Mindfully Training AI Models Using Public Data

The Ethical Web Data Collection Initiative (EWDCI) is an industry-led consortium of web data collectors focused on strengthening public trust, promoting ethical guidelines, and helping businesses and their customers make informed data extraction choices. We're setting the standard for ethics in the process widely known as "data scraping", with the goal of enhancing trust—a key component of a free, fair, and open Internet.

The EWDCI appreciates this opportunity to respond to the ICO's call for consultation on generative AI. As a foundational element in the AI-training process, we as web data collection companies are looking for legal clarity ourselves, as well as seeking a seat at the table, as it were, in outlining the types of data and data-collection methodologies that will be considered legal under any new British law.

Collecting Public Data Makes Sense

Our suggestions herein refer only to the public or private nature of data; issues like copyright and intellectual property are different discussions for another day. It is our position that legitimate interests exist for using publicly-available personal data to train an AI model, as long as safeguards are built into the process. For example, there's great value in training an AI to understand what a human being looks like by learning real faces, as well as that a person's name is just that—a name. Existing frameworks such as GDPR and UK GDPR are starting points for establishing appropriate guardrails for data collection, anonymization and output; and establishing those guardrails will require input from the technologists that build and deploy data-collection tools.

Some Personal Data is Also Public Data

When considering personal data, it is understandable to immediately consider that data to be private. However, two generations' worth of interactive online activity have led to a digital landscape full of data that is both personal and public: in essence, we live our lives online. Given the vast amounts of personal data that people voluntarily make public, there should be a reasonable expectation that public data could be used to train AI models. Allowing AI companies to use this data lawfully for processing and/or allowing a carve-out in the law for personal data manifestly made public (similar to other jurisdictions)—along with sensible guardrails—would be the most reasonable way to protect personal data while also not chilling commerce. Some sensible guardrails could include minimizing data retention time, as well as posted notices on AI websites regarding what sorts of public data they obtain, and how they use it. Should legislation emerge that places all personal data off-limits to collection for use in

training artificial intelligences, then UK-based AI models will be inferior by design to those based in virtually any other jurisdiction.

This is a Business-Competitiveness Issue
Just as the United Kingdom enjoyed success in previous digital revolutions, the emergence of easily-accessible AI is a race—one that the UK can win or lose. We also understand that striking the proper balance between maintaining personal privacy and achieving smarter artificial intelligence outcomes is not only the right thing to do, but will also shape the very nature of tomorrow's AI. That said, most countries will indeed draw the distinction between publicly-available personal data and private personal data; our advice is that the UK do so as well, or risk being left behind.

We appreciate the opportunity to comment, and we hope that this is only the beginning of a long and productive conversation for all involved stakeholders.

Christian Dawson
Co-Founder and Executive Director
Ethical Web Data Collection Initiative
Internet Infrastructure Coalition
web: ethicalwebdata.com
email: contact@ethicalwebdata.com
tel: +1 (202) 524-3183

About the EWDCI
The Ethical Web Data Collection Initiative (EWDCI) is an industry-led consortium of web data collectors focused on strengthening public trust, promoting ethical guidelines, and helping businesses and their customers make informed data extraction choices. This international, member-driven consortium is developing an accreditation program developed to bring greater accountability and build consumer confidence in the data collection industry.

The EWDCI is a working group of the Internet Infrastructure Coalition (i2Coalition), which works with Internet infrastructure providers to advocate for sensible policies, design and reinforce best practices, help create industry standards, and build awareness of how the Internet works.

https://ethicalwebdata.com
https://i2coalition.com