

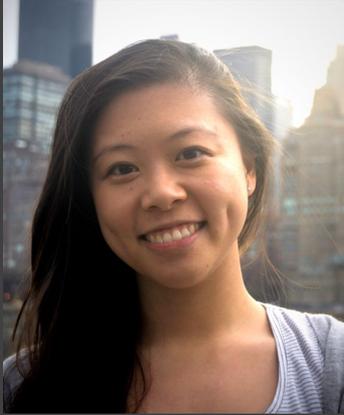
The Alan Turing Institute

Participatory data stewardship in AI

Jennifer Ding (she/her) and Anne Lee Steele (she/her)
12 December 2023



Introductions



Jennifer Ding

Senior Researcher, Research
Applications
The Alan Turing Institute



Anne Lee Steele

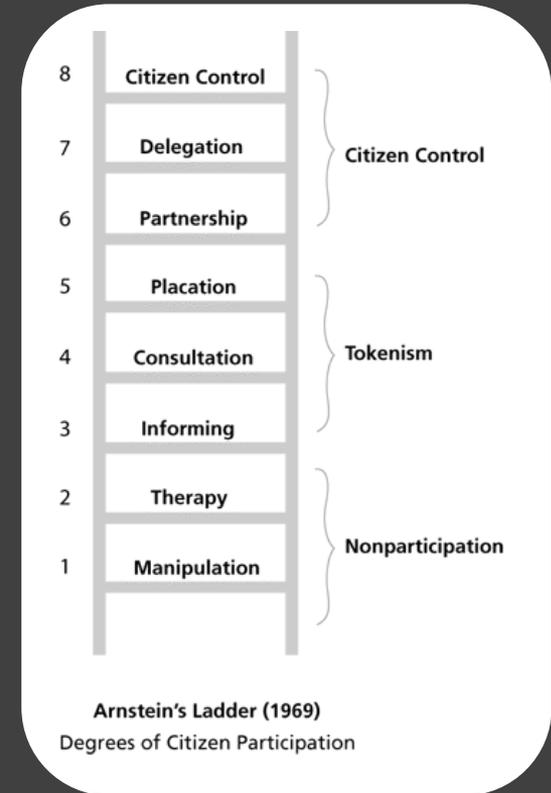
Research Community Manager
The Turing Way,
The Alan Turing Institute

Participatory approaches to data and data stewardship

- *What histories of ‘participatory’ work pre-date current AI initiatives & collaboratives?*
- *Where does ‘open source’ fit in?*
- *How do these initiatives address these ideas both separately & together?*

Participatory design: a (very) brief history

Modern understanding of “meaningful” and “participatory engagement” in public processes began as a reaction to top-down public (urban) planning



Participatory approaches to science & scientific research

Public participation and citizen science emerged in response to norms in public health and medical experiments

Example: The Tuskegee Experiment (1932 - 1972)



Free and open source (F/OSS): a (very) brief history



Created by hacker communities after “unbundling” of academic computing to share software.

Evolved and inspired other ‘open’ sub-movements - for open access, open government, open data, etc.

Open evolved into a broader idea, with societal implications: for public good

Participatory approaches to data and data stewardship

Emergent frameworks for enabling public participation in data usage

Project

Participatory data



Department of Health and Social Care, on behalf of NHS Transformation

Designing Data Stewardship Models for Artificial Intelligence (AI) R&D

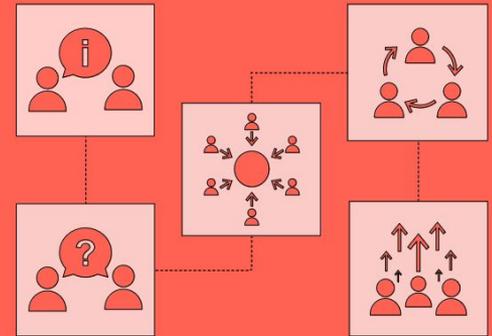


Participatory data stewardship

A framework for involving people in the use of data

September 2021

Data for the public good



Participatory approaches to data and data stewardship

*Different organisations and groups have divergent (and convergent) understandings of what **participation** looks like*

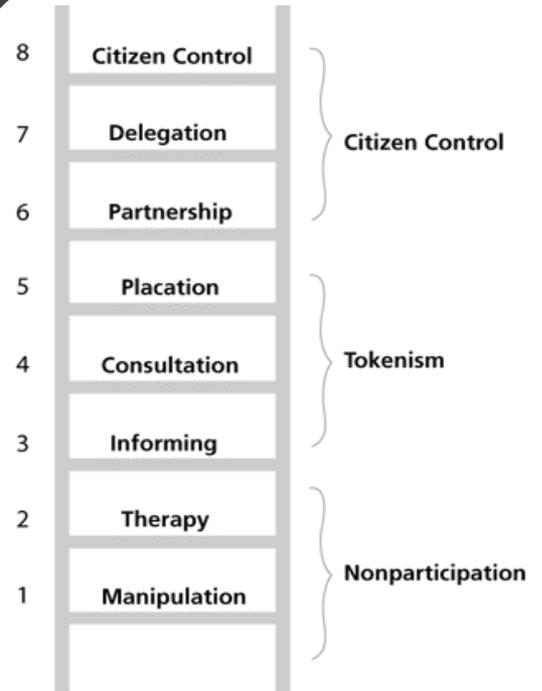
arXiv > cs > arXiv:2301.08488

Computer Science > Computers and Society

[Submitted on 20 Jan 2023]

Towards Openness Beyond Open Access: User Journeys through 3 Open AI Collaboratives

Jennifer Ding, Christopher Akiki, Yacine Jernite, Anne Lee Steele, Temi Popo



Arnstein's Ladder (1969)

Degrees of Citizen Participation

The Alan Turing Institute

The Turing Way



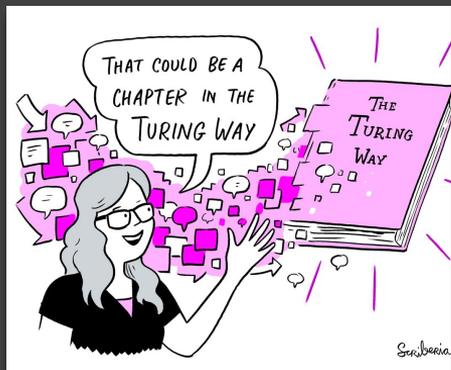
The Turing Way



A community-led project on data science.

We are an open source, open collaboration and globally distributed community working together to **make data science accessible, reproducible and beneficial.**

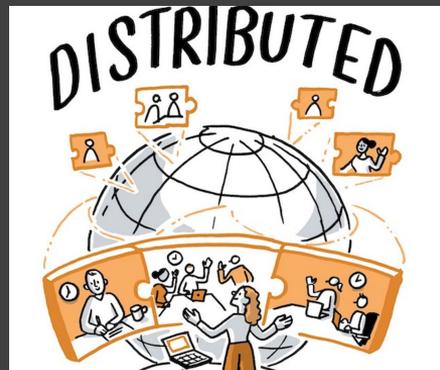
<https://the-turing-way.start.page/>



Co-created Book



Community-based



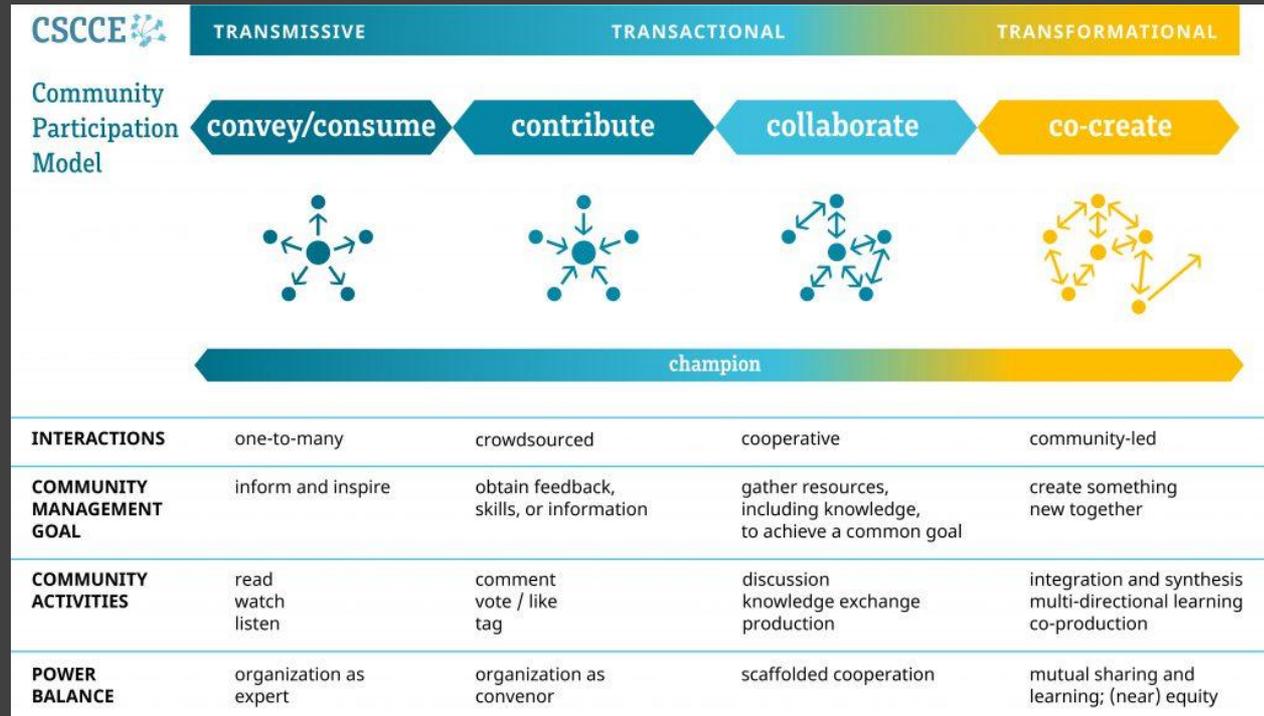
Participatory process



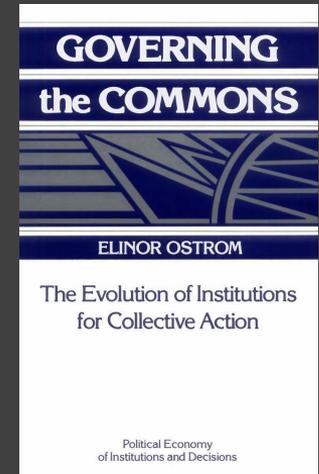
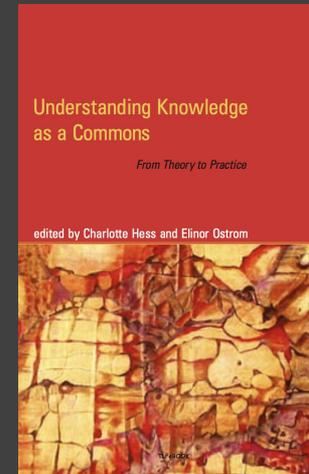
Culture change

TTW: community-participation model

Community-led projects and approaches to growth and themed work: 'participation' defined broadly

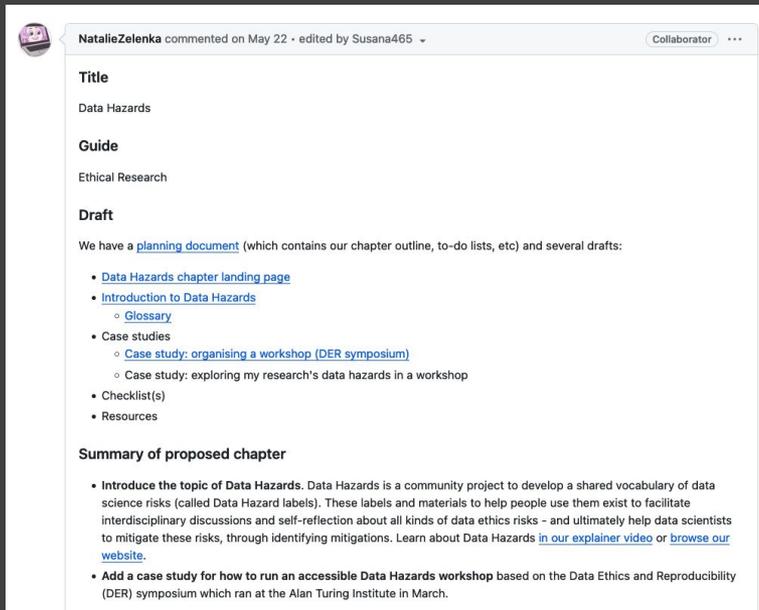


Community Participation and knowledge commons



Governance: rules to care for resources & community

TTW: community-participation model



NatalieZelenka commented on May 22 · edited by Susana465 · Collaborator · ...

Title
Data Hazards

Guide
Ethical Research

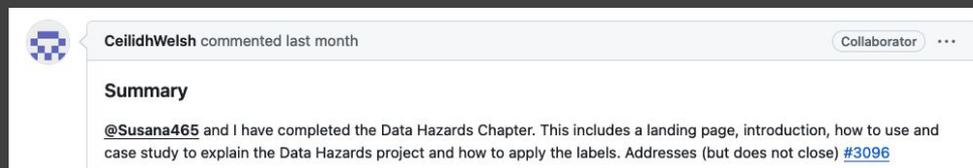
Draft
We have a [planning document](#) (which contains our chapter outline, to-do lists, etc) and several drafts:

- [Data Hazards chapter landing page](#)
- [Introduction to Data Hazards](#)
 - [Glossary](#)
- Case studies
 - [Case study: organising a workshop \(DER symposium\)](#)
 - Case study: exploring my research's data hazards in a workshop
- Checklist(s)
- Resources

Summary of proposed chapter

- **Introduce the topic of Data Hazards.** Data Hazards is a community project to develop a shared vocabulary of data science risks (called Data Hazard labels). These labels and materials to help people use them exist to facilitate interdisciplinary discussions and self-reflection about all kinds of data ethics risks - and ultimately help data scientists to mitigate these risks, through identifying mitigations. Learn about Data Hazards [in our explainer video](#) or [browse our website](#).
- **Add a case study for how to run an accessible Data Hazards workshop** based on the Data Ethics and Reproducibility (DER) symposium which ran at the Alan Turing Institute in March.

Co-creation of Chapters & content

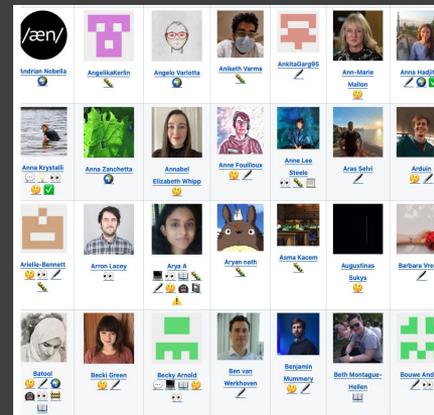
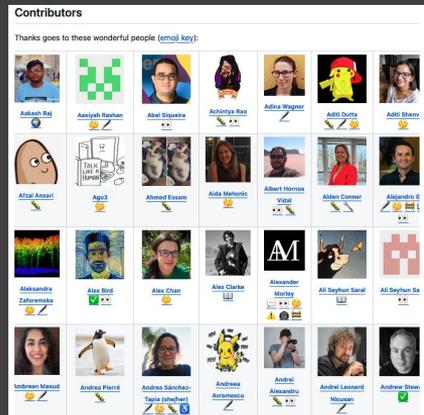
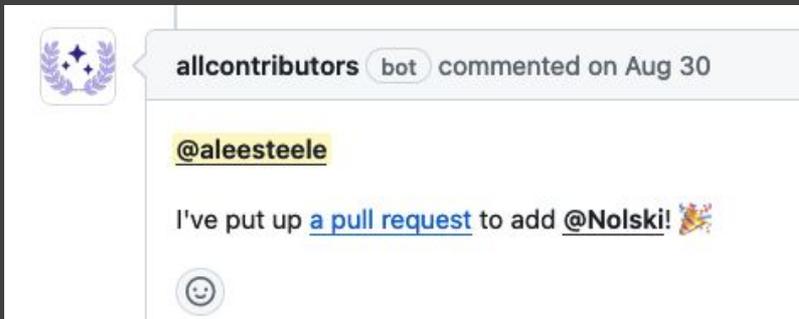
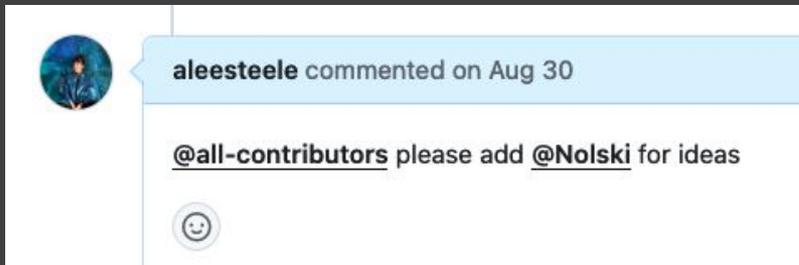


CeilidhWelsh commented last month · Collaborator · ...

Summary

[@Susana465](#) and I have completed the Data Hazards Chapter. This includes a landing page, introduction, how to use and case study to explain the Data Hazards project and how to apply the labels. Addresses (but does not close) [#3096](#)

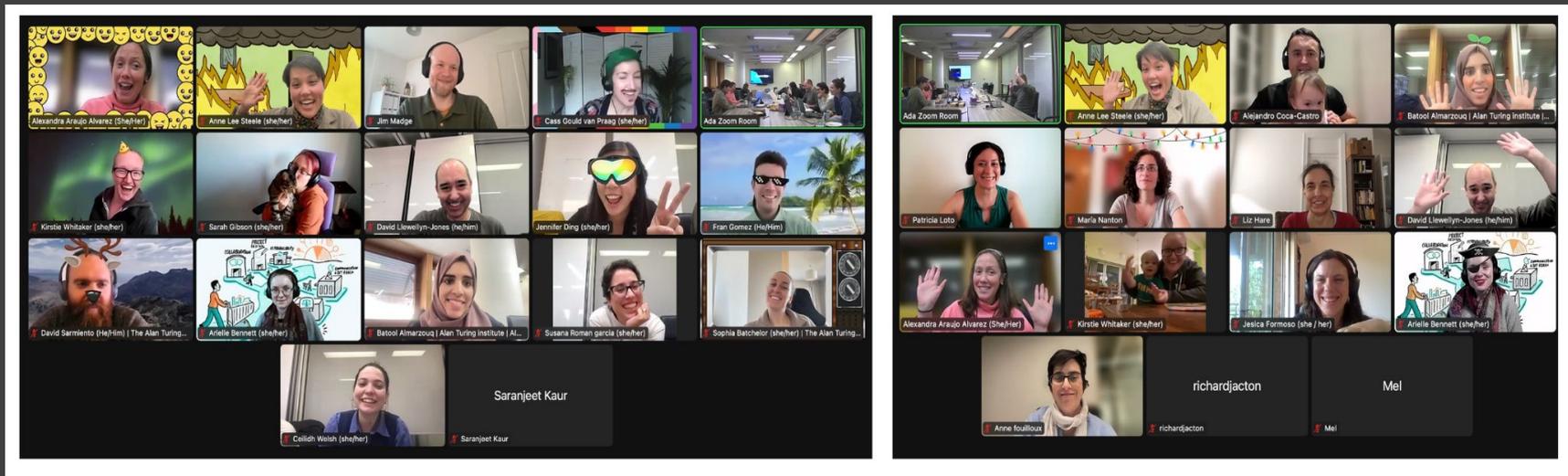
TTW: community-participation model



Contribution & Acknowledgement methods

TTW: community-participation model

Community calls: Book Dash, Collaboration Cafes, Fireside Chats, Community Share-outs



TTW: community-participation model

Community-led
project working
groups and
informal
interest groups



Project Growth & Recognitions

- 450+ contributors, 3500+ monthly users - supported by 25+ core members (governance)
- 2 awards and engagement with 25+ orgs
- References in 40+ peer reviewed articles and 100s of publications, reports, policy documents
- Replication/extension by 3 Turing & 7+ external projects



**The
Alan Turing
Institute**

**Participatory AI: open
source AI collaboratives**



Open Source AI Collaborative

- **Distributed, volunteer-led teams** creating alternative AI development pathways grounded in “open” principles
- Focus on activities relevant for their respective communities:
 - building **multilingual LLMs** (BigScience Workshop, Aya Initiative)
 - Building **code LLMs** (StarCoder)

Towards Openness Beyond Open Access: User Journeys through 3 Open AI Collaboratives

Jennifer Ding
The Alan Turing Institute
jding@turing.ac.uk

Christopher Akiki
Leipzig University
christopher.akiki@uni-leipzig.de

Yacine Jernite
Hugging Face
yacine@huggingface.co

Anne Lee Steele
The Alan Turing Institute
asteele@turing.ac.uk

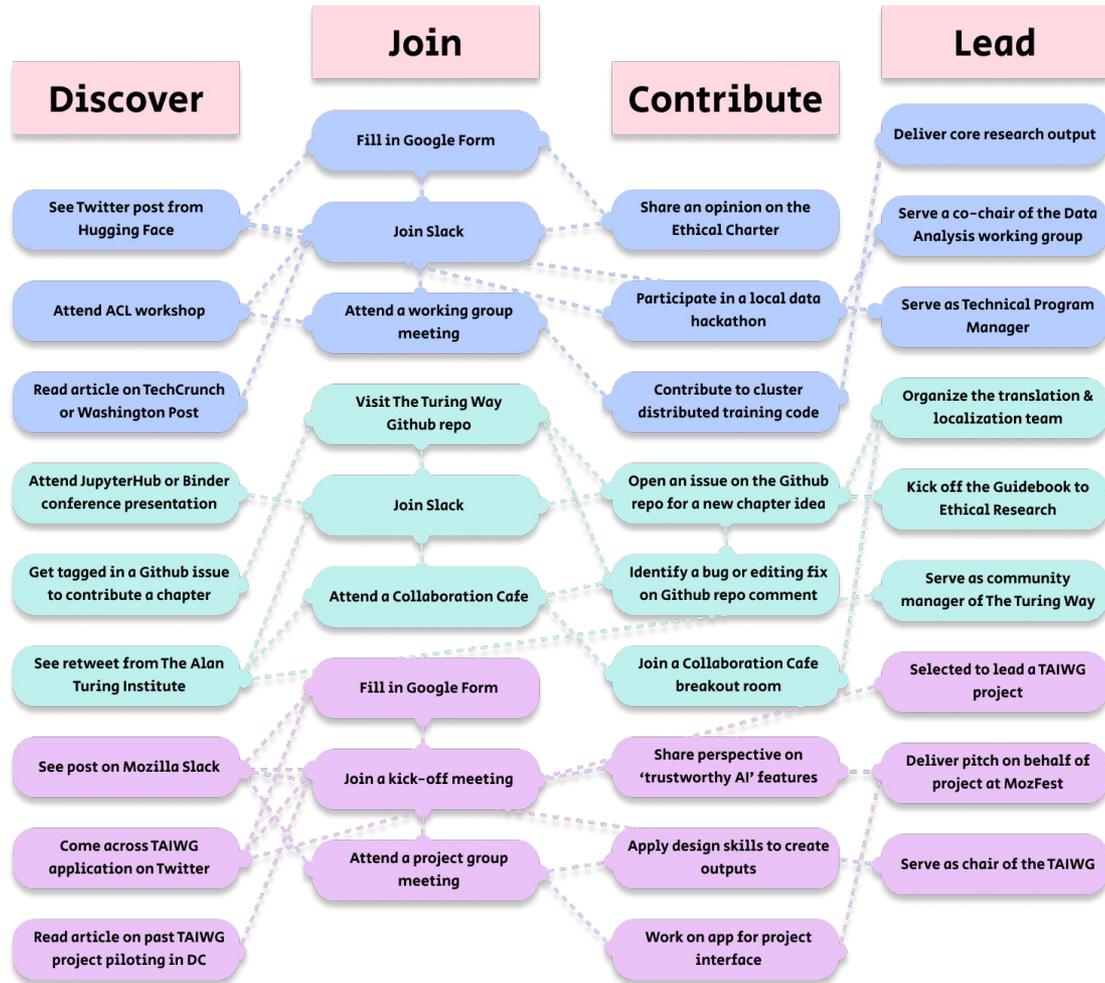
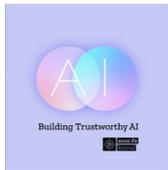
Temi Popo
Mozilla Foundation
temi@mozillafoundation.org

Abstract

Open Artificial Intelligence (Open AI) collaboratives offer alternative pathways for how AI can be developed beyond well-resourced technology companies and who can be a part of the process. To understand how and why they work and what additionality they bring to the landscape, we focus on three such communities, each focused on a different kind of activity around AI: building *models* (BigScience workshop), *tools/ways of working* (The Turing Way), and *ecosystems* (Mozilla Festival’s Building Trustworthy AI Working Group). First, we document the community structures that facilitate these distributed, volunteer-led teams, comparing the collaboration styles that drive each group towards their specific goals. Through interviews with community leaders, we map user journeys for how members discover, join, contribute, and participate. Ultimately, this paper aims to highlight the diversity of AI work and workers that have come forth through these collaborations and how they offer a broader practice of openness to the AI space.

Towards Openness Beyond Open Access Neurips workshop paper:
<https://arxiv.org/abs/2301.08488>

User journeys through open AI communities



**The
Alan Turing
Institute**

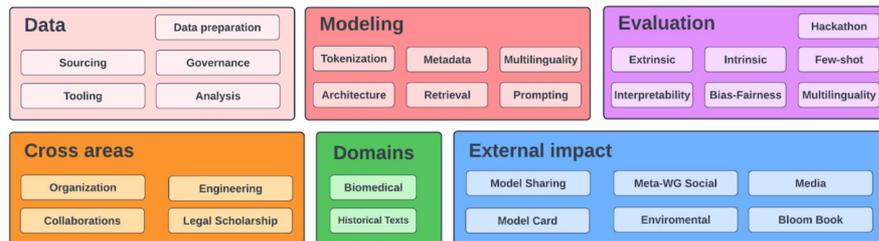
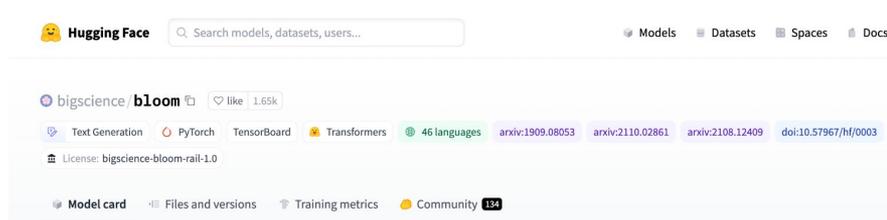
BigScience Workshop



BigScience Workshop - BLOOM LLM

"Open in the way people can understand what you're talking about and can try it themselves..."

BigScience Workshop Working Groups



BigScience Large Open-science Open-access Multilingual Language Model
Version 1.3 / 6 July 2022



Join/follow

Twitter: [@BigScienceW](https://twitter.com/BigScienceW)
Website home: <https://bigscience.huggingface.co>
[Join the newsletter](#)
[Participate in the workshop](#)
email: [bigscience-contact \[at\] googlegroups \[dot\] com](mailto:bigscience-contact[at]googlegroups[dot]com)

BigScience Data Governance

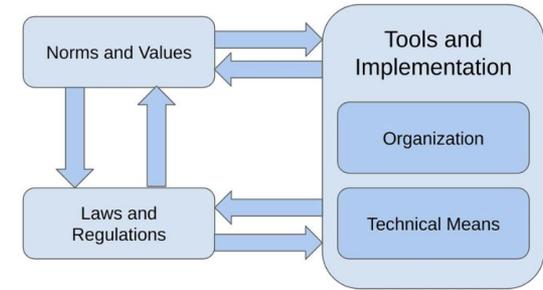
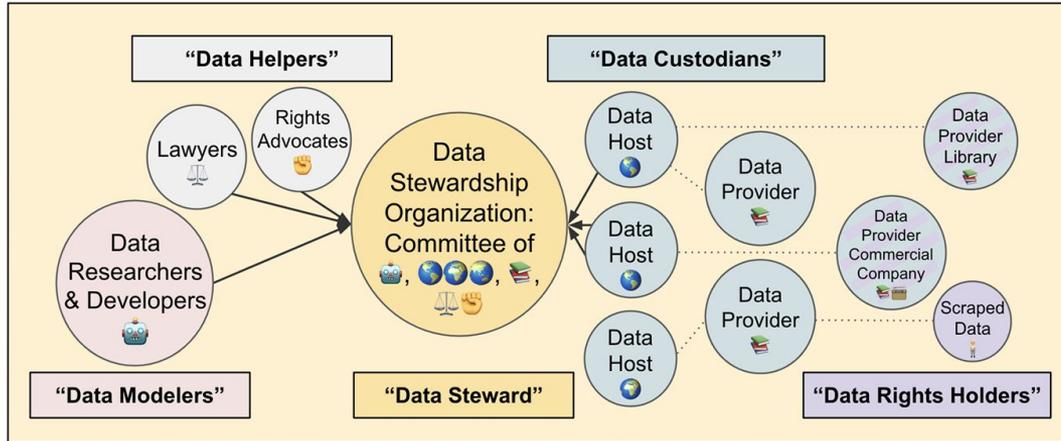


Fig. 2. Collaborative governance mechanisms rely on interacting pillars.

Fig. 1. Overview of the Data Stewardship Organization and Actors

Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAcCT '22). Association for Computing Machinery, New York, NY, USA, 2206–2222. <https://doi.org/10.1145/3531146.3534637>

Catalogue of Tools & Metrics for Trustworthy AI

These tools and metrics are designed to help AI actors develop and use trustworthy AI systems and applications that respect human rights and are fair, transparent, explainable, robust, secure and safe.

Contribute to the catalogue

BigScience Catalogue of Language Data and Resources

Website

Technical Educational Procedural

Uploaded on Sep 15, 2022

Organisation(s): BigScience

The main goal of the catalogue is to support the creation of the BigScience dataset while adhering to the values laid out by the various data working groups: collecting diverse resources (Data Sourcing), supporting information required for open and easily usable technical infrastructure (Data Tooling), and respecting the privacy of data subjects and the rights of data owners (Data Governance).

As per 14 December 2021, the catalogue contained 192 entries with 432 different language tags (each entry can have multiple language tags). The most frequent language tags are those of the BigScience target language groups. English is the most frequent language across all entries. The most frequent varieties of Arabic are Modern Standard Arabic and Classical Arabic, the most frequent Indic languages are Hindi, Bengali, Telugu, Tamil and Urdu, and the most frequent NigerCongo languages are Swahili, Igbo, Yoruba and isiZulu.

Select entries to visualize

Select entries by category, language, type of custodian or media

You can select specific parts of the catalogue to visualize in this window. Leave a field empty to select all values, or select specific options to only select entries that have one of the chosen values.

You can filter the catalogue to only visualize entries that have certain properties, such as:

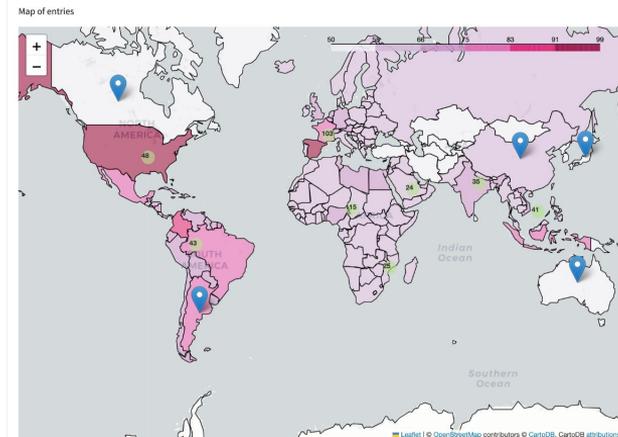
Choose an option

Your query matched 252 entries in the current catalogue.

I want to visualize

- Where the organizations or data custodians are located
- Where the language data creators are located

🌸 - BigScience Catalog of Language Resources



BigScience Catalogue of Language Data and Resources:

<https://oecd.ai/en/catalogue/tools/bigscience-catalogue-of-language-data-and-resources>

<https://huqqinqface.co/spaces/bigscience/SourcingCatalog>

BigScience



The BigScience OpenRAIL-M License

Carlos Muñoz Ferrandis - BigScience Legal/Ethical Working Group
Danish Contractor - BigScience Model Governance Working Group

Disclaimer: This post is not intended to be legal advice from any of the authors.

In collaboration with the [RAIL initiative](#), we are excited to release the BigScience OpenRAIL-M [license](#) – a license with behavioral use restrictions that can be applied to any AI model being released. We hope the AI community will find this license useful for releasing their AI Models.



The Turing Way

Search this book...

- Welcome
- [Guide for Reproducible Research](#) ^
- Overview v
- Open Research v
- Version Control v
- [Licensing](#) ^
- License Compatibility
- Ethics-informed Licensing



Open & Responsible ML Licenses

The “open source” approach to collaborative software development has permeated and influenced AI development and licensing practices. It is a common practice of ML developers to use open source licenses to release their ML models. This is due to the fact that open source licenses have become a standard practice when it comes to the sharing of artefacts in the entire ICT space (for example, software; datasets; models; apps). ML developers might colloquially refer to “open sourcing a model” when they make its weights (trained model parameters) available by attaching an official open source license, or any other open software or content license such as Creative Commons.

However, open source licenses do not take the technical nature and capabilities of the ML model as a different artefact to software/source code into account, and are therefore ill-adapted to enabling a more responsible use of ML models.

In order to balance the principles from open source with a growing demand of responsible ML development, use, and access, a new branch of ML licenses called Responsible AI Licenses (RAIL) emerged in 2019 with the [RAIL Initiative](#). Research initiatives such as [BigScience](#) and companies such as [Hugging Face](#) have decided to join efforts and push towards this direction along with the RAIL Initiative.

BigScience Open RAIL-M: <https://bigscience.huggingface.co/blog/bigscience-openrail-m>

TTW Chapter on Licensing ML models: <https://the-turing-way.netlify.app/reproducible-research/licensing/licensing-ml.html>

**The
Alan Turing
Institute**

BigCode



BigCode - StarCoder LLM

“Open not only for transparency but accountability”



The Stack is an open governance interface between the AI community and the open source community.

Am I in The Stack?

As part of the BigCode project, we released and maintain [The Stack](#), a 6 TB dataset of permissively licensed source code over 300 programming languages. One of our goals in this project is to give people agency over their source code by letting them decide whether or not it should be used to develop and evaluate machine learning models, as we acknowledge that not all developers may wish to have their data used for that purpose.

This tool lets you check if a repository under a given username is part of The Stack dataset. Would you like to have your data removed from future versions of The Stack? You can opt-out following the instructions [here](#).

The Stack version:
v1.2

Your GitHub username:
dingailing

Check!

Yes, there is code from 6 repositories in The Stack:

- [dingailing/citi-map](#)
- [dingailing/simple-ChatBot](#)
- [dingailing/street-light-map](#)
- [dingailing/the-turing-way](#)
- [dingailing/tramTracker](#)
- [dingailing/turing-commons](#)

Opt-out

If you want your data to be removed from the stack and model training open an issue with [this link](#) (if the link doesn't work try right a right click and open it in a new tab) or visit <https://github.com/bigcode-project/opt-out-v2/issues/new?&template=opt-out-request.md>.

StarCoder



How to join?

We are excited to invite AI practitioners from diverse backgrounds to join the BigCode project! Note that BigCode is a *research collaboration* and is open to participants who

1. have a professional research background and
2. are able to commit time to the project.

In general, we expect applicants to be affiliated with a research organization (either in academia or industry) and work on the technical/ethical/legal aspects of LLMs for coding applications.

You can apply here to the BigCode project!

BigCode Data & Project Governance Card

THE BIGCODE PROJECT GOVERNANCE CARD

Sean Hughes^{1,*} Harm de Vries² Jennifer Robinson¹
Carlos Muñoz Ferrandis^{3,*} Loubna Ben Allal³ Leandro von Werra³
Jennifer Ding⁴ Sebastien Paquet² Yacine Jernite³

¹ServiceNow ²ServiceNow Research ³Hugging Face ⁴The Alan Turing Institute

Corresponding authors (*) can be contacted at contact@bigcode-project.org

Abstract

This document serves as an overview of the different mechanisms and areas of governance in the BigCode project. It aims to support transparency by providing relevant information about choices that were made during the project to the broader public, and to serve as an example of intentional governance of an open research project that future endeavors can leverage to shape their own approach. The first section, Project Structure, covers the project organization, its stated goals and values, its internal decision processes, and its funding and resources. The second section, Data and Model Governance, covers decisions relating to the questions of data subject consent, privacy, and model release.



The Turing Way

Search this book...

- Welcome
- Guide for Reproducible Research
- Guide for Project Design**
- Overview of Project Design
- Project Design Checklist
- Creating Project Repositories
- Personas and Pathways
- File Naming Convention
- Code Styling and Linting
- Sensitive Data Projects
- Managing Sensitive Data Projects
- Working on Sensitive Data Projects
- Data Governance**
- Data Governance for the Machine Learning Pipeline**
- BigCode Data Governance
- Case Study



Data Governance for the Machine Learning Pipeline

A Machine Learning (ML) pipeline consists of a series of activities including the collection of data, training of an ML model, and the deployment of the model into use. Data is integral throughout the ML process and the methods for which data is collected, annotated, processed, and shared will impact individuals and communities who may be represented in or the creators of the data, as well as data users who would like access to the data.

This chapter will cover examples of data governance practices for ML for different steps in the pipeline, which may include but not be exclusive to:

- Data Collection
- Data Management
- Data Processing

Data Collection

Many ML models are trained using datasets collected by a research team, which may be proprietary, or by using an open dataset that is available for download (sometimes with restrictions on its usage, such as only being available for use in academic settings). The deep learning (DL) family of models, in particular, relies on massive corpuses of data such as text, code, images, sound, and other media. The process of data collection depends on the type and volume of data required and sources for acquisition. For example, a project that uses patient health record data versus a project that uses a dataset of millions of social media posts will require different processes for gaining access to data as well as actually collecting and storing the data. Many DL models rely on data scraped from the internet due to the sheer volume of digital content that is available on the web. For example, [ImageNet datasets](#) are sourced from web images from image hosting websites like Flickr, and [LAION datasets](#) come from web crawling sources like Common Crawl. These methods of data collection through web scraping have raised issues regarding data quality and bias due to the nature of using uncurated

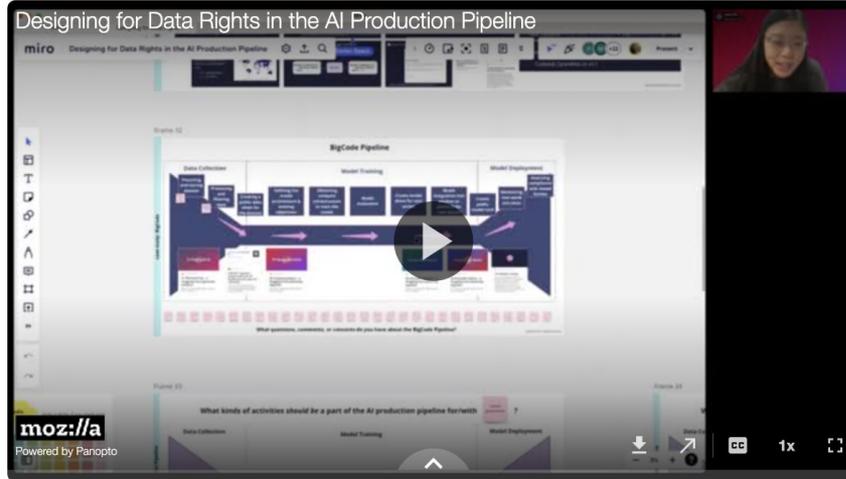
Project Governance paper: <https://arxiv.org/pdf/2312.03872.pdf>

TTW Chapter on Data Governance for the ML pipeline:

https://the-turing-way.netlify.app/project-design/data-governance/bigcode_casestudy

WORKSHOP

Designing for Data Rights in the AI Production Pipeline



EN Allies in Practice Data Stewardship Recorded

Building trustworthy AI requires building public trust in how AI is developed. While the majority of AI production/resources are concentrated within a few companies in even fewer countries, alternative spaces are emerging for more people to participate in creating, applying, and governing data and data-generated ML models. New initiatives such as BigScience and BigCode seek to change extractive methods of AI production, replacing secretive web scraping with data stewardship and other data rights-affirming tools, practices, and systems.

SCHEDULE

PAST

02:00 PM → 03:00 PM

March 24
Europe/London

LINKS

Miro Board

[miro.com/app/board/uXjVMeuvLR8
=?share_link_id=159151239611](https://miro.com/app/board/uXjVMeuvLR8=?share_link_id=159151239611)

copy

FACILITATORS



Anne Lee Steele
The Alan Turing Institute



Jennifer Ding
The Alan Turing Institute



Yacine Jernite
Hugging Face

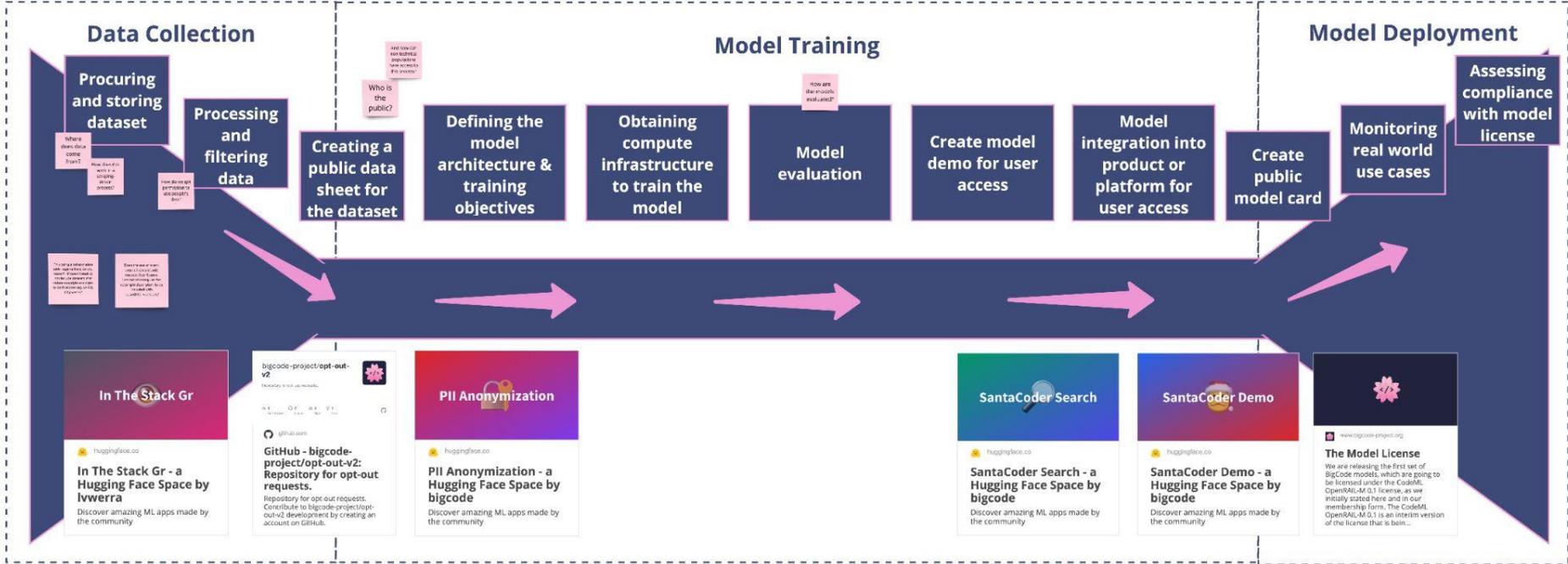


MozFest workshop Miro board:

https://miro.com/app/board/uXjVMeuvLR8=?share_link_id=159151239611

BigCode Pipeline

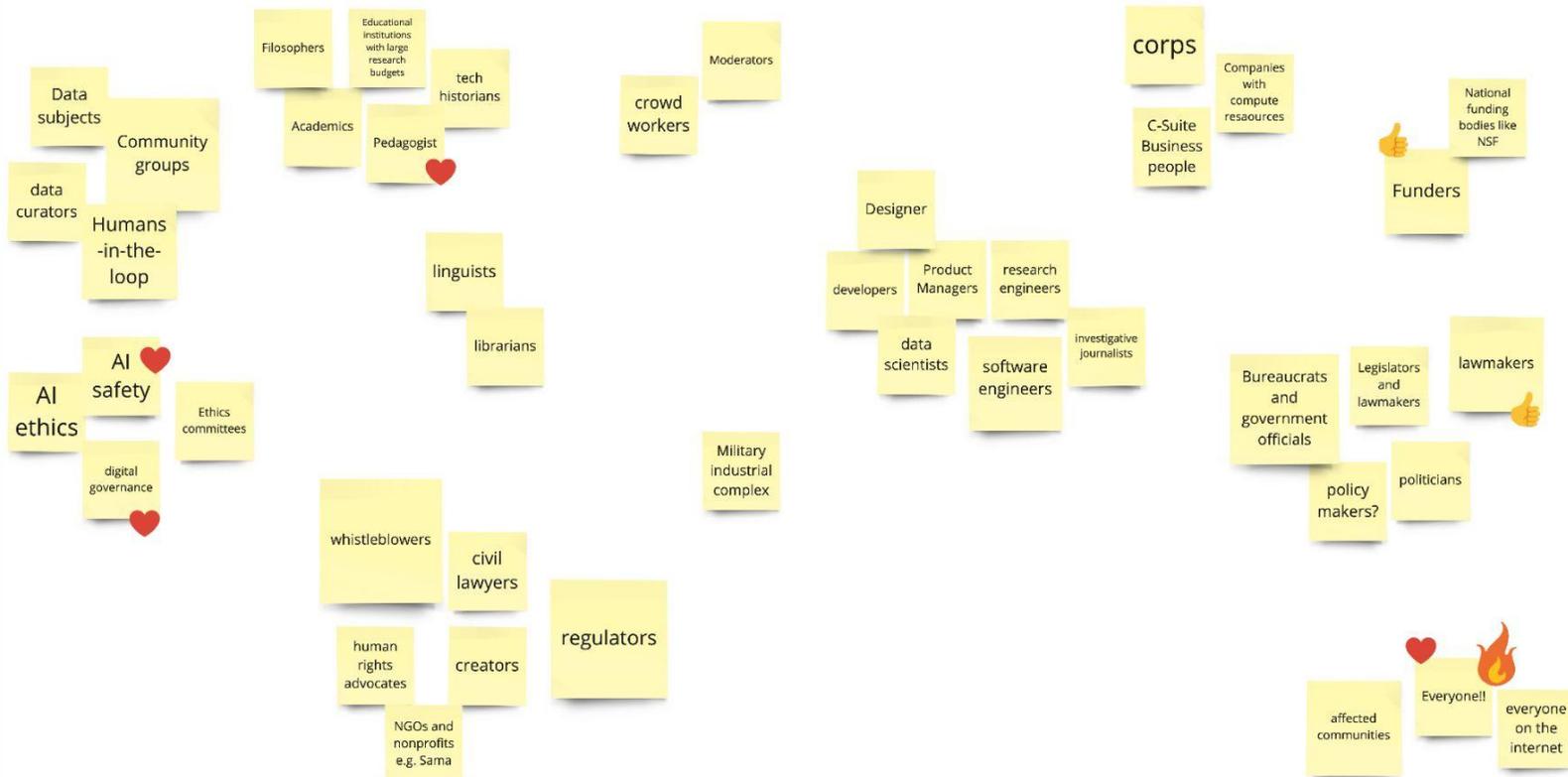
case study: BigCode



What questions, comments, or concerns do you have about the BigCode Pipeline?

Who is part of or can shape the AI production pipeline?

who



Drag and duplicate if you like what someone else wrote



Who is impacted by the AI production pipeline?

who

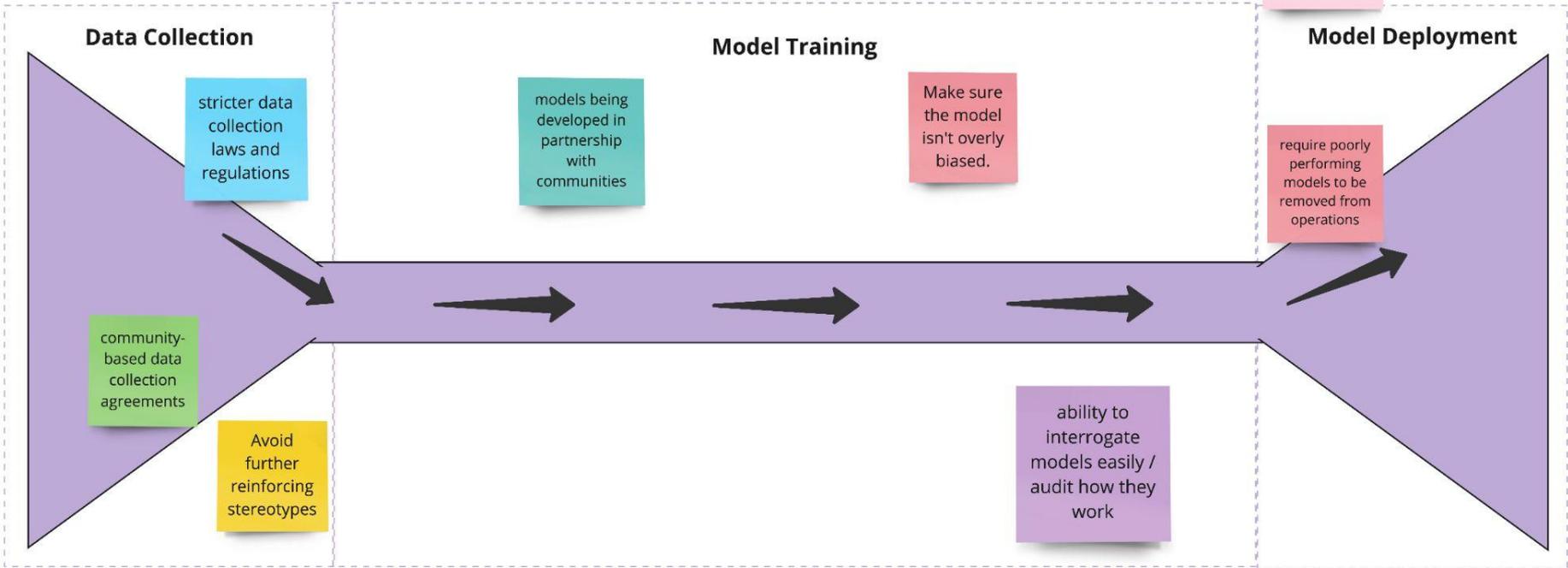


Drag and duplicate if you like what someone else wrote



What kinds of activities *should be* a part of the AI production pipeline for/with

Policed populations ?



The Alan Turing Institute

**Other participatory data &
AI initiatives**



← Cohere For AI

AYA: Accelerating Multilingual Progress

Join us



← Cohere For AI | 

Languages are not treated equally by researchers. Some languages have received disproportionate attention and focus in NLP.

Language	# of papers per million speakers	# of speakers (in millions)
Irish	5235	0.2
Basque	2430	0.5
German	179	83
English	63	550
Chinese	11	1,000
Hausa	1.5	70
Nigerian Pidgin	0.4	30

Van Esch et al. 2022



Closing the Contribution Chapter - December 15

On December 15, the Aya community will be hosting a live virtual event to celebrate the end of the Contribution Chapter of the project! Aya Project Lead, Sara Hooker, will share her final remarks on the project, and Community Lead, Madeline Smith, will announce the final Top 50 Language Champions! Aya Language Ambassadors and Regional Leads will share their messages of celebration and gratitude for their wonderful communities. Join us on December 15 to celebrate the end of this chapter with our dedicated contributors around the world, as we become one step closer to revolutionizing the world of AI forever!

Add the event to your Calendar

Watch live on Discord

Watch live on LinkedIn

Opening Remarks!

Sara Hooker, Aya Project Lead

Top 50 Quality Champions Announcement

Madeline Smith, Aya Operations Lead

Celebratory Remarks from Aya Language Ambassadors

Irem Ergun (Turkish)
Surya Guthikonda (Telugu)
Birin (Rin) Intachuen (Thai)
Jay Patel (Gujarati)
Emad Alghamdi (Arabic)
Ramith Hettiarachchi (Sinhala)
Nathanael Rakotonirina (Malagasy)
Iñigo Parra Martin (Basque)
Deividas Mataciunas (Regional Lead, Europe)
Yong Zheng-Xin (Malay)
Olanrewaju Samuel (Yorùbá)
Joseph Wilson (French)
Shivalika Singh (Regional Lead for Asia)
Abinaya Mahendiran (Tamil)

Cohere for AI Aya Initiative: <https://sites.google.com/cohere.com/aya-en/home>



Home + Research + Research projects

Developing a data charter with Camden Council

A project to co-develop a data charter that can help ensure the responsible use of citizen data

Learn more ↓

Project status

Ongoing

Related programmes

Public Policy

Introduction

The Public Policy Programme's Ethics Team has partnered with **Camden Council** and **Involve**, supported by funding from **Wellcome**, to deliver a series of education-based public engagement workshops – known collectively as a Resident Panel. The panel was diverse and represented the community, and has supported Camden's residents with the development and update of **a data charter** that can help ensure that the council's ongoing use of citizen data works for the common good.

Camden data charter: <https://www.camden.gov.uk/data-charter>

Browsealoud Accessibility News Camden Sign in / R

The 2023 panel looked at:

- the Data Charter
- Camden's commitments to the Data Charter
- the work that had been done since the first panel

Why have we created a Data Charter?

We believe that data rights are human rights. We make [use of a wide range of data](#) to improve residents' experiences with council services.

We know it is not always easy to have trust in organisations that use people's personal information. We want to lead the way and act as a responsible custodian of data.

All organisations in Camden need to work together to have the biggest impact on building trust in data sharing.

[Watch this video for more information about why we have created a Data Charter](#)

You can also find out about [how we developed the Data Charter](#).

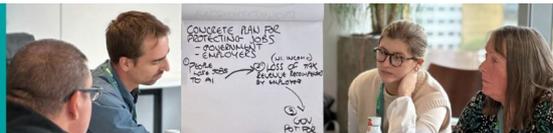
Data Charter principles

The principles of the Data Charter are:

1. Build trust through transparency
2. Provide accountability and oversight
3. Make sure data is secure, safe and ethical
4. Make sure data is used for public good and be mindful of residents' data
5. Be beneficial for all by using an outcomes-based approach
6. Camden's external partners should sign up to the Data Charter principles
7. Protect individuals' rights and privacy
8. Ensuring that the information we share about data use is clear and accessible

CONNECTED
BY DATA

People's Panel Recommendations



Connected by Data is the campaign for communities to have a powerful voice in the governance of data and AI. We work to put communities at the centre of data narratives, practices and policies.

We organised the **People's Panel on AI** to address the glaring absence of public voice in the UK AI Safety Summit, and to demonstrate the importance and feasibility of embedding deliberative public participation in future AI debates and decision making.

On 3rd November 2023, after deliberations at the AI Fringe, and considering outputs from the UK AI Safety Summit, the People's Panel on AI made the following recommendations:

- 1 A global governing body for AI** to bring together citizens, impartial experts and governments from across the world, and ensure regulatory collaboration that includes the global south.
- 2 A system of governance for AI in the UK that places citizens at the heart of decision making** drawing on input from scientists, researchers, ethicists, civil society, academia and industry to inform and provide evidence for government and citizens to then work together on decisions.
- 3 Awareness raising about AI across society.** From the classroom to the home. From the workplace to the community. Highlighting risks such as addiction to social media, as well as the opportunities that AI offers.
- 4 A safe transition, with training, to support people into a world of work alongside AI, ensuring no-one is left behind.**
- 5 A continued national conversation on AI, including retaining the People's Panel to keep public voice live in a fast-changing AI landscape.** We citizens can do jury service and as such are already trusted to make life-impacting and significant decisions.

- 6 Focus on inclusive collaboration,** to set out a vision of life where AI is used to enhance and balance human needs.
- 7 Stakeholders acting with transparency at all times.** An example of this might include a 'black box flight recorder' approach to AI models: protecting intellectual property but shared when things go wrong.

The importance of public deliberation

Public deliberation goes beyond methods like surveys, polling or user research in order to support extended dialogue with informed members of the public. This can:



Uncover values that the public want to see driving the use and governance of AI, helping to focus research and development on shared priorities.



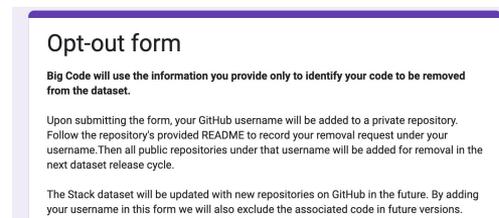
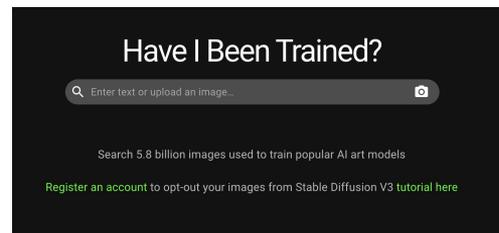
Build trust by delivering a deeper understanding of how AI tools and services impact everyday life in different communities.



Inform decision making enabling all stakeholders, including members of the public, to take an equal seat at the table when charting a path to responsible and ethical AI.

The Spectrum of Open Data Governance

- Transparency
- Versioning
- Training & Upskilling
- Consent: Opt-Out
- Consent: Opt-In
- Co-design & decision making



The Alan Turing Institute

Participatory data stewardship in AI

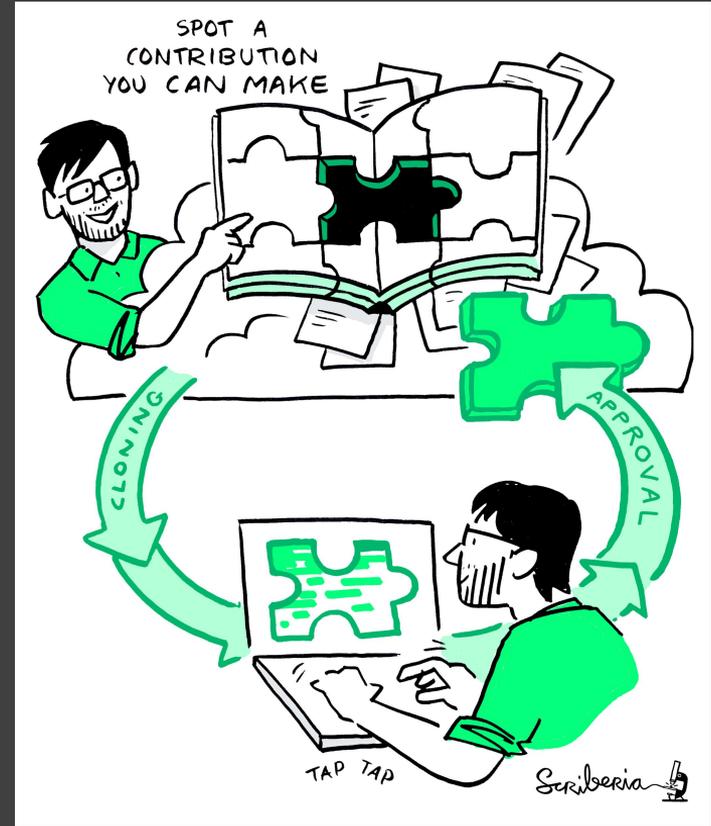
Jennifer Ding (she/her) and Anne Lee Steele (she/her)
jd Ding@turing.ac.uk asteel@turing.ac.uk



Meeting you where you are!

- i. Join the community
- ii. Learn a new skill
- iii. Share your skills
- iv. Collaborate with others
- v. Mentor others' contributions
- vi. Represent this community

We value your participation!



The Alan Turing Institute

Acknowledgements:

- The Turing Way team, Kirstie Whitaker, Malvika Sharan, Anne Lee Steele
- *The Turing Way* community, contributors & collaborators

- Book: the-turing-way.netlify.com
- Twitter: twitter.com/turingway
- Newsletter: tinyletter.com/TuringWay
- Slack: tinyurl.com/jointuringwayslack
- GitHub: github.com/alan-turing-institute/the-turing-way
- Original artwork by Scriberia: <https://doi.org/10.5281/zenodo.3332807>



Description of event - by Lisa

The final DFL Community Call in 2024 will feature Anne Lee Steele of the Alan Turing Institute, who will talk about Big Science and The Turing Way, two projects that offer alternative pathways for how AI can be developed beyond a handful of powerful companies, using open, participative methods.

BigScience is by now, a well known research initiative that created the BLOOM (BigScience Large Open-science Open-access Multilingual) Language Model. It is hosted by, and in partnership with HuggingFace.

The Turing Way is a distributed community of researchers and practitioners from data science related fields who actively contribute to a handbook of tools and best practices to ensure that conducting open, responsible, localised, and collaborative data science is "too easy not to do." I personally reference it all the time, especially to talk about the nuances in [open licensing as it relates to AI](#), from models, to datasets, to traditional software licenses.

The call will take place on Monday, December 11, at 3pm UTC. Sign up here to attend! In the meantime, you can learn more about Big Science and [The Turing Way](#) on the respective links, and take a look [at this paper](#) which features both. We look forward to seeing you there.

Notes (5 December)

- Update on the broader ecosystem - and the number of projects that have happened since then
- Presenting the ecosystem of work from before?
- Seeing how the projects has changed - but there was never really a clear line
- BLOOM and TTW are very niche parts of the ecosystem - what ecosystem are we pointing to? Is it open? Was focusing on participatory AI.
- Participatory ladder: <https://www.citizenshandbook.org/arnsteinsladder.html>
- Public comment → engaging the public is about people reading and providing thoughts, users we interview as opposed to being a part of the project
- BLOOM, Big Code, Luther, Aya
- Open science → Participatory → Open
- Going beyond specific field of science →
<https://the-turing-way.netlify.app/reproducible-research/licensing/licensing-ml.html>
<https://arxiv.org/abs/2301.08488>
-

Notes (5 December)

- Stewardship
- https://en.wikipedia.org/wiki/Data_governance
- https://en.wikipedia.org/wiki/Data_steward
- <https://medium.com/data-stewards-network/open-data-governance-and-open-governance-interplay-or-disconnect-d869e6f8d0ce>
- <https://opendatawatch.com/publications/open-data-for-official-statistics-history-principles-and-implentation/>
-

Notes (5 December)

Data Futures Lab: <https://foundation.mozilla.org/en/data-futures-lab/>

Projects from 2022 - early 2023

TTW Chapter: <https://the-turing-way.netlify.app/reproducible-research/licensing/licensing-ml.html>

Paper: <https://arxiv.org/abs/2301.08488>

MozFest workshop: https://www.youtube.com/watch?v=Z3gl_BUgF94

TTW Fireside Chat: <https://www.youtube.com/watch?v=e8EB1ocxt4>

JD:

https://docs.google.com/presentation/d/1G_TDYg6lfa6gKmOzifdfEPmxwfgM0h4-Dxe6ZHUmm8U/edit?usp=sharing

<https://zenodo.org/records/8028175>

ALS:

https://docs.google.com/presentation/d/1PPHJmUCf5Ws_It9TQyVvx45NNge0TWJ3DXzWtewHw-6A/edit#slide=id.g35f391192_00

https://docs.google.com/presentation/d/1-7zsHN_Bsp2S2m07S5Pa5awkK5_hINRxSMbxdGBT9PQ/edit#slide=id.g2522d0819ff_0_265