Openness in AI Production

Upsides, Downsides, and Middle Ground Approaches

David Gray Widder, Carnegie Mellon University
Invited talk at the *Allen Institute for AI*, 6 April 2023

Born in Tillamook Oregon, grew up in Berlin & Singapore.

BS in CS and Liberal Arts at University of Oregon,
PhD candidate at at Carnegie Mellon University.

I spent much of the start of my PhD studying open source communities, and am now thinking about trust and ethics in AI.

I maintain a conceptual-realist **painting** practice on issues of **appropriation/inspiration**, and **surveillance/observation**.

I organize against **workplace surveillance**, and for **graduate student worker power**.

In Fall ’23, I’ll be a post doc at **Cornell Tech** in NYC, thinking about privacy and norms in AI systems. Collaborations welcome!

Today

Releasing AI openly can enable good, but also enable harm. What to do?

1. **Deepfakes** case study: does a community releasing software with prolific misuse feel accountable for this misuse? [Widder+, FAccT'22]

2. Generalizing this: how does the distributed "**AI supply chain**" complicate accountability? [Widder & Nafus, *Big Data & Society*]

3. A summary of **upsides** and **downsides of openness**, drawing on [Solaiman '23]

4. I advocate for **Middle Ground Approaches** to openness and for **Swiss cheese thinking**

Case Study: An Open Source Deepfake Tool

"*Free* software" has an explicitly **political, anti-corporate** history [Coleman '12]
which "*Open* source" eschews on "**pragmatic, business case grounds**"

In Open Source Developers (often volunteers) build software and make it freely available **without any restrictions on use**, **nor ability to know when it is used** ["non discrimination to persons, or field of endeavor", Open Source Initiative]

Open AI is open source? What openness means has changed

Deepfake videos spoof one person's face on another person's body, sometimes for satirical or artistic purposes, but **96% of online deepfakes are non-consensual porn of women**, causing job loss, anxiety, and illness [Ajder '15, Maddocks '20]

We interviewed 11 developers of an open source Deepfake creation tool about their sense of *Agency* and *Responsibility* to address downstream harm

Freedom 0: For Use By Anyone, For Anything?

Open source licenses enforce strong norms against restricting downstream use, which limited participants' feelings of agency to control downstream harmful use

Participants recognize that centralized control would help prevent misuse.

This maximizes agency for the software *Users*, but minimizes *Developers' agency to* decide what their system should be used for

Wider open source norms acts as a frame for understanding one's own ethical responsibility

"**I cannot stop people [from] using my software for stuff which I don't agree with**. [Open Source's] positive is also it's negative"

"Some of these server-based [Deepfake] apps [...] actually have filters [for] nude pictures. [...] That's a different kind of setup **because [...] of the centralized control, [...] they could implement filters**"

Setting and Enforcing Counter-Norms

After choosing an open source license, participants felt they had few other opportunities for agency

They **set norms against harmful uses**, in public statements where the code is downloaded and in communication channels

They **enforce community norms**, banning people from community forums and chatrooms who admit using it for porn

Intrinsically, some of this was motivated by own personal sense of ethics, but extrinsic: also to avoid **deplatforming** on GitHub and Discord platforms -> **Platform power**

Power: community leaders over members, and **platforms over community leaders**

"One of the points in our [public statement] is that [the project] is not for changing faces without consent […] Again, **we can't force our users to do anything**"

"So there's not a lot actively I can do. [...] But what I can do is discourage it and not [...] offer advice, and actively **block people looking for that advice within forums and domains that I have control over**"

“Technological Inevitability”

Participants view their role in developing Deepfake software as insignificant in the context of of available alternatives

Some viewed other “competitor” Deepfake tools as in a “race”

OpenAI: “competitive landscape”.
It is not a race to build the new thing.

Laws against Deepfake Videos, or restrictions on Deepfake tools were viewed as resisting their inevitable proliferation

Accepting the *Technological Imperative* “implies a suspension of ethical judgement or social control: individuals and society are seen as serving the requirements of a technological system which shapes their purposes” [Chandler ’95]

“We knew that that sort of thing was going to come about **whether or not I participated** in [this project]”

“If you ban something, it just goes underground”

“This genie’s out of the bottle.”

“**Nothing [can] stop the steam engine that is progress. And technology, it’s only getting better, faster**”

“Technological Neutrality”

Participants suggested that neutral tools can be used for good or bad, ethics up to the user

Reveals an *instrumentalist* view: tools are “value-neutral”

“Guns can be tossed around like frisbees”,[Selinger ’12] and you might use a frisbee to kill someone if you tried hard enough

But **affordances** built (or not built) make certain uses easier or harder, affecting how it is ***likely*** to be used

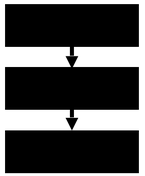
“You can’t really blame the project, cause **it’s like blaming the people that make the paint and the canvas**”

“For people that [want to make porn] they’re not very into [...] how it works. They just want the end result. [...] **Right now you have to do quite a bit of manual stuff** and you have to set up the whole environment”

The AI supply chain

Releasing AI openly can enable good, but also enable harm. What to do?

1. **Deepfakes** case study: does a community releasing software with prolific misuse feel accountable for this misuse? [Widder+, FAccT'22]

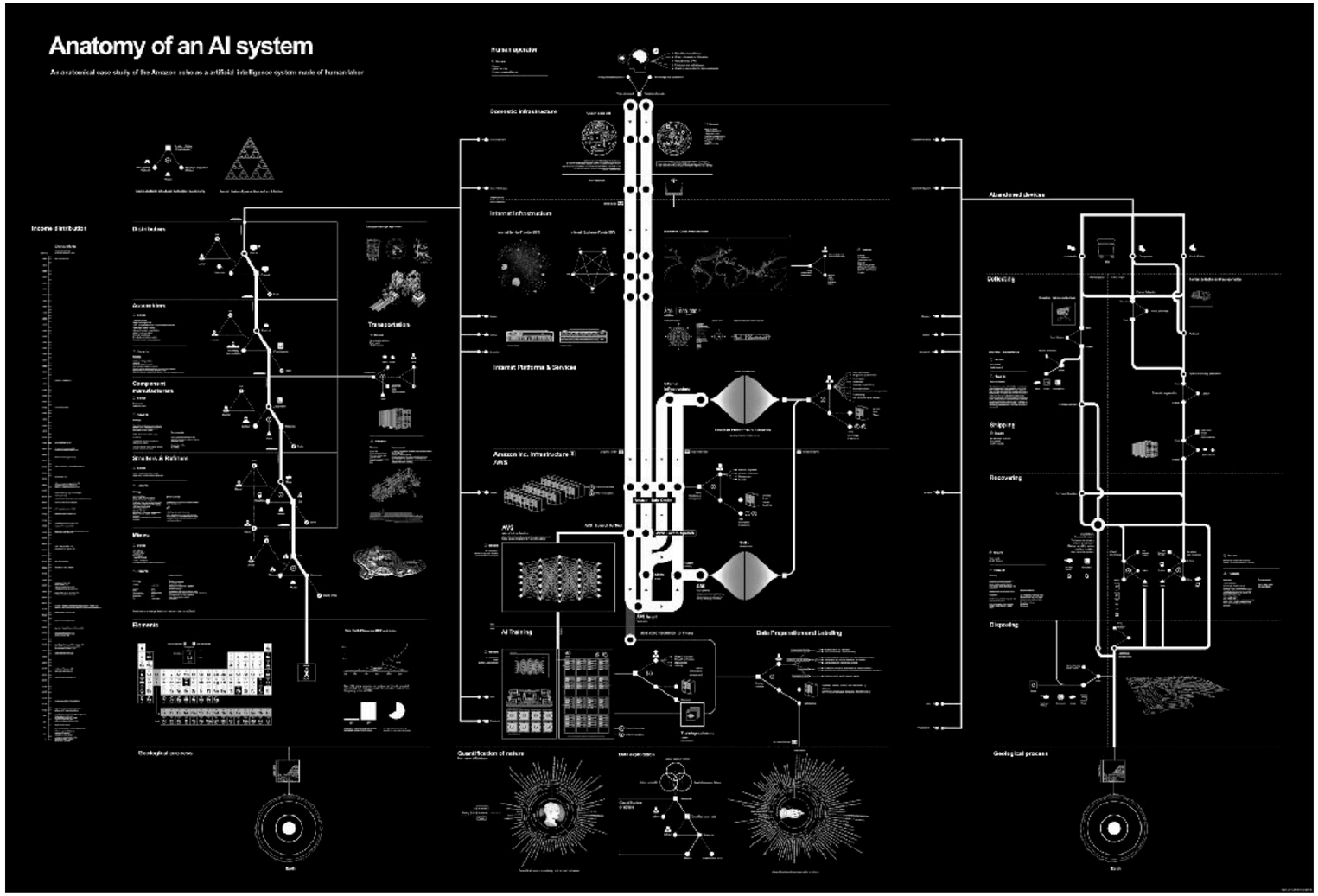


2. Generalizing this: how does the distributed "**AI supply chain**" complicate accountability? [Widder & Nafus, to appear in *Big Data & Society*]

3. A summary of **upsides and downsides of openness**, drawing on [Solaiman '23]

4. I advocate for **Middle Ground Approaches** to openness and for **Swiss cheese thinking**

Generalizing away from the Deepfake case: **modules be everywhere!**



Kate Crawford and Vladan Joler, Anatomy of AI

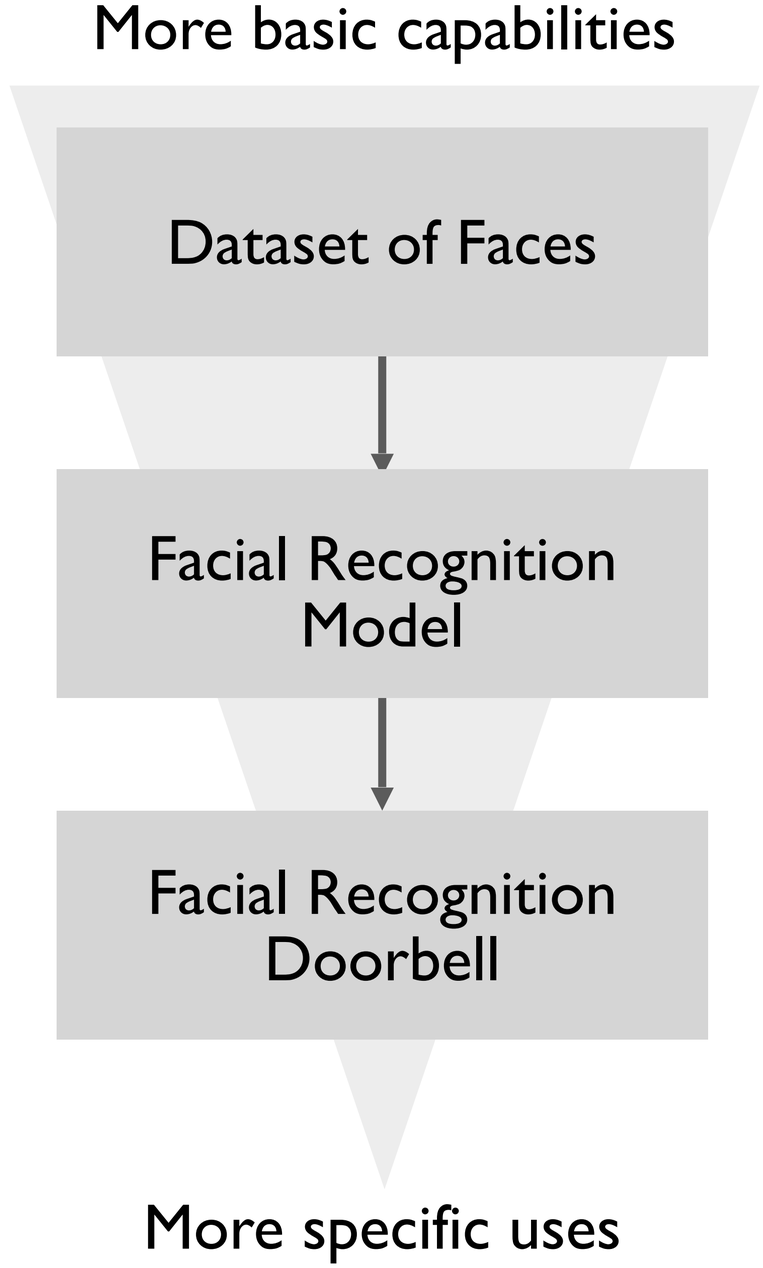
Modularity **is a technical and social practice that makes it easier to disavow harm.**

Software Modularity means users of your module need only understand modules interface but not internal workings, minimizes friction in reuse, **ideal of "general purpose"**

But **modularity has ethical implications**: allows disavowal of concerns outside the module, and division of labor

AI developers: Rely on **upstream** datasets and "fundamental" models, but **disavow and rarely scrutinize their flaws**

Release what they build openly, for anyone to use for anything **downstream**, while **disavowing these uses**



Implementation vs **Use**-Based Harms

Implementation-Based: harm inherent in how the system is built, eg gender biased credit allocation algorithms, or self driving cars not recognizing pedestrians in wheelchairs

Fixed with better datasets, or technofixes to make systems Fair, Accountable, or Transparent

Use-Based: harm inherent in how a system is used: drone strikes in Google's Project Maven, or Deepfake porn

The harm can't be fully eliminated by implementation fixes, or building the system differently.

Ethical AI narrowly scope to fairness and other *implementation harms*, because *use* is cast as an out-of-scope business decision,[Greene '19] or as "policing downstream use".

This explicit framing can help question whether use-based harms are really "out of scope"

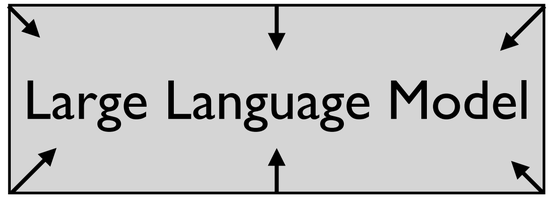
But! **Affordances** affect Use: Design affects how tools are *likely* to be used, even if unable to rule out harm altogether. But, this control is often disavowed.

Transparency & Accountability: for *Use* or *Implementation*?

For "**Implementation harms",** which can be fixed by changing data/ code, **open source is great**: [Grodzinsky '03]

you can inspect each line of code, each datapoint. This transparency helps scrutinize for and mitigate implementation harms

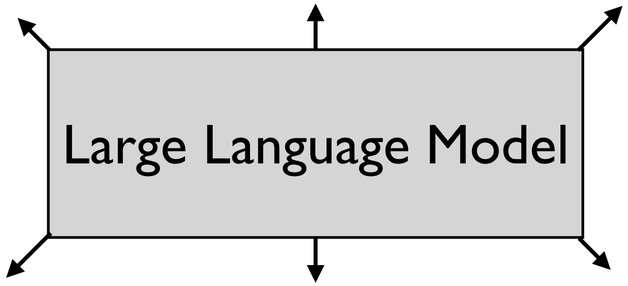
For systems determining major life outcomes (eg, recidivism, access to credit), open scrutiny supports fairer system, allows accountability



For **Use harms,** open source is problematic:

anyone can use your code without asking, so downstream uses are not transparent

nor can you enforce usage restrictions or hold users accountable for harm, so no *use accountability* for harm resulting from these uses



Participants accept responsibility for their module, but not how it is used.

More basic capabilities

Technique to regularize model accuracy

"a procedure […] a new way to optimize your machine learning model and depending on the data set you use, **the application domain you pick can be potentially endless**"

"nothing that would concern me [except] **general ways in which you can abuse machine learning**."

Model "benchmarks", "showcases", "demos"

"there is a very little interest in the […] the meaning of translation, but rather [more interest in] the **performance numbers**"

"an engineer working [in the] machine translation area, **he or she is aware of** […] **the bias**"

VR Training Software for Department of Defense

"It's a concern to me because there could be flaws in the code, security risks, quality risks, and effectively, **if anything goes wrong, it looks bad on us.**"

"We're not going to have a random [person] buy our products and begin using it. There's always going to be **some level of** […] **customer qualification**"

"**I get to turn a blind eye** to certain social aspects, because we have program managers that tend to be the buffer [between us and the user]"

More specific uses

The AI Supply Chain helps us *Locate Accountability*.

Responsibly developing tech must be “**a boundary-crossing activity**, taking place through the deliberate creation of situations that allow for the meeting of different partial knowledges”

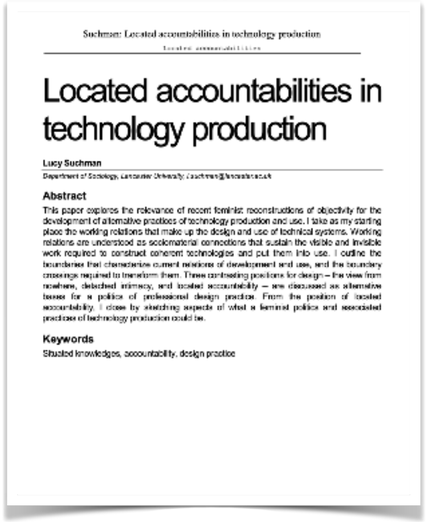
Requires a shift “**from a view of design as the creation of discrete devices,** or even networks of devices, to a view of systems development as **entry into the networks of working relations**”

What holds ethics together is outside of the modularized supply chain: personal and company reputation reputation concerns, delivering value to end users, seeing them as people.

What if we thought of a chain of modules as something that enables a view from somewhere, to see where action can take place?

The AI Supply Chain view situates even relatively “general purpose” AI libraries or frameworks in the context of the downstream harms they potentiate or constrain.

Its messy, but we hold suppliers of physical goods accountable for their supply chain, eg upstream: Nike, and downstream: weapons export.


Suchman: Located accountabilities in technology production

Located accountabilities in technology production

Lucy Suchman

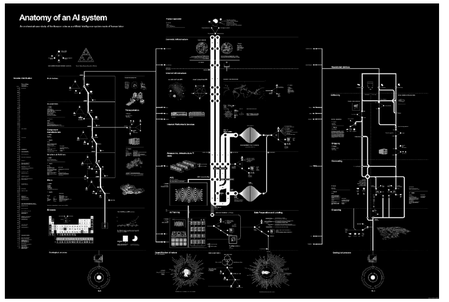
Department of Sociology, Lancaster University, l.suchman@lancaster.ac.uk

Abstract

This paper explores the relevance of recent feminist reconstructions of objectivity for the development of alternative practices of technology production and use. I take as my starting place the working relations that make up the design and use of technical systems. Working relations are understood as sociomaterial connections that sustain the visible and invisible work required to construct coherent technologies and put them into use. I outline the boundaries that characterize current relations of development and use, and the boundary crossings required to transform them. Three contrasting positions for design – the view from nowhere, detached intimacy, and located accountability – are discussed as alternative bases for a politics of professional design practice. From the position of located accountability, I close by sketching aspects of what a feminist politics and associated practices of technology production could be.

Keywords

Situated knowledges, accountability, design practice



Upsides and downsides of openness

Releasing AI openly can enable good, but also enable harm. What to do?

1. **Deepfakes** case study: does a community releasing software with prolific misuse feel accountable for this misuse? [Widder+, FAccT'22]

2. Generalizing this: how does the distributed "**AI supply chain**" complicate accountability? [Widder & Nafus, to appear in *Big Data & Society*]

3. A summary of **upsides and downsides of openness**, drawing on [Solaiman '23]

4. I advocate for **Middle Ground Approaches** to openness and for **Swiss cheese thinking**

The Gradient of Generative AI Release: Methods and Considerations

Irene Solaiman
Hugging Face
irene@huggingface.co

Abstract

As increasingly powerful generative AI systems are developed, the release method greatly varies. We propose a framework to assess six levels of access to generative AI systems: fully closed; gradual or staged access; hosted access; cloud-based or API access; downloadable access; and fully open. Each level, from fully closed to fully open, can be viewed as an option along a gradient. We outline key considerations across this gradient: release methods come with tradeoffs, especially around the tension between concentrating power and mitigating risks. Diverse and multidisciplinary perspectives are needed to examine and mitigate risk in generative AI systems from conception to deployment. We show trends in generative system release over time, noting closedness among large companies for powerful systems and openness among organizations founded on principles of openness. We also enumerate safety controls and guardrails for generative systems and necessary investments to improve future releases.

OpenAI's U-turn on Openness: "*we were wrong*"

How it started…
2015

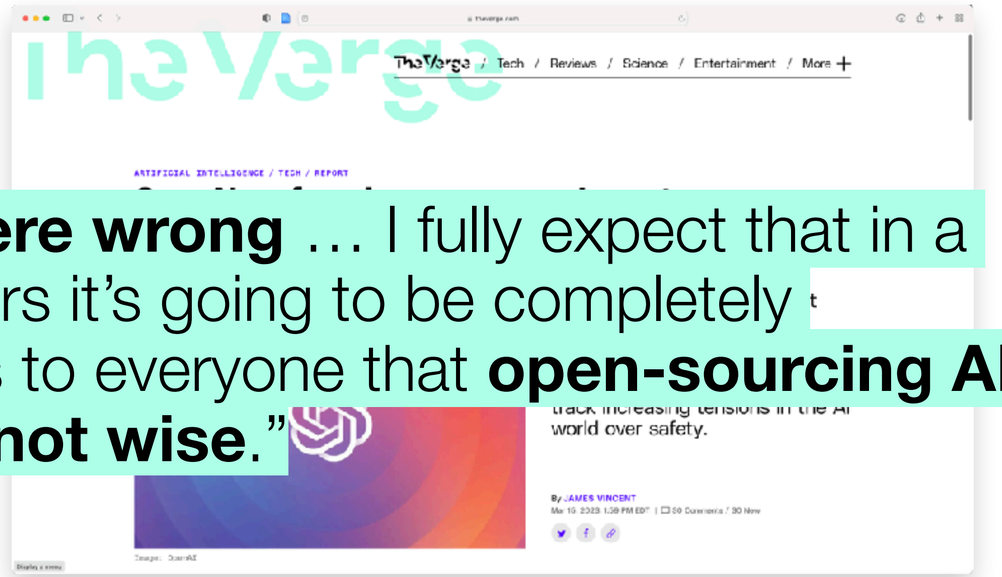
→

…How its going
2023

OpenAI

Because of AI's surprising history, it's hard to predict when human-level AI might come within reach. When it does, it'll be important to have a leading research institution which can prioritize a good outcome for all over its own self-interest.

We're hoping to grow OpenAI into such an institution. As a non-profit, our aim is to build value for everyone rather than shareholders. Researchers will be strongly encouraged to publish their work, whether as papers, blog posts, or code, and our patents (if any) will be shared with the world. We'll freely collaborate with others across many institutions and expect to work with companies to research and deploy new technologies.



"**We were wrong** … I fully expect that in a few years it's going to be completely obvious to everyone that **open-sourcing AI is just not wise**."

2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly

"Given both the **competitive landscape** and the **safety implications** of large-scale "models like GPT-4, this **report contains no further details** about the architecture (including model size), hardware, training compute, dataset e, construction, training method, or similar"

technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.[2] We plan to make further technical

👎 Downsides of AI Openness

👎 **The big one**: people can **misuse** the system in ways that cause harm (eg, Deepfake porn! Spam! Fake news!)

👎 If you think **enabling AI development can be itself harmful** (for example, by automating jobs and leading to increased economic inequality), openness may lead to AI being developed faster

❌ If you believe “Artificial General Intelligence” is possible *and* undesirable (I don’t think it is possible), openness may mean AGI is developed faster

❓ For discussion later: what else goes here?

👍 Ethical Upsides of AI Openness

👍 Openness may **“democratize” access to powerful AI**, thereby reducing concentration of power

⚠️ Caveat: access to data or code doesn’t mean you have the **compute** to make use of it, or the **power or skills** to put AI to use in the real world

👍 Openness is important for **replicability** through the **Scientific Method**, an important way that we agree about what is true in the world.

➡️ Implication: I don’t think non-open AI should be accepted into scientific literature

👍 More perspectives perspectives “in the room” to **scrutinize the system for harms**, enameling wider scrutiny, especially from perspectives not highly represented among AI developers

❌ “Competitive landscape”: aversion to sharing your IP openly is not an *ethical* argument argument in my view, though it may be an economic argument / (dis)incentive. I look poorly on “nonprofits” like “Open”AI which make this argument.

❌ Nation state concerns: (ie, “but what about China?”) I don’t see this as a convincing ethical argument, as I think AI nationalism is on balance unethical

❓ For discussion later: what else goes here?

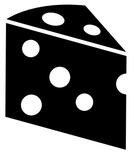
Middle ground approaches and Swiss cheese thinking

Releasing AI openly can enable good, but also enable harm. What to do?

1. **Deepfakes** case study: does a community releasing software with prolific misuse feel accountable for this misuse? [Widder+, FAccT'22]

2. Generalizing this: how does the distributed "**AI supply chain**" complicate accountability? [Widder & Nafus, to appear in *Big Data & Society*]

3. A summary of **upsides and downsides of openness**, drawing on [Solaiman '23]



4. I advocate for **Middle Ground Approaches** to openness and for **Swiss cheese thinking**

Overview of Middle Ground Approaches to AI Openness

1. Licensing for ethics
2. Norm setting and community governance

3. Technological restrictions
4. Usage monitoring

5. Release gating
6. Staged or partial release

? For discussion later: what else goes here?

Middle Ground Approaches

1. Licensing for ethics

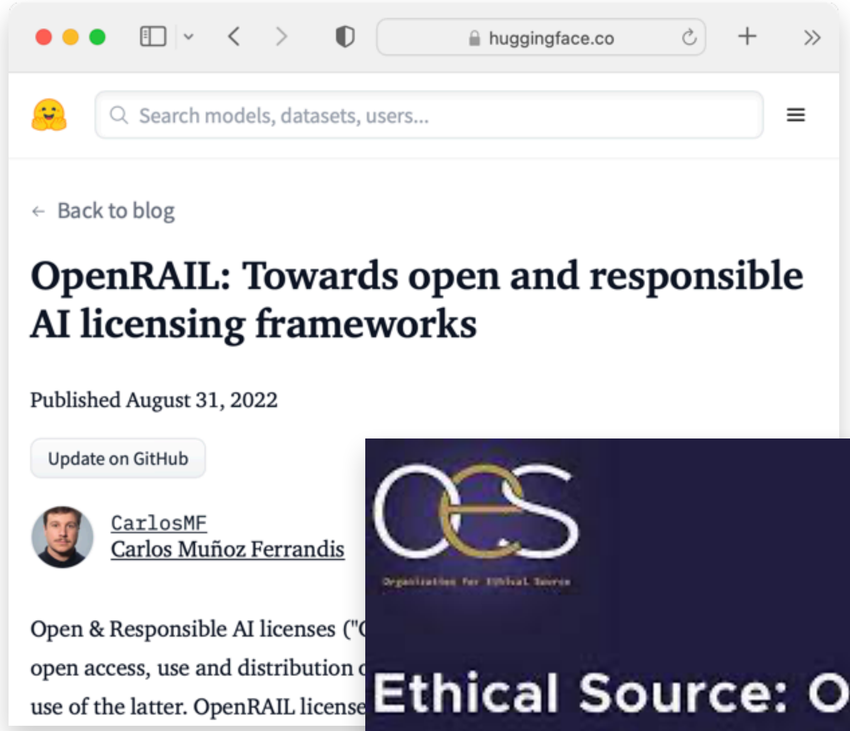
Hugging Face's Open Responsible AI Licenses (RAIL)

Organization for Ethical Source's Hippocratic License

Behavioral Use Licensing [Contractor+ 2022]

Objection: Enforcement may be tricky.
Rebuttal: misuse isn't just done by individuals, who may ignore legal licenses and use it in secret. Companies can misuse things too, and they have lawyers who listen to licenses.
Also: Licenses help set norms, which themselves are powerful!

Objection: Ethical licenses may dissuade adoption, eg how some companies won't use "viral" GPL licenses.
Rebuttal: Do we want wide adoption, if some of that adoption is for unethical uses?



2. Norm setting and community governance

No, they aren't foolproof, but norms are powerful! They set default behavior, scaffold agreement what is ok and what isn't, are a primary way that communities set bounds of ethical behavior.

How to set norms:
Public platform: As researchers at ~fancy~ institutions, we have **powerful** personal and institutional platforms. We can use this to promote certain uses we believe are beneficial, and criticize uses we believe are harmful. We ought to use this!

Community governance: Have an acceptable use policy for support forums, etc, and ban who use your tech for harm.

Licenses set strong norms! Right now, open source sets norm about disavowing use, but they can also be used to set other norms! (see previous slide)

3. Technological restrictions

Of varying degrees of hardness.

Soft restrictions prevent casual misuse. Eg: open source deepfake software could come with code that detects and quits when it detects pornographic use.

A technically skilled user could remove this code, but not everyone has technical skill!

Possible hard technological restrictions: cryptographic key to use software, blockchain

"For people that [want to make porn] they're not very into [...] how it works. They just want the end result. [...] **Right now you have to do quite a bit of manual stuff** and you have to set up the whole environment"

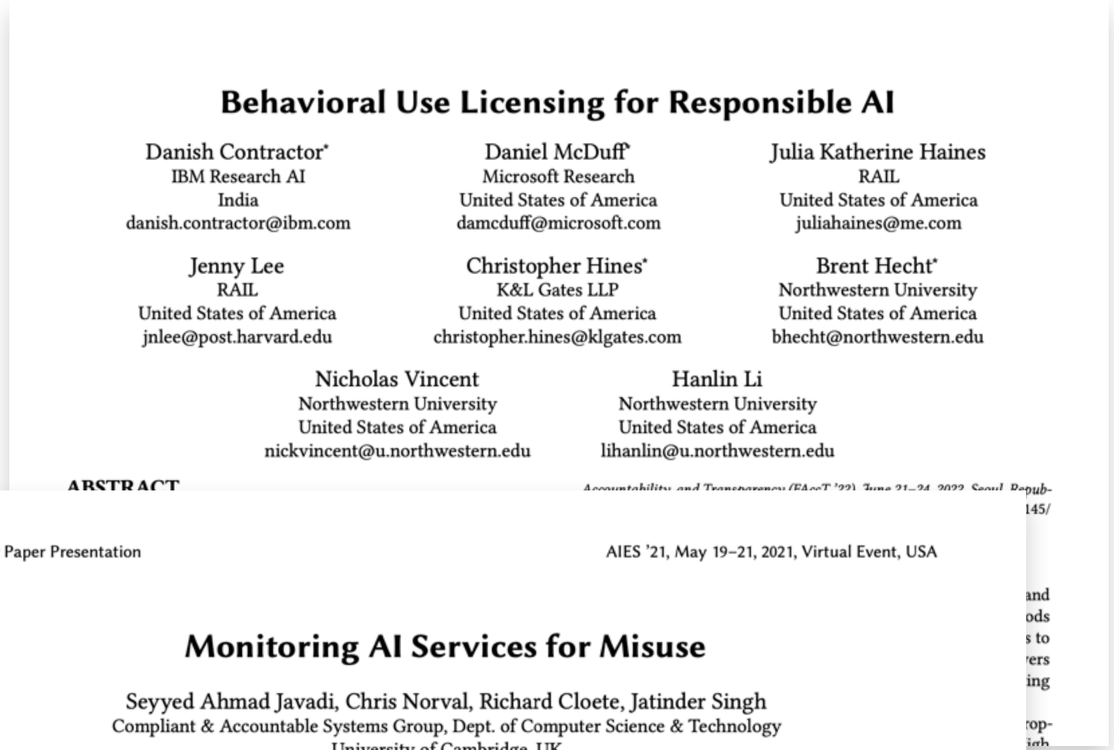
Middle Ground Approaches

4. Usage monitoring

May work better on SaaS

For code, some have suggested blockchain/ DRM/ cryptographic approaches. Unsure about this.

Also may have ethical downsides, depends on trusting the monitor!



Behavioral Use Licensing for Responsible AI

Danish Contractor*
IBM Research AI
India
danish.contractor@ibm.com

Daniel McDuff*
Microsoft Research
United States of America
damcduff@microsoft.com

Julia Katherine Haines
RAIL
United States of America
juliahaines@me.com

Jenny Lee
RAIL
United States of America
jnlee@post.harvard.edu

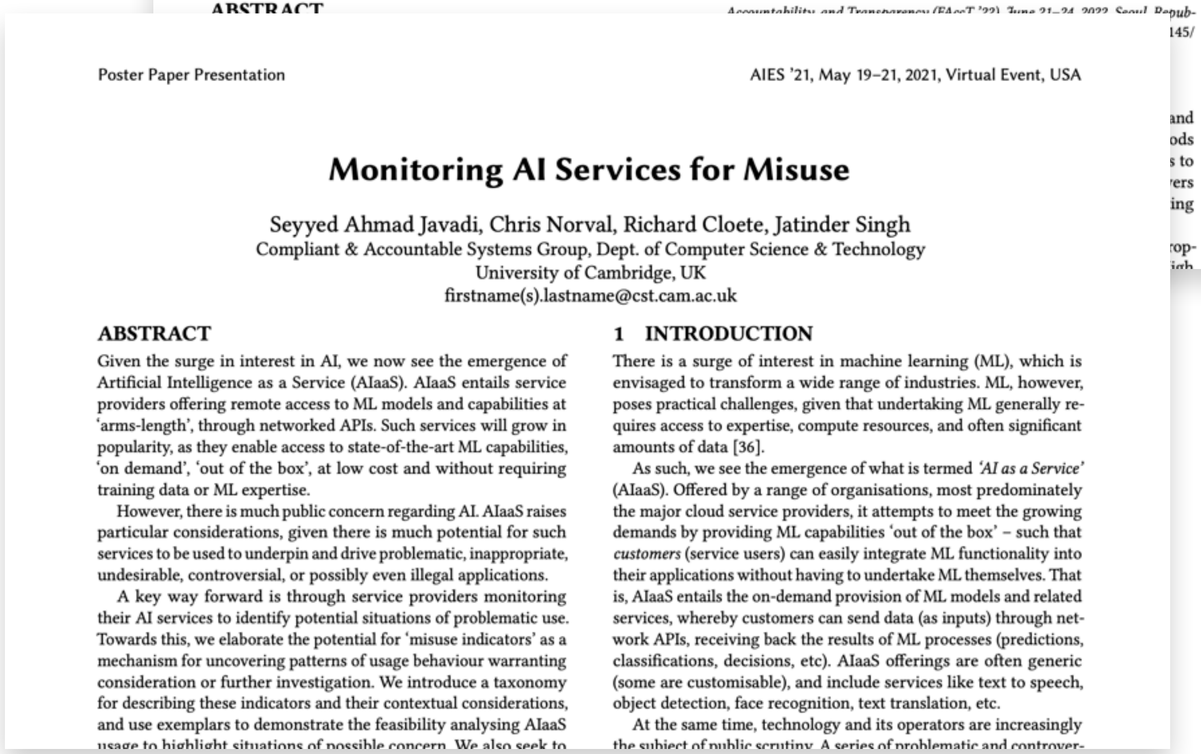
Christopher Hines*
K&L Gates LLP
United States of America
christopher.hines@klgates.com

Brent Hecht*
Northwestern University
United States of America
bhecht@northwestern.edu

Nicholas Vincent
Northwestern University
United States of America
nickvincent@u.northwestern.edu

Hanlin Li
Northwestern University
United States of America
lihanlin@u.northwestern.edu

ABSTRACT



Poster Paper Presentation

AIES '21, May 19–21, 2021, Virtual Event, USA

Monitoring AI Services for Misuse

Seyyed Ahmad Javadi, Chris Norval, Richard Cloete, Jatinder Singh
Compliant & Accountable Systems Group, Dept. of Computer Science & Technology
University of Cambridge, UK
firstname(s).lastname@cst.cam.ac.uk

ABSTRACT

Given the surge in interest in AI, we now see the emergence of Artificial Intelligence as a Service (AIaaS). AIaaS entails service providers offering remote access to ML models and capabilities at 'arms-length', through networked APIs. Such services will grow in popularity, as they enable access to state-of-the-art ML capabilities, 'on demand', 'out of the box', at low cost and without requiring training data or ML expertise.

However, there is much public concern regarding AI. AIaaS raises particular considerations, given there is much potential for such services to be used to underpin and drive problematic, inappropriate, undesirable, controversial, or possibly even illegal applications.

A key way forward is through service providers monitoring their AI services to identify potential situations of problematic use. Towards this, we elaborate the potential for 'misuse indicators' as a mechanism for uncovering patterns of usage behaviour warranting consideration or further investigation. We introduce a taxonomy for describing these indicators and their contextual considerations, and use exemplars to demonstrate the feasibility analysing AIaaS usage to highlight situations of possible concern. We also seek to

1 INTRODUCTION

There is a surge of interest in machine learning (ML), which is envisaged to transform a wide range of industries. ML, however, poses practical challenges, given that undertaking ML generally requires access to expertise, compute resources, and often significant amounts of data [36].

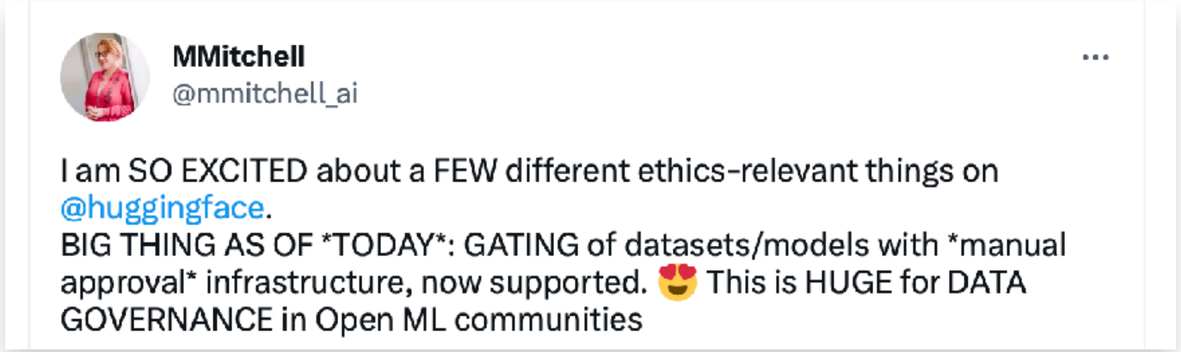
As such, we see the emergence of what is termed *'AI as a Service'* (AIaaS). Offered by a range of organisations, most predominately the major cloud service providers, it attempts to meet the growing demands by providing ML capabilities 'out of the box' – such that *customers* (service users) can easily integrate ML functionality into their applications without having to undertake ML themselves. That is, AIaaS entails the on-demand provision of ML models and related services, whereby customers can send data (as inputs) through network APIs, receiving back the results of ML processes (predictions, classifications, decisions, etc). AIaaS offerings are often generic (some are customisable), and include services like text to speech, object detection, face recognition, text translation, etc.

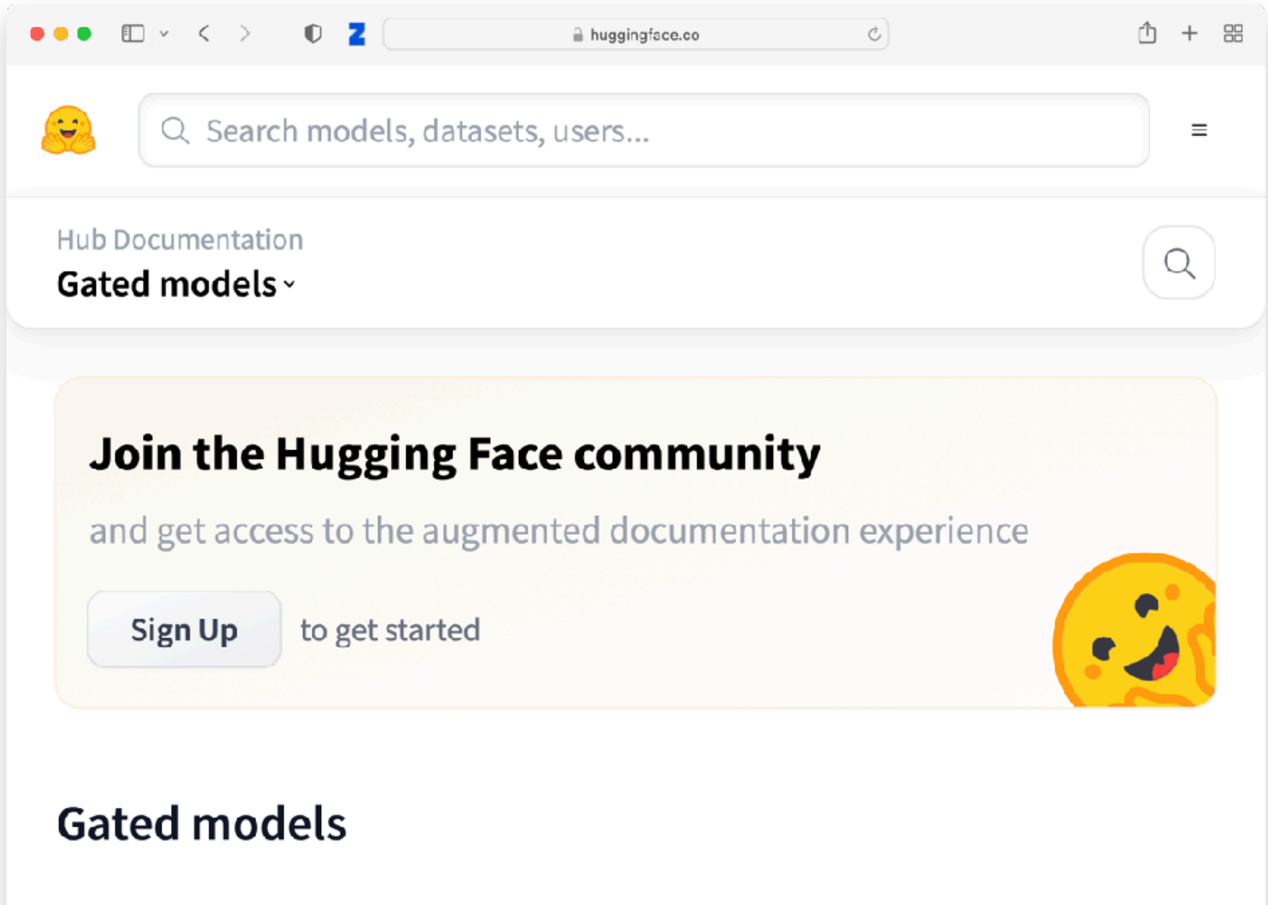
At the same time, technology and its operators are increasingly the subject of public scrutiny. A series of problematic and controver-

5. Release gating

Only releasing to certain people, eg, members of the scientific community, those who you trust.

Allows you to set and enforce norms!





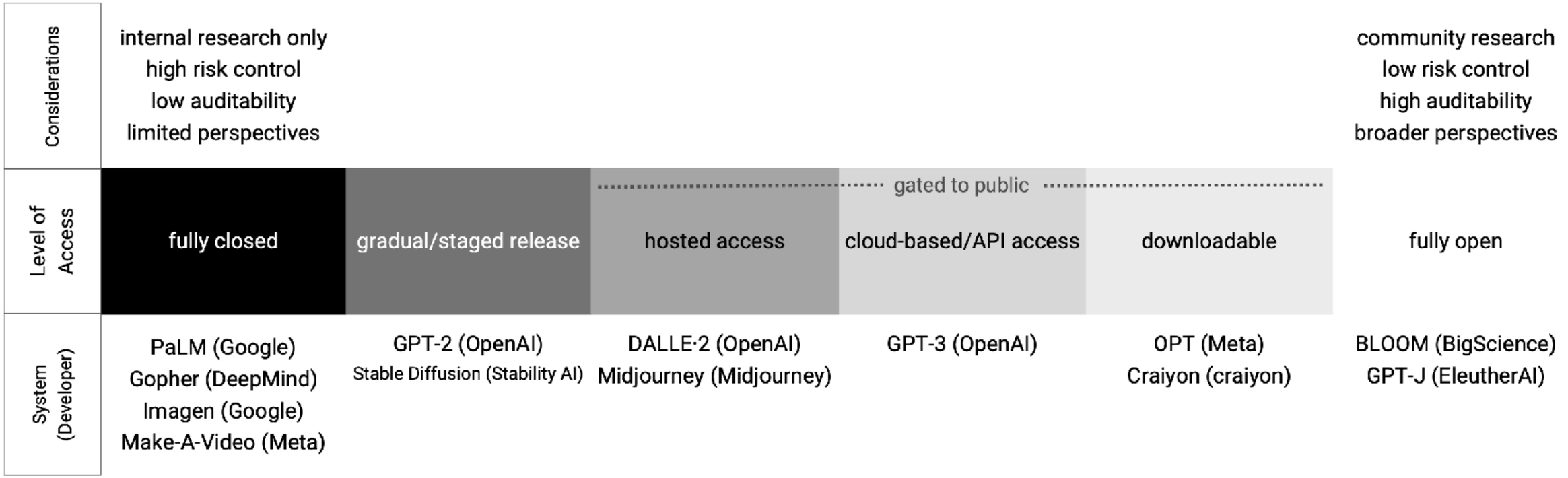
6. Staged or partial release

Partial: Eg, Weights, but not code. Data, but not models.

Can allow some scrutiny, but prevent some misuse.

Staged: release more parts publicly (eg: API access → data → weights → code) over time

Can allow a “see how things go” approach, more caution, care as new tech is released into world.



From Solaiman '23

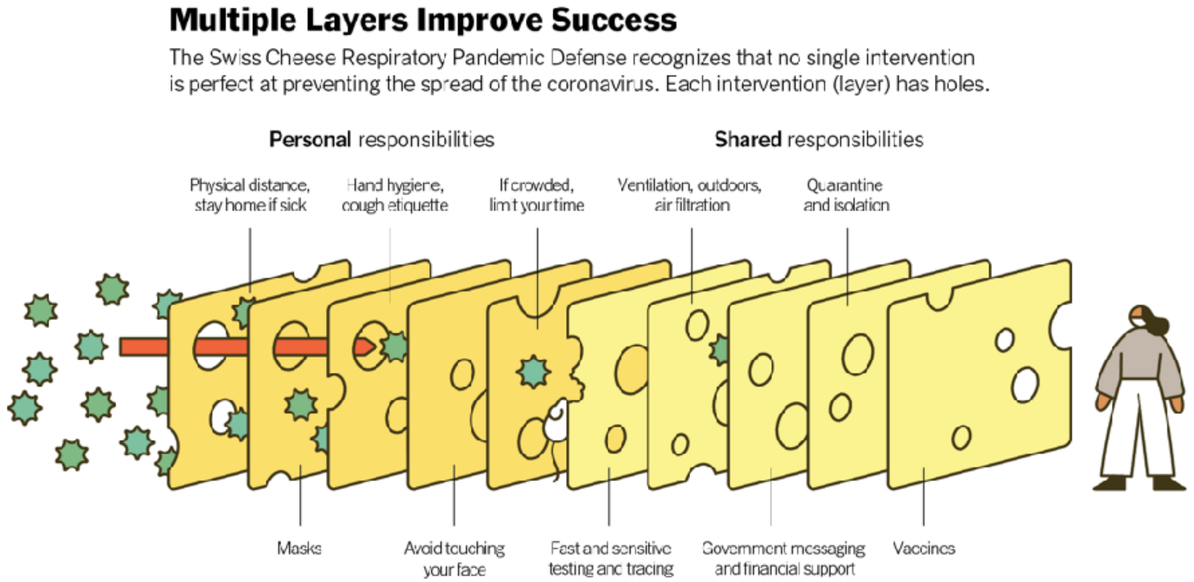
Preventing Misuse is about **shades of grey** and **Swiss cheese thinking**!

~~Silver bullets~~ → shades of grey: Even if you can't stop all misuse, middle ground approaches are better than doing nothing.

Swiss cheese thinking!

Learning from **software security:** No measure can make a system 100% secure. But multiple layers of security can stop some hacks, which is better than stopping none.

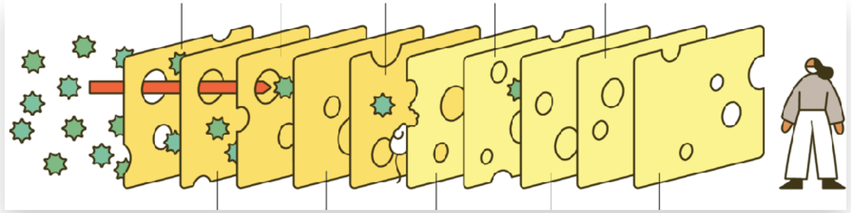
Learning from **public health:** Masks are not 100% effective against covid, but stop some cases, and that is meaningful.



Source: Adapted from Ian M. Mackay (virologydownunder.com) and James T. Reason. Illustration by Rose Wong

On balance, I believe: we should **default towards openness** in AI production, but **NOT let us think this allows us to disavow misuse.**

We must **adopt middle ground** approaches to openness, and further develop new ones. This is a hard problem, taking time and effort to get right.



Underlying papers:
davidwidder.me/deepfakes.pdf
davidwidder.me/supply-chain.pdf

I'd love to connect!
dwidder@cmu.edu • @davidthewid •
@davidthewid@hci.social • www.davidwidder.me

➡ SOON Cornell Tech in NYC: norms and privacy in AI.

Feedback and **critique please**!

👍🏼 Upsides of openness

- May **reduce concentration of power**
  Caveat: **compute, power and skills** needed to make this meaningful
- **Replicability, Scientific Method**
  -> non-open AI papers should **not** be accepted as science
- **Wider scrutiny** by more (perhaps disenfranchised) perspectives, since **AI can affect one's life chances**

❎ "Competitive landscape"

❎ nation-state concerns

👎🏼 Downsides of openness

- The big one: **misuse**!
- Some think **AI development can be itself harmful**

❎ If you believe AGI is possible and harmful, openness may mean AGI is developed faster

Middle ground approaches:

- Licensing for ethics
- Norm setting
- Technological restrictions
- Usage monitoring
- Release gating
- Staged or partial release