JOHAN LINÅKER (RISE)

Open Source AI

– An undefined divide between community and vendor development

First, what’s
Open Source Software?

Photo by Billy Huynh | https://unsplash.com/photos/v9bnfMCyKbg

Liberally licensed, Collaboratively developed software

Photo by Billy Huynh | https://unsplash.com/photos/v9bnfMCyKbg

Liberally licensed software

- Software available under an Open Source Software license
- License that follows the Open Source Definition and is approved by the Open Source Initiative (http://opensource.org)
- Anyone, for whatever reason, may inspect, use, modify the source code and redistribute
- Different conditions apply per license requirements

Photo by Fabian Fauth | https://unsplash.com/photos/1_EedIBc6jY

Collaboratively developed software

- Software developed as projects by networks of individuals and organizations, aka. Open Source Communities
- ”Members” of the community commonly both users and developers
- Are united by a common vision and goal around the Open Source Software.

Photo by Annie Spratt | https://unsplash.com/photos/QckxruozjRg

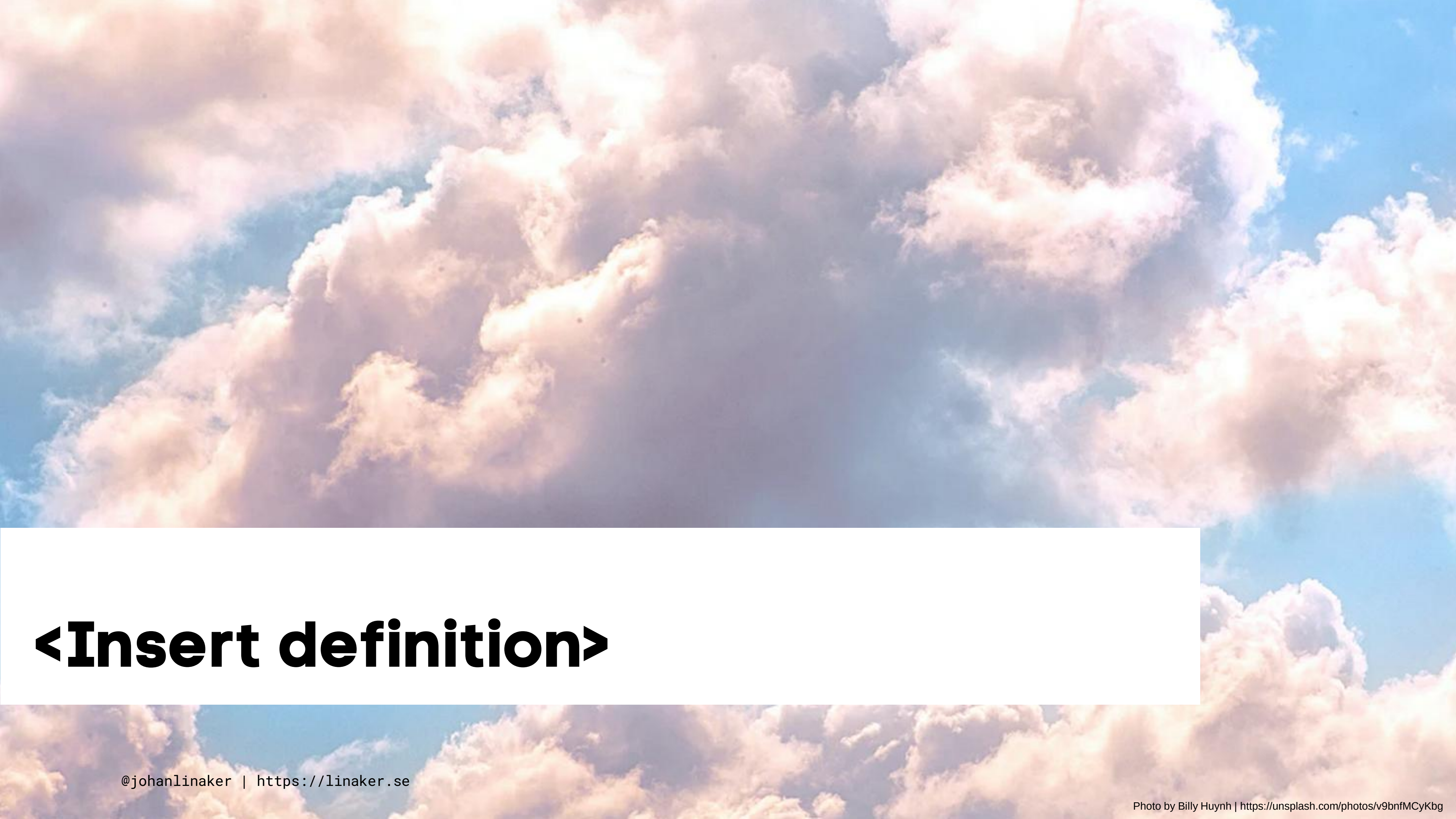
Open development process

- Informal structure pending on community
- Focus is on openness
  - Whoever can contribute
  - Influence through merit
  - Self-appointment of tasks
- Traditional development
  - Carried out in silos
  - Influence though hierarchical status
  - Appointment of tasks

Photo by Leon | https://unsplash.com/photos/Oalh2MojUuk

So, what's
Open Source AI?

Photo by Billy Huynh | https://unsplash.com/photos/v9bnfMCyKbg



<Insert definition>

Photo by Billy Huynh | https://unsplash.com/photos/v9bnfMCyKbg

Open Source AI (Systems)

- Definition being developed by the OSI
  - See: https://opensource.org/deepdive
- Open community effort working towards reaching consensus among key stakeholders
- Building on the four freedoms, and the AI systems definition by OECD

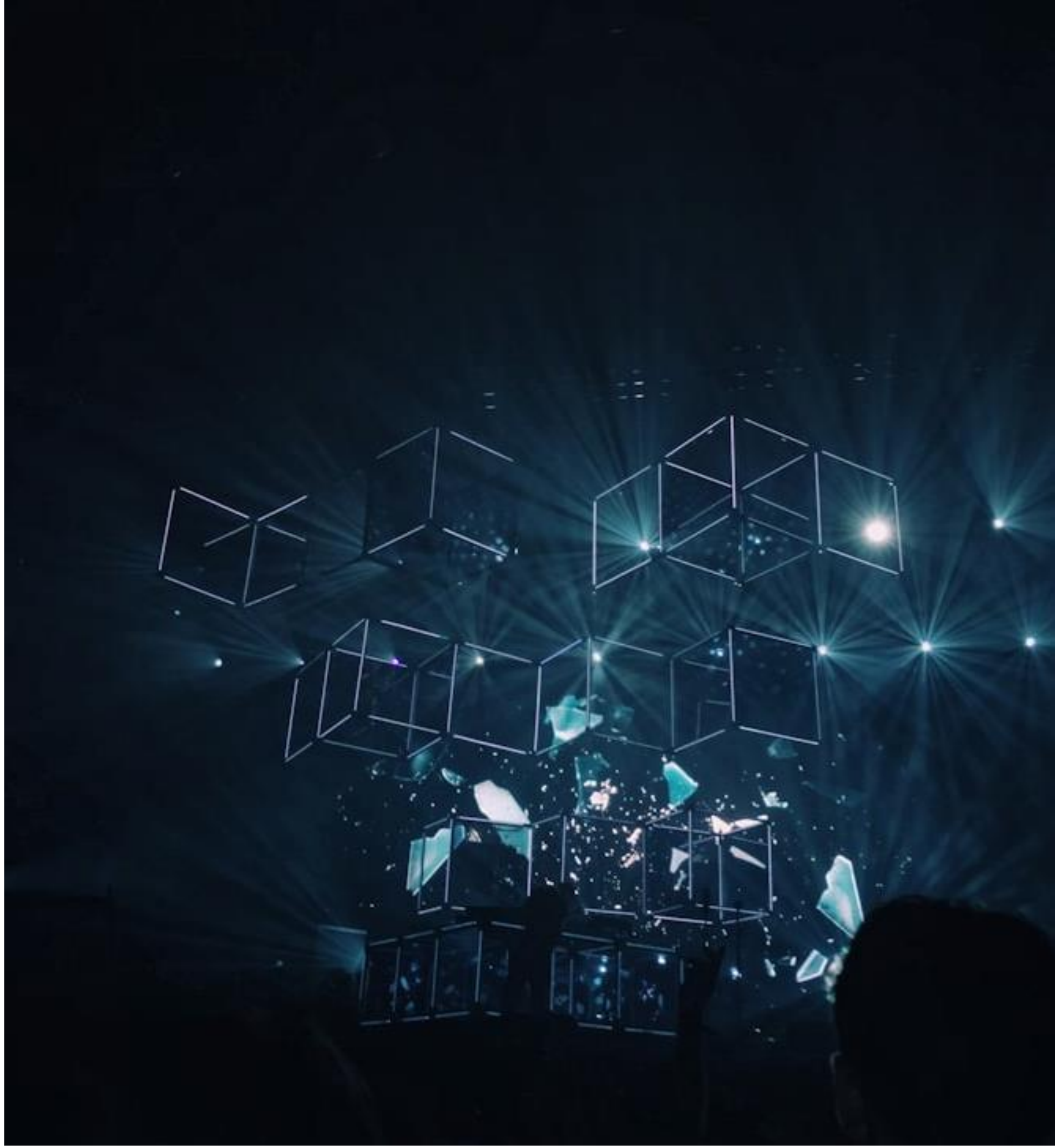


Photo by fabio | https://unsplash.com/photos/geometric-shape-digital-wallpaper-oyXis2kALVg

Open Source AI Systems

- *”To be Open Source, an AI system needs to be available under legal terms that grant the freedoms to:*
  - ***Use** the system for any purpose and without having to ask for permission.*
  - ***Study** how the system works and inspect its components.*
  - ***Modify** the system for any purpose, including to change its output.*
  - ***Share** the system for others to use with or without modifications, for any purpose.”*
- See: https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-5

Photo by Damir Kopezhanov | https://unsplash.com/photos/person-holding-green-and-silver-hand-tool-Z5tXjpTqwR8

Out of scope

- *“The Open Source AI Definition doesn’t say how to develop and deploy an AI system that is*
  - *ethical,*
  - *trustworthy or*
  - *responsible,*
- *although it doesn’t prevent it.*
- *We support the efforts to discuss the responsible development, deployment and use of AI systems, including through appropriate government regulation, as a separate conversation.”*
- See: https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-5

Photo by pine watt | https://unsplash.com/photos/person-hand-holding-photo-frame-3_Xwxya43hE

Many models referred to as "open source"

- But what is open? Are you able to
  - *Use the system for any purpose and without having to ask for permission?*
  - *Study how the system works and inspect its components?*
  - *Modify the system for any purpose, including to change its output?*
  - *Share the system for others to use with or without modifications, for any purpose?*

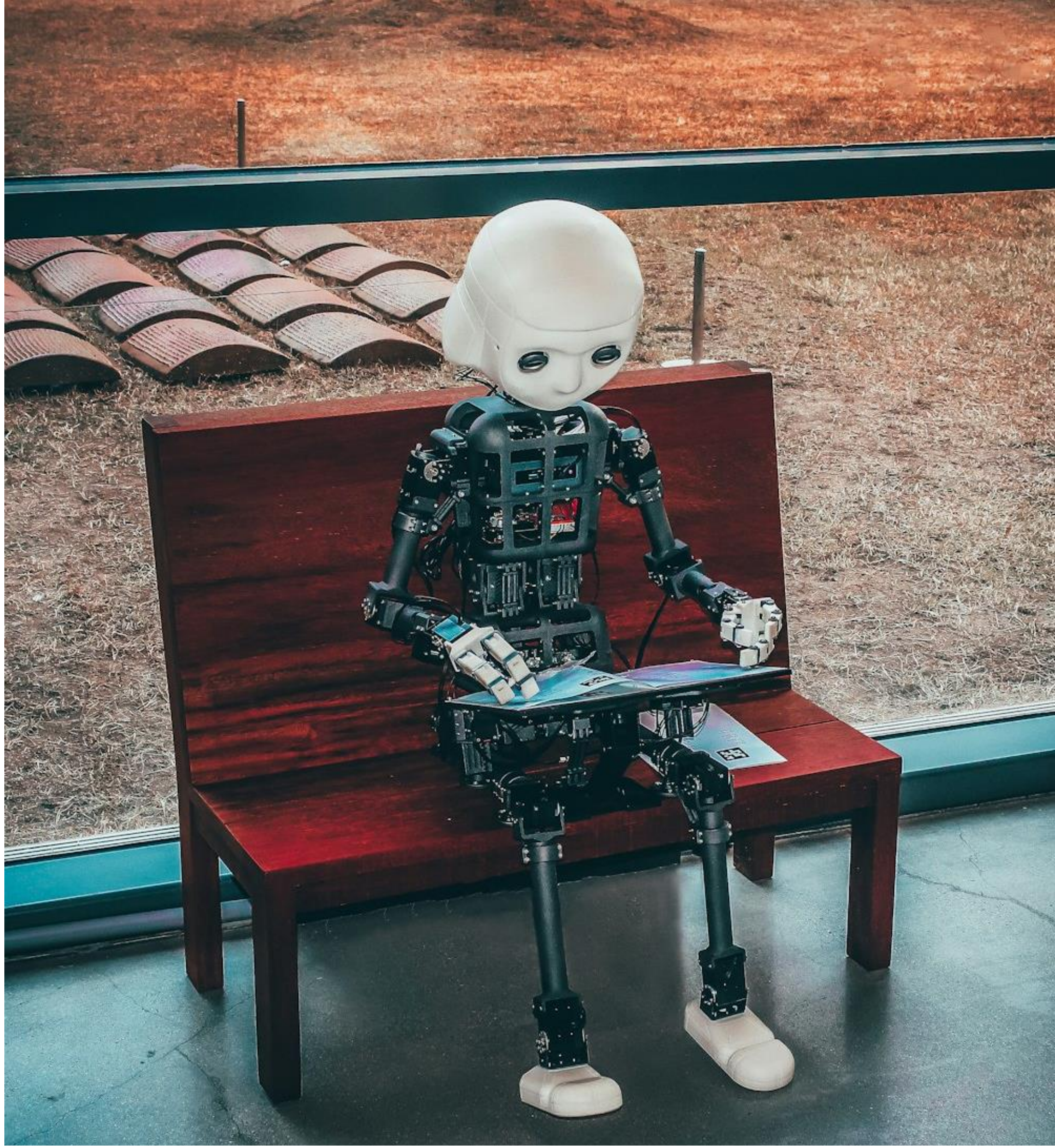
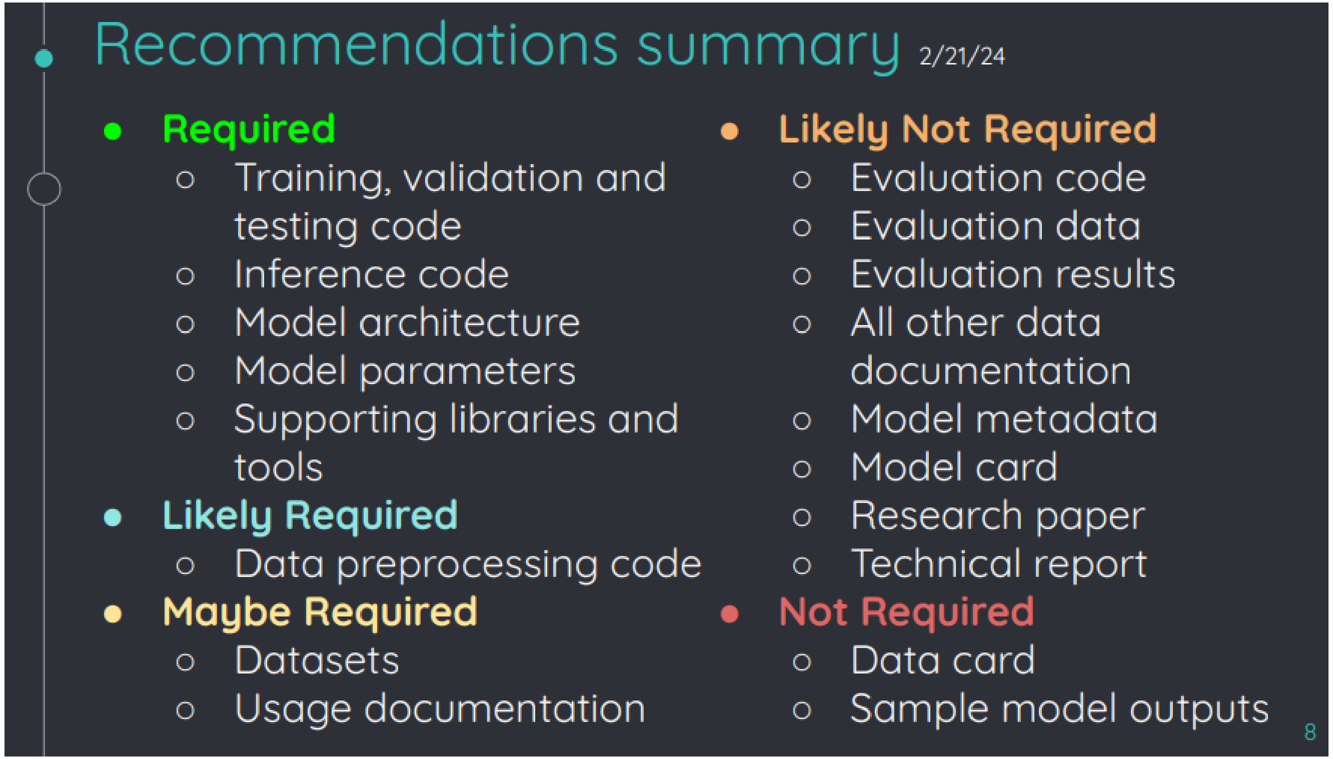


Photo by Andrea De Santis | https://unsplash.com/photos/black-and-white-robot-toy-on-red-wooden-table-zwd435-ewb4

“For [a machine learning system] to be open. I need to be able to question it,”
- Julia Ferraioli

https://aibusiness.com/ml/amazon-ml-expert-what-makes-machine-learning-truly-open-source

Ongoing evaluation of models



Recommendations summary 2/21/24

- **Required**
  - Training, validation and testing code
  - Inference code
  - Model architecture
  - Model parameters
  - Supporting libraries and tools
- **Likely Required**
  - Data preprocessing code
- **Maybe Required**
  - Datasets
  - Usage documentation
- **Likely Not Required**
  - Evaluation code
  - Evaluation data
  - Evaluation results
  - All other data documentation
  - Model metadata
  - Model card
  - Research paper
  - Technical report
- **Not Required**
  - Data card
  - Sample model outputs

https://opensource.org/wp-content/uploads/2024/02/osi_townhall_4.pdf

Collaborative development varies

- Presence and form for collaboration may differ based on the component:
  - data (e.g., for training, validation, and testing),
  - source code (e.g., for training and inference),
  - model architecture (e.g., for design choices and hyperparameters), and
  - documentation (e.g., for training procedure and evaluation).

Photo by Desola Lanre-Ologun | https://unsplash.com/photos/woman-and-man-sitting-in-front-of-monitor-IgUR1iX0mqM

Complexity in development

- Many components needed
- Development is costly, e.g.,
  - Collecting and processing data, and
  - Training the model
- Usually limited to resourceful, or venture-backed firms or research institutes

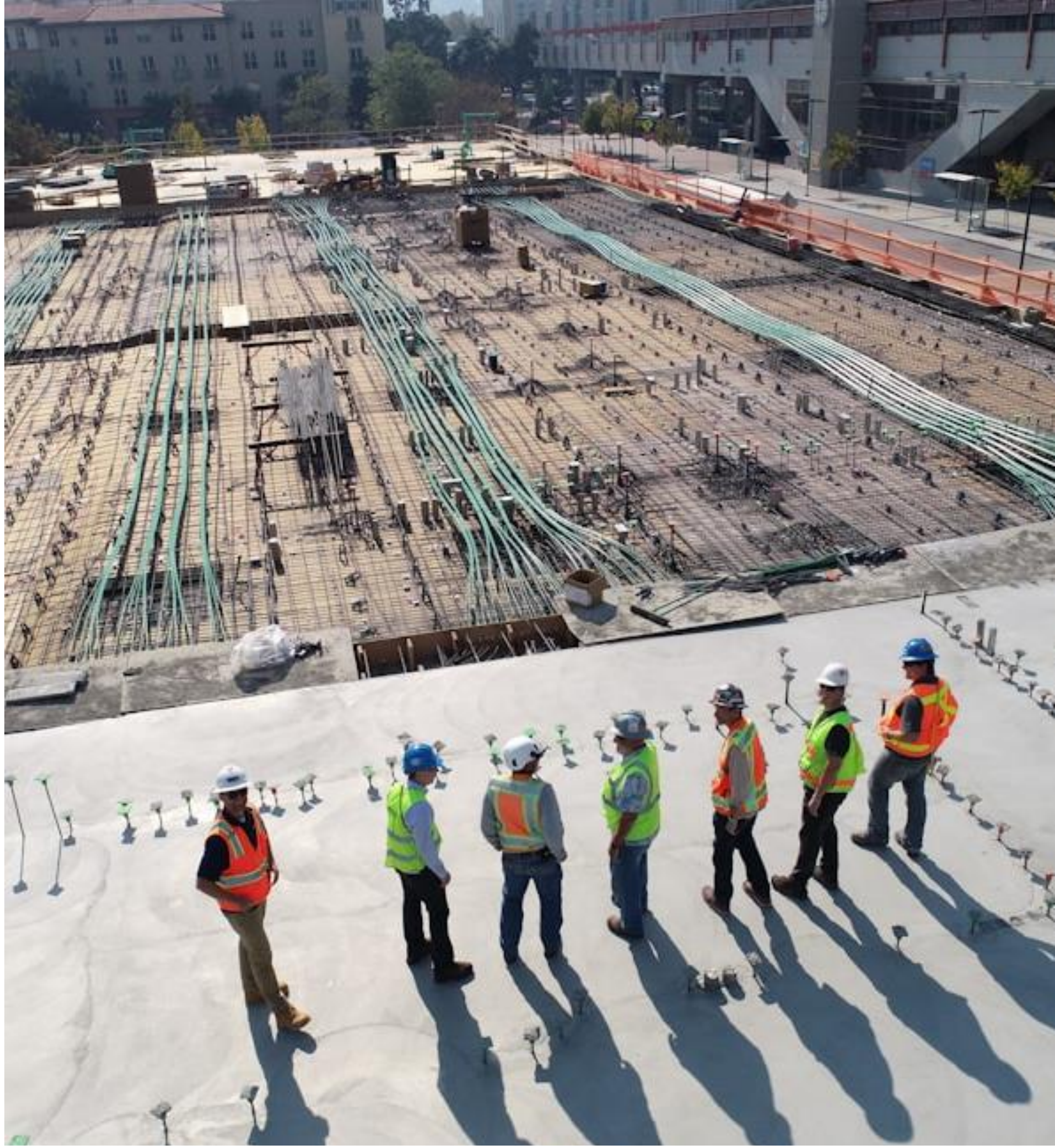


Photo by Scott Blake | https://unsplash.com/photos/seven-construction-workers-standing-on-white-field-x-ghf9LjrVg

Single-vendor vs. community models

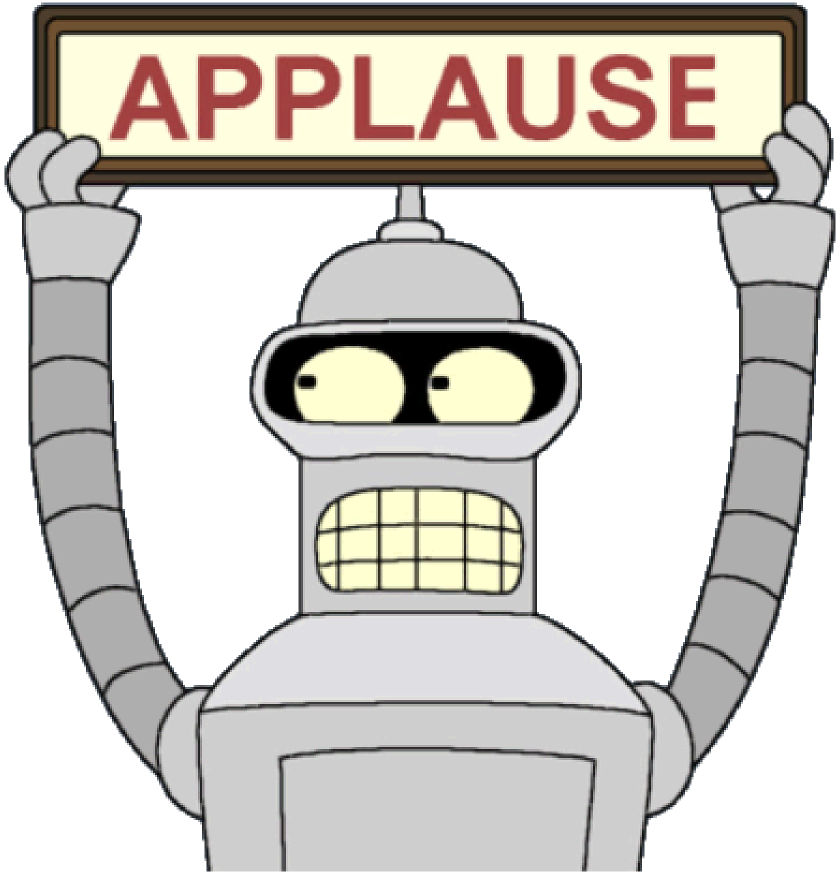
- A sliding scale without set definitions
  - Big tech: Llama by Meta,
  - Startups: Mistral, Aleph Alpha
  - Research Institutes: Falcon by Technology Innovation Institute, OLMo by AI2
  - ”Community”: EIutherAI (heavily backed) and BigScience Workshop (Hugging Face)

Photo by Katrin Hauf | https://unsplash.com/photos/white-and-brown-weighing-scale-at-0-XyGfxCVbjCA

Many outstanding questions

- What can we consider as open? And what parts needs to be available?
- How can the community aspects from open source software development expand to that of an AI system?
- How can we balance the tradeoff between benefits and challenges? For example,
  - Cost-efficiency, innovation, transparency, sovereignty vs.
  - Unethical, illigal use cases, propaganda, disinformation, integrity, national security

Photo by Evan Dennis | https://unsplash.com/photos/forest-trees-marked-with-question-marks-i--IN3cvEjg


APPLAUSE