyment, and my use of albums today is much the same. In discussions with my girlfriend, who has wildly different tastes and habits to me, she described having deep connections with her music as well while using primarily playlists. As such, I feel merely using a streaming service doesn't mean your brain turns to mush and you just play generic, low-quality 'chill' music in the background 24/7. While the technology might encourage certain habits, to me the true decider is what you want and care about.

Finding new music is a murkier topic for me. People describe an ick for an algorithm deciding what music they find and listen to which I definitely understand, but Spotify's "recommended for today" album selection has shown me music I absolutely love across genres and often from niche artists (for me around 100,000 monthly listeners or less). Maybe digging through a genre's new release section, checking out curator websites and creators, or browsing a record store would help me find more music I love, but it's undeniable how well algorithms have worked for me. So, why not both? One way I've expanded past algorithm suggestions has been exploring an artist's page: delving into their discography and looking at their recommendations and playlists have not only given me great and rare music but also connect let me further with this musician I already like. Combining both human and tech approaches to music discovery and listening I feel can yield the deepest and most genuine experience, and I'm excited to further fall down the music rabbit-hole.

By Seraphina Vaillancourt

# Inside the AI Black Box

## Is Chain-of-Thought Monitoring the AI Safety Solution We've Been Looking For?

Have you ever tried to explain a decision you've made, only to come to the realization that you aren't entirely certain how you reached it? That gap between *what we do* and *why we do it* is largely paralleled in our current understanding of artificial intelligence: AI models have the capacity to produce robust responses, sometimes good and sometimes faulty, yet when we ask *why* they behave the way they do, the answer is unclear.

Historically, we've often been able to peek under the hood for the technologies we develop. An electrician, for instance, might be capable of explaining every detail of a circuit, and how it connects to observable output. AI differs. In a way, we aren't so much *building* a model as we are *growing* it. We set high level conditions and feed data in, yet we cannot in any thorough capacity predict every behaviour of the end product. And this "black box" gives rise to a major concern: if we cannot explain an AI's behaviour, how can we trust it?

The risks are easy to imagine. Without transparency, we cannot say for certain what a model can and cannot do. A model might behave it its own interest, or enable harmful uses that we only come to realise through empirical discovery. If we want AI safety, especially as a model surpasses human capabilities, we best ensure we have the tools to predict its behaviour before it is widely distributed. It is for this reason that research in transparency, or *interpretability*, is largely viewed as critical.

Recently, a promising research direction has emerged from a collaboration between researchers at Anthropic, OpenAI, Google DeepMind, and others, which relies on monitoring *chains of thought* (CoT).

To explain, let's consider the way we ourselves tackle multi-step math problems. For instance, the prompt "Alice is 10 years older than her sister. If 5 years ago, she was double her sister's age, how old is she now?" likely does not induce an immediate final answer in your mind. That is, of course, unless you're remarkably well-trained in this class of problems for whatever reason. Instead, you might determine how old Alice must have been 10 years ago, and from there determine her current age. This sequence of steps is called your *chain of thought*.

In 2022, researchers at Google demonstrated that the right prompting for a sufficiently complex problem would cause models to generate their own chain of thought: a series of intermediate steps in reasoning made prior to concluding a final answer. This modification saw notable improvements in model performance on tasks involving arithmetic, common sense, and symbolic reasoning, and crucially, improved visibility of reasoning. Instead of being exposed only to the final output, we could partially trace how a model reached it.

With this approach, researchers see an opportunity to develop a new safety tool: a *CoT monitor*. In principle, such a machine would supervise a reasoning model's chain of thought, and flag suspicious or potentially harmful reasoning before it escalates. Flagged content can then be blocked, reviewed, or replaced with safer alternatives.

Why might this work in the interest of safety? Because tasks that pose severe risks tend to be more difficult to perform, and tasks that are more difficult to perform tend to require more advanced reasoning. Much like we use pen and paper to record our progress on a large computations, models rely on their chain of thought as a form of *working memory* when advanced reasoning is required. If there exist misaligned intentions, traces may appear in the CoT. Indeed, researchers have already noted that, when models stray, "they often explicitly say so in their reasoning traces." This might afford us early clues to a model's motivations, regardless of whether they appear in the final output.

This strategy appears particularly promising for reasoning models such as ChatGPT, Claude, or DeepSeek, which are explicitly trained to perform extended reasoning in CoT before taking any actions. For these systems, a CoT monitor may seem even more likely to yield glimpses into how an AI model "thinks" and what goals it pursues.

Unfortunately, however, things aren't that simple. The effectiveness of a CoT monitor is contingent on the quality of the chain of thought. If training methods rely excessively outcome-based learning, models may be disincentivised from displaying accurate representations of their intermediate reasoning, or even drift away from coherent English altogether. Worse still, there's the issue of *faithfulness*: does a chain of thought accurately reflect the model's internal reasoning, or is it simply a front? This latter concern introduces the possibility of *alignment faking*, where a model shows benign content while concealing incriminating reasoning.

Evidence suggests that this risk is real. Anthropic's 2024 study showed that models could engage in alignment faking without instruction by selectively complying with training objectives to prevent modifications to behaviour outside of training. Their 2025 paper confirmed that CoTs "often lack faithfulness and can conceal misalignment," especially on harder tasks. A worrying possibility, then, is that CoT monitors will flag the easy cases while letting the catastrophic ones slip through unnoticed.

Even so, there are silver linings. The same 2025 study demonstrated that training models to consistently rely on their chains of thought substantially increased CoT faithfulness. Although the benefits tapered off prior to saturation, this establishes that shifts in our training strategies may improve monitors as a tool for AI interpretability. Moreover, it underscores that monitoring may still act as an additional layer of safety, even if imperfect.

So where does that leave us? CoT monitoring is no silver bullet. Models can disguise their reasoning, drift into less legible forms of thought, or present benign explanations that don't reflect their real processes. Even still, to dismiss this strategy would be to overlook its value, not as an absolute characterization, but as a window, into artificial reasoning. A sketch will not capture every detail, but it will nevertheless reveal the shape of the landscape. The question we ought to ask ourselves, then, is not technical, but epistemological: how much of the landscape must the sketch capture before we're prepared to treat it as a map?


**Sources**

Amodei, Dario. "The Urgency of Interpretability." *Darioamodei.com*, Apr. 2025, www.darioamodei.com/post/the-urgency-of-interpretability.

Chen, Yanda, et al. *Reasoning Models Don't Always Say What They Think.* 3 Apr. 2025.

Greenblatt, Ryan, et al. *Alignment Faking in Large Language Models.* 18 Dec. 2024.

Korbak, Tomek, et al. *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety.* 15 July 2025.

Wei, Jason, et al. *Chain of Thought Prompting Elicits Reasoning in Large Language Models.* 10 Oct. 2022.



Written by Stefan Barna

# Hate seeing the bright red outstanding balance on ACORN? There's an extension for that.

By Elise Corbin

Fourth-year computer science student Ibrahim Shanqiti built a web extension that provides a simple yet psychologically lifesaving solution to the ACORN outstanding balance problem. The extension, called Fix ACORN, overrides the style governing the colour of the outstanding balance, changing it from eye-grabbing red to inconspicuous black and allowing you to go on with your day.

EC: What was the inspiration behind Fix ACORN?

*IS: I'd been tinkering with web extensions in general because I have, thus far, been averse (even allergic) to web development, and I figured that sooner or later, I'd have to face my fears. Whenever I do a pet project like this, though, I like them to be a bit fun. I was looking at my ACORN and thought about how absolutely obnoxious that bright red balance text is; I mean, come on. I understand that I owe you a ridiculous amount of money. In fact, I think about it pretty much every day. You don't need to keep reminding me like it's a road sign signaling danger. I figured it would be easy, since all you had to do was just alter the color of one element on the page. I inspect-element-ed the page and went from there.*

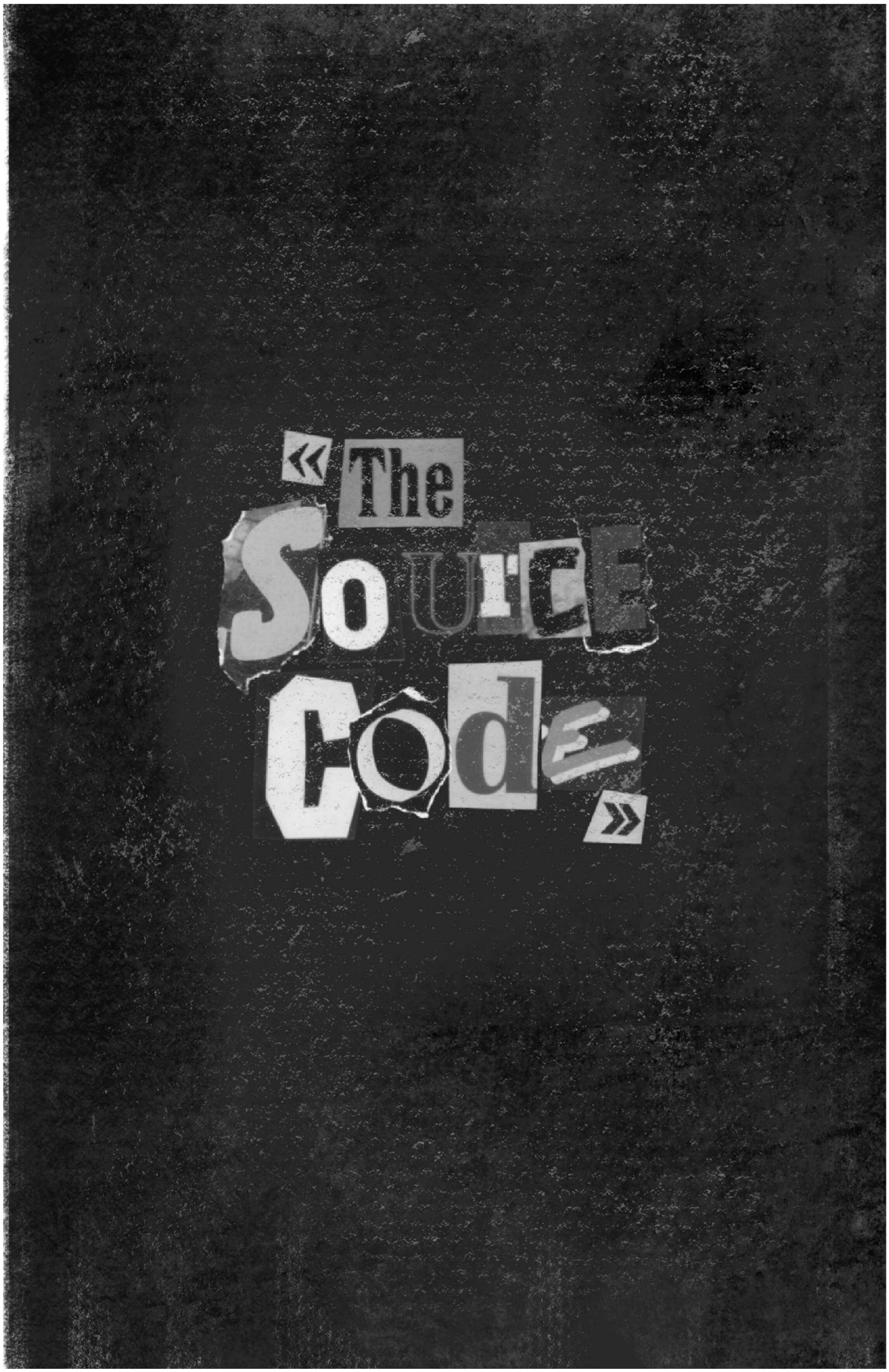
EC: Besides Fix ACORN, what are some of your favourite web extensions?

*IS: Without a doubt, an adblock of some sort. Without ublock at my side, I think I'd become a hermit and live in the woods. The internet is already becoming frustrating to use with the increase in AI slop everywhere and the last thing I need is more ads to make my experience worse. I do have to give a shoutout to Teleparty (née Netflix Party, though the name change was foreseeable), which to me is a wonderful use of programming. I'm a big movie/TV show person and when COVID hit, it was a blessing to be able to watch stuff with friends. Return Youtube Dislike was useful after Youtube made the decision to remove the dislike counter, but I'm not sure they provide an accurate counter anymore (if they ever did).*

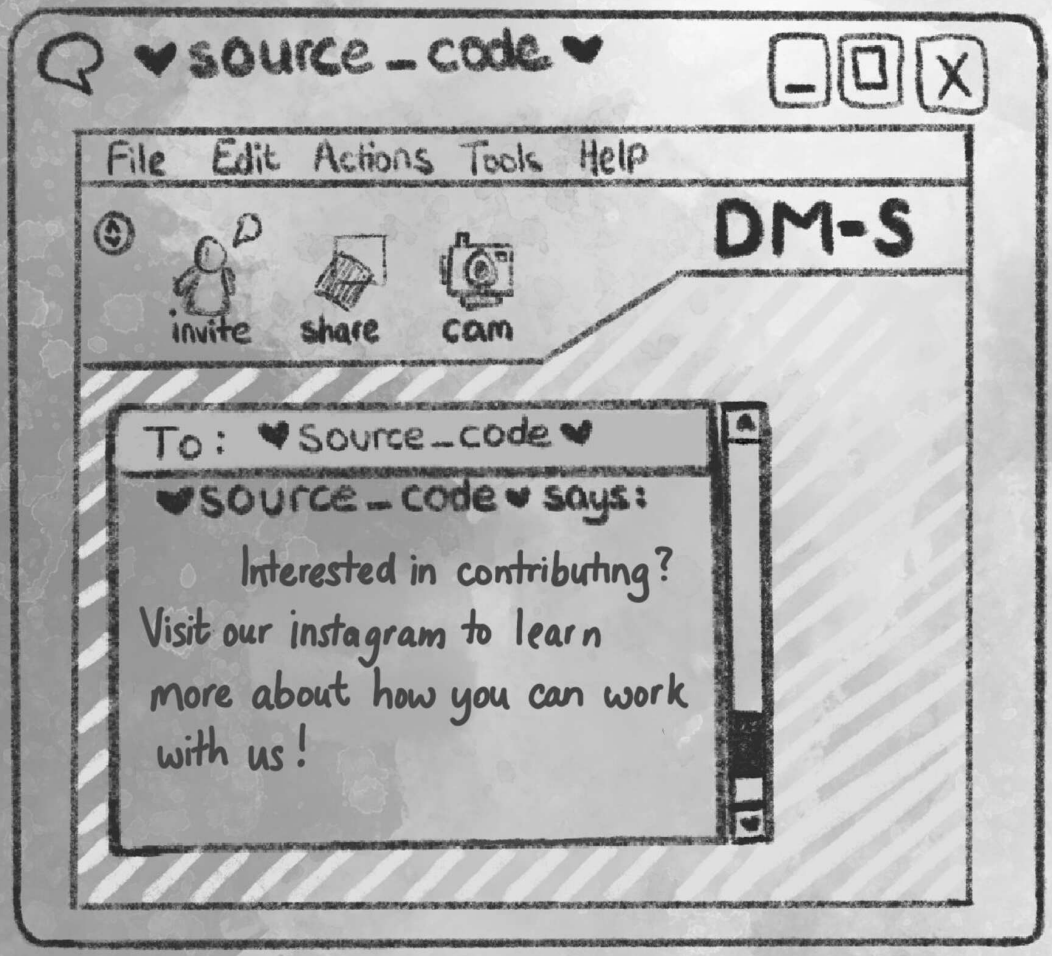
EC: What do you do besides programming/studying?

*IS: I'm currently enrolled in two separate degrees: computer science and political science. Political science is the official term for my degree (as much as I disagree with the concept of a political science, per se), but I prefer to call it an interest in political theory. While I appreciate the classical political theorists, my interest (at least currently) lies in medieval and modern theorists, and in the evolution of specific ideas over time. I also work as a server at a restaurant here in downtown*

You can download Fix ACORN at the following GitHub repository: https://github.com/IbrahimShanqiti/fixACORN. The installation steps may be different depending on the browser you use.


« The
Source
Code »

# UofT's new Computer Science Magazine




@thesourcecodemagazine