The Artefacts of Intelligence: Governing scientists' contribution to AI proliferation

Toby Shevlane

Faculty of Law (Centre for Socio-Legal Studies), University of Oxford

Supervisor: Dr Bettina Lange

Abstract: This DPhil dissertation is about attempts to govern how artificial intelligence (AI) researchers share their work. There is growing concern that the software artefacts built by AI researchers will have adverse impacts on society if made freely available online. AI research is a scientific field, and openly sharing these artefacts is routine and expected, as part of the functioning of the scientific field. Recently, there have been a number of occasions where members of the AI research community have trialled new ways of sharing their work, in response to their concerns that it poses risks to society. The case study follows: the 'staged release' of the GPT-2 language model, where more capable models were gradually released; the platform through which researchers and developers could access GPT-3, the successor to GPT-2; and a wave of new ethics regimes for AI conference publications. The study relies on 42 qualitative interviews with members of the AI research community, conducted between 2019 and 2021, as well as many other publicly available sources such as blog posts and Twitter. The aim is to understand how concerns about risk can become a feature of the way AI research is shared. Major themes are: the relationship between science and society; the relationship between industry AI labs and academia; the interplay between AI risks and AI governance regimes; and how the existing scientific field provides an insecure footing for new governance regimes.

Word count: 75, 378

INTRODUCTION

The progress of science and technology has revolutionised society time and time again, and yet this process is usually haphazard. Technology has changed the way we travel, interact with each other, and changed our relationship with the planet. Technology has enabled the formation of the modern state, and technology has created existential risk. The development and diffusion of new technologies is arguably the most impactful thing we do, and yet the process appears to be less tightly managed than, say, banking, sports competitions, or the selling of alcohol. Regulation, if it comes, usually applies to the sale or use of new technologies; in comparison, the creation and diffusion of new technologies, especially if done within science, is more of a Wild West.

The development of artificial intelligence (AI) is no different. The ultimate goal of many AI scientists is to create machines that are generally more intelligent than humans. Even before reaching this Holy Grail, today's AI systems are already capable of a wide range of applications: surveillance, drug discovery, internet search, content recommendation, and so on. These applications are built using methods developed by AI scientists, and sometimes even using specific AI systems that AI scientists make freely available. AI scientists train systems that can be adapted to a wide range of tasks, and then upload those systems to the internet; these systems can then be customised and integrated into software systems all over the world. AI is an unusual scientific field: not only do AI scientists build the artefacts that they study (rather than finding them in nature), but those artefacts are machine-like things that can be repurposed and deployed by onlookers.

"How can we build better AI systems?" This is the overarching research question that guides the AI scientific field. Research projects commonly take the following form. The scientists

train a deep learning model using familiar techniques, but making some change (e.g. to the training process, or the architecture of the model). They evaluate how well this model performs on a set of standard tasks (e.g. reading comprehension or image recognition), comparing its performance against similar models. This is an experiment, and the scientists write it up in a paper. The paper describes the innovation, gives some justification for why it works well, and states the results. The scientist also has the code that they used to run the experiment, including the code that trains the model. The code is, at least from the scientist's perspective, analogous to a biologist's lab notebook, and sharing this code allows the experiment to be reproducible. Moreover, the best description of the model is the model itself, i.e. the code and all the numerical parameter values, of which there may be billions, which comprise the trained AI system. Sharing the model allows other scientists to verify that the model does indeed perform as described. All this code is uploaded onto sites like GitHub, where it can be freely downloaded. In this way, many of the world's most advanced AI systems are created and shared as byproducts of the scientific process. These AI systems emerge and travel the world with a kind of "diplomatic immunity", initially treated as scientific knowledge objects rather than hazardous machines.

If the status quo prevails, when humans build machines that are more generally intelligent than us, the institutional passage through which these machines will be shepherded — during their creation and dissemination — will be a scientific one. The machines will be disseminated in a manner not dissimilar to how a scientist would share, for example, the description of an optical experiment, or photos of a new species of beetle. As they leave the lab, the machines will be judged against scientific standards, such as whether they really perform as well as the authors claim, or whether the authors have adequately cited the relevant literature. The machines will be easy to download, copy, share, edit, and use — so that other scientists can judge and build upon the work. This will quickly lead to widespread

proliferation. Where states step in and regulate, it will be to regulate how companies use AI systems in the services they provide (as the EU Commission has recently proposed),[1] which is much more familiar terrain than the regulation of scientific activity or the diffusion of scientific knowledge. Under this model, the transition into a post-AI world will not be a managed one. Humanity will stumble into a world filled with human-level or superhuman-level AI systems, without having conducted any process for collective decision-making around what AI should be developed and what should go undeveloped, how it should be designed, and who should have access to it and for what purpose. Later, state regulation will step in, but only once all the most important questions have been answered.

1. Governing scientific research outputs

This dissertation follows recent attempts to impose order on the dissemination of AI research artefacts. A number of institutional entrepreneurs within the AI research community have attempted to integrate their concerns about AI risks into the way AI research outputs are shared. This has involved departing from the standard paradigm in which AI researchers are expected to promptly open source the software created during a published research project. There have been some minor successes in reconfiguring the way AI research is shared, although on the whole, progress has been slow. Those concerned with the risks of widespread AI proliferation have been met with resistance, both from the AI research community and, on a deeper level, from the institutions they seek to reform.

On a general level, I am trying to understand whether science — as a place where new, potentially dangerous technologies sometimes get developed — can be governed in the

[1] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final). 2021.

interests of society. This question, of the governability of science, has already been studied in various forms. Some scholars have studied how public opinion can be afforded a greater role in the setting of scientific priorities.[2] Others have focussed on the epistemic problem of anticipating the impacts of new technologies before they become irreversible.[3] Other scholars have studied how scientists use rhetoric to insulate themselves against government intervention.[4] Scholars have also looked at attempts to integrate into major scientific projects the expertise of social scientists, ethicists, and lawyers — can scientific development be guided, from the outset, by these voices?[5] My dissertation contributes to this general body of literature on the governability of science.

More specifically, I am investigating a particular type of governance: structuring who gets what access to the knowledge and artefacts created during the research process. Here, I rely on Stephen Hilgartner's concept of a "knowledge-control regime".[6] I explain this concept below, but for now, a knowledge-control regime is a governance regime that allocates rights, obligations, and restrictions over knowledge — for example, military classification or patent protection. The central knowledge-control regime operative within AI research is the conference publication regime, through which AI researchers publish papers. This goes hand-in-hand with the second-most central knowledge-control regime in AI research, which is that

[2] Alan Irwin, 'Constructing the Scientific Citizen: Science and Democracy in the Biosciences', *Public Understanding of Science* 10, no. 1 (1 January 2001): 1–18, https://doi.org/10.3109/a036852.

[3] David H Guston, 'Understanding "Anticipatory Governance"', *Social Studies of Science* 44, no. 2 (15 November 2013): 218–42, https://doi.org/10.1177/0306312713508669; David Collingridge, *The Social Control of Technology* (London: Frances Pinter, 1980).

[4] Thomas F. Gieryn, 'Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists', *American Sociological Review* 48, no. 6 (1983): 781–95, https://doi.org/10.2307/2095325; J. Benjamin Hurlbut, 'Remembering the Future: Science, Law, and the Legacy of Asilomar', in *Dreamscapes of Modernity*, ed. Sheila Jasanoff and Sang-Hyun Kim (University of Chicago Press, 2015), 126–51, https://doi.org/10.7208/chicago/9780226276663.003.0006.

[5] Stephen Hilgartner, Barbara Prainsack, and J. Benjamin Hurlbut, 'Ethics as Governance in Genomics and Beyond', in *The Handbook of Science and Technology Studies*, ed. Ulrike Felt et al., Fourth, Ebook Central (Cambridge, Massachusetts: The MIT Press, 2017), 823–51.

[6] Stephen Hilgartner, *Reordering Life : Knowledge and Control in the Genomics Revolution*, Inside Technology (Cambridge, Massachusetts: The MIT Press, 2017).

researchers are expected to open source the AI software that they write about in papers. These regimes are designed such that scientists can effectively evaluate and build upon each other's work. My dissertation addresses the question: **how have these scientific knowledge-control regimes been adapted to address AI risks**? In other words, how does risk become a *feature of* the knowledge-control regimes that govern AI research?

Stephen Hilgartner's 2017 book, *Reordering Life*, studies knowledge-control regimes in the Human Genome Project. The study is about the relationship between knowledge-control regimes and transformative scientific change. Hilgartner studies how knowledge-control regimes, such as libraries of genetic data, coevolve alongside changes in the underlying science, such as new genome mapping technologies. My research is a simple extension to this work, in that I have added an extra ingredient: risk. Many people, including many AI scientists, are concerned about various risks of AI, ranging from concrete harms taking place today, to risks of future, more powerful systems (see below). I study the interaction between AI risks and knowledge-control regimes. As with Hilgartner's study of the Human Genome Project, I am simultaneously attentive to how changes in the underlying science can also have an impact upon knowledge-control regimes.

The question is **how AI risks are integrated into the knowledge-control regimes** ("KCRs") that govern AI research. A range of lower-level questions fall underneath. What do risk-focussed KCRs look like? How do these KCRs distribute responsibilities for tackling risks across different actors? How do they rewrite the social contract between the scientific community and broader society? Are some scientific knowledge objects easier to control than others — and so how might changes in the underlying science produce opportunities for governance? Is it possible to build risk-focussed KCRs on top of the field's existing (academic) regime for sharing research outputs, or is it impossible to teach an old dog new

tricks? How does the anarchic structure of the scientific field limit the kinds of changes that can be effectively made to KCRs?

To further clarify the focus of the thesis, the table lays out different possible research questions and whether or not they are the focus:

Question	**Focus of this thesis?**
***Why** govern?* I.e. What is the best rationale for governing the outputs of AI research?	No. In the case study, the motivation for governance is the societal risks posed by AI (described below). I do not focus on substantiating these risks.
***Who** should govern?* I.e. which actors should be empowered to govern the outputs of AI research?	No. The case study focuses on governance coming from *within* the AI research community rather than government regulation (see below). But I do not study the normative question of which actors would be most appropriate to govern AI research.
***How** does governance take place?* I.e. What form do the various governance attempts take, and how do they evolve?	**Yes**. This is the main focus of the thesis, with an emphasis on the interplay between risk and knowledge-control regimes.
***Can** AI research be governed?* I.e. is the field of AI research, in its current form, capable of being governed?	Secondary focus. A recurring theme is the difficulty of governing how AI researchers share their work.

The thesis is mainly *descriptive*, in that I describe new, risk-focussed KCRs and how they function. The thesis is also partly *explanatory*, in that I offer some explanation for the emergence of these new KCRs – beyond the obvious explanation of "researchers and labs were simply responding appropriately to the real risks of AI". For example, I argue that the existing institutional landscape – including the existing KCRs – constrains the development of new, risk-focussed KCRs (see below, section 5). The thesis is not seeking to be normative, although in the concluding chapter I do briefly offer some reflections on how AI should be governed given the persistent difficulties with addressing AI risks via the channels attempted in the case study.

2. Knowledge-control regimes in AI research

Hilgartner defines a knowledge-control regime as "a sociotechnical arrangement that constitutes categories of agents, spaces, objects, and relationships among them in a manner that allocates entitlements and burdens pertaining to knowledge."[7] The Creative Commons licence is an example, as are therapist-patient confidentiality rules. Knowledge-control regimes can be created by legal rules, but other substrates are available, such as informal norms and practices. Some knowledge-control regimes restrict the flow of knowledge, while others aim to spread knowledge more widely.

A knowledge-control regime positions different agents, such as the author of a conference paper or the owner of a patent. The regime structures the relationship between agents analogous to a contractual agreement. Hilgartner borrows from the legal philosopher Wesley Hohfeld in unpacking these different possible relationships.[8] For example, one scientist might

[7] Hilgartner, 9.

[8] Wesley Newcomb Hohfeld, *Fundamental Legal Conceptions*, ed. Walter Wheeler Cook (New Haven: Yale University Press, 1919).

have a *right* to access the data behind a published experiment, and this entails that the authors have a *duty* to share the data upon request. That is one category of relationship: rights and corresponding duties. Another category is where one agent is free to take an action (which Hohfeld calls a "privilege"), e.g. he is free to pass on information to trusted confidantes; and this entails that other agents have no right to prevent the behaviour.[9] These relationships parcel out control over knowledge, delineating who can do what with certain knowledge objects, such as accessing, editing, using, and sharing the knowledge. Sometimes, this has the effect of creating some kind of ownership over property — for example, patent ownership, or the ownership of digital art via non-fungible tokens.

In AI research, the central knowledge-control regime is the system for publishing papers at conferences, which I will call the "conference publication regime". There are certain high-prestige AI conferences that are international in nature and cover AI or machine learning very broadly: NeurIPS, ICML, and ICLR. There are also conferences known for particular areas within AI, such as CVPR for computer vision, and NAACL for natural language processing. Authors submit short papers, typically around 8 pages. Papers usually have several authors; often, the first author will be a junior researcher who did most of the painstaking labour, and the last author will be a senior researcher who supervised the project. The authors can come from anywhere in the world, and they do not have to be based at a university. The papers are peer reviewed on the basis of scientific merit. The accepted papers are published on the conference's web page. The authors can be invited to present their work at a poster session or a talk at the conference. Authors also often upload their papers to the preprint site arXiv, in tandem with going through the conference publication regime, and this is generally tolerated by the conferences.

[9] Hohfeld's framework also contains second-order relationships: an agent may (or may not) have the power to make changes to the existing setup of rights and privileges.

Under the conference publication regime, nobody owns the knowledge contained in the papers — it "belongs to the literature". The authors have no right to step in and prevent other researchers from using the ideas in later work; they can only demand that their contribution be acknowledged through a citation. If the authors want control over how their work is applied in the world, they have no recourse through the conference publication regime. They would have to turn to other regimes, where possible, such as patent protection. Patent protection is compatible with the conference regime, in that it is possible to publish and then apply for a patent on the published technique, although this is not a widespread practice (the industry labs sometimes do it).[10]

The conference publication regime plays a foundational role in structuring AI research. First, it provides strong incentives for researchers to be productive. Publishing at a major conference gives a researcher status, a sense of achievement, and it looks good on the CV. Job advertisements for AI researcher roles will often prefer that the applicant has, for example, 'First-author publications at peer-reviewed AI conferences (e.g. NeurIPS, CVPR, ICML, ICLR, ICCV, and ACL)'.[11] Second, and relatedly, somebody's existing publications provides a way of assessing their expertise or progress in an area. Internally, corporate research labs will sometimes look at publications as a way of assessing career progression for promotion, and the same thing happens at universities. Third, the conference publication regime helps the researcher to organise their workflow. A researcher's calendar will have the submission deadlines for their favourite conferences; they can spend a few months working on a paper, submit it to a conference, and then move onto the next. Finally, conference papers

[10] Nathan Calvin and Jade Leung, 'Who Owns Artificial Intelligence? A Preliminary Analysis of Corporate Intellectual Property Strategies and Why They Matter' (Working paper, February 2020), https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-working-paper-Who-owns-AI-Apr2020.pdf.

[11] This example is from a job advert for a Research Scientist at Facebook's AI research team (accessed Jan 2022).

allow for collaboration between researchers across time and space. Papers are narrowly focussed and build upon one another. Polanyi compared the scientific process to different people collectively building a jigsaw puzzle.[12] The conference publication regime (and the broader culture of publishing papers) hosts this large-scale collaboration. Overall, this regime is the most important institutional structure in AI research. It allows for a kind of structured anarchy, comparable to (as Polanyi pointed out) the market as a form of social organisation. In just a single institutional framework, AI researches are given incentives to produce research, a way of assessing each other's expertise and status, a way to structure the working year, a means of collaborating across time and space, and even, in many cases, a social life — after all, the conferences are great opportunity to hang out with people you haven't seen in a while. It is little wonder, then, that AI researchers appear to be obsessed with papers.[13]

Alongside conference publication, there is an informal regime where researchers upload their code and models to online repositories like GitHub. This dovetails with the conference publication regime: authors will often include links to such repositories at the bottom of their papers, and this is viewed favourably by reviewers. The NeurIPS conference has a code submission policy, where they ask that, where relevant, authors submit the code that they used to train and evaluate their models.[14] Sometimes, it will be easy to take the code and use it to train a model that is very similar to that featured in the paper. However, sometimes – even if the code is very comprehensive – running this computation will take many weeks and

[12] Michael Polanyi, 'The Republic of Science', *Minerva* 1, no. 1 (1 September 1962): 54–73, https://doi.org/10.1007/BF01101453.

[13] On Twitter, common topics of conversations are, for example: (a) "I am excited to share my latest paper…"; (b) complaints about shoddy peer reviewing; and (c) solidarity with other researchers in the face of an upcoming submission deadline. Also, if you ask an AI researcher, "how do you keep up to date with what's going on in the community?" they will often assume you are asking: "how do you keep up to date on reading the latest literature?"

[14] The NeurIPS 2021 code submission policy reads: "If any of the main contributions of your accepted paper depends on an experimental result, it's best practice for responsible research to include code that produces this result." See: 'NeurIPS 2021 Code and Data Submission Guidelines' (NeurIPS, 2021), https://neurips.cc/Conferences/2021/PaperInformation/CodeSubmissionPolicy.

thousands of dollars or more. This is becoming increasingly expensive with the trend towards training large models. In such cases, researchers are generally expected to open source the model. Practically speaking, this means uploading a file containing code that defines the structure of the model alongside a very long list of numbers. Those numbers are called the "parameters" or "weights" of the model, and they specify the strength of the connections between different neurons inside the neural network. These weights are discovered during the training process.

Even though researchers use GitHub to share their models, they do not build these models as open source projects, with online developers chipping in along the way. Rather, the researchers upload their finished project to GitHub.[15] If researchers directly collaborate with individuals from different labs, this is an academic collaboration, as coauthors on a paper. The project's GitHub page will refer to the paper – for example, linking to a preprint listed on arXiv, listing the authors, and showing how to cite the paper. In other words, scientific publication is still the dominant KCR.

Finally, models are usually trained using publicly available datasets. However, sometimes researchers will create their own datasets; if so, they are generally expected to share this too. As with code and models, this expectation is not always strictly adhered to; for example, Google and Facebook have certain in-house datasets that their researchers use without sharing.[16]

[15] This reflects the 'cathedral' rather than the 'bazaar' model of development. See Eric S Raymond, *The Cathedral and the Bazaar : Musings on Linux and Open Source by an Accidental Revolutionary*, First edition (Beijing; Cambridge, Mass.: O'Reilly, 1999).

[16] For example, Google has the JFT-300M internal dataset of labelled images.

I have briefly sketched out the existing knowledge-control regimes within AI research. Risks to society are not given any specific role in the setup.

3. Risks

In this thesis, "risk" refers broadly to a possibility of future harm. The sociological literature on risk sometimes takes a narrower concept: using 'risk' to refer to a particular cultural approach towards assessing potential harm, emphasising harms that are quantifiable or lend themselves to technical assessment.[17] I use the broader concept in order to capture the diversity in how people in my case study think about risks from AI, which are often difficult to quantify or technically assess, especially when they relate to risks of future AI systems. Nonetheless, the thesis still retains a sociological approach to studying risk in one key respect: risks are treated as partly social phenomena. In other words, risks are subject to debate and contestation between actors, they motivate changes to knowledge-control regimes, and they can be reflected in the resultant new knowledge-control regimes. It is this social process that is of primary interest; and so I do not seek to give my own position on what the risks of AI look like.

The conversation about AI and risk is very diverse, with many different kinds of risks discussed. One way of carving up the space of risks is to look at potential *applications* of AI that could lead to harmful outcomes:

[17] Langdon Winner, *The Whale and the Reactor: A Search for Limits in an Age of High Technology* (Chicago, UNITED STATES: University of Chicago Press, 1986), http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=557593; Hurlbut, 'Remembering the Future'.

- **Weapons.** AI has many potential military applications, including missile targeting, drones, and battlefield analysis and decision support.[18] The concern is that AI could increase the destructive potential of military force, or remove elements of human judgement (and restraint) from the use of force. There are also indications that AI could be used for the creation of new chemical weapons.[19]
- **Surveillance and political oppression.** AI is applicable to various forms of surveillance: identifying individuals (e.g. facial recognition) and making sense of what is happening in a stream of data (e.g. a video).[20] AI for processing text and image data is also applicable to censorship. Therefore, there are concerns that AI systems could strengthen authoritarian regimes in their political oppression; and that biases in AI systems will lead to discriminatory policing.
- **Misinformation.** AI could be used to create "bots" that spread misinformation online.[21] This is because today's AI systems are becoming capable of generating human-like text.[22] There are also concerns about AI-generated media such as videos and images, which could also be used to spread misinformation.
- **Automated decision-making.** AI can be used to automate decisions on behalf of companies and government authorities.[23] AI can introduce defects into these decisions, including gender or racial biases.

---

[18] Kai-Fu Lee, 'The Third Revolution in Warfare', The Atlantic, 11 September 2021, https://www.theatlantic.com/technology/archive/2021/09/i-weapons-are-third-revolution-warfare/620013/.

[19] Fabio Urbina et al., 'Dual Use of Artificial-Intelligence-Powered Drug Discovery', *Nature Machine Intelligence* 4, no. 3 (March 2022): 189–91, https://doi.org/10.1038/s42256-022-00465-9.

[20] Toby Shevlane and Allan Dafoe, 'AI as a Social Control Technology' (Unpublished manuscript, 2021).

[21] Samuel C. Woolley and Philip N. Howard, eds., *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*, Oxford Studies in Digital Politics (New York: Oxford University Press, 2018), https://doi.org/10.1093/oso/9780190931407.001.0001.

[22] Alec Radford et al., 'Better Language Models and Their Implications', *OpenAI Blog* (blog), 24 February 2019, https://openai.com/blog/better-language-models/.

[23] Karen Yeung, 'Algorithmic Regulation: A Critical Interrogation.', *Regulation & Governance* 12, no. 4 (December 2018): 505–23.

- **Labour displacement.** A wide range of potential AI applications involve the substitution of human labour, including (for example) truck driving. One risk is if such automation leads to unemployment and economic inequality.[24]

Another approach to carving up different AI risks is to look at the mechanism leading up to the harm. One helpful framework separates between accidents, deliberate misuse, and "structural risks":[25]

- **Accidents.** These are harms not intended by the people using the AI system. This would include cases where the AI system is biassed (e.g. on racial or gender grounds) and so produces unfair outcomes when deployed. It also includes the concern that highly capable, future AI systems will pursue goals not intended by the developers (see below).
- **Misuse.** These are harms intended by the users, e.g. cases where AI systems are used for criminal activity, terrorism, or disinformation. The category of "misuse" will, for many people, include states using AI for unethical forms of surveillance and for warfare; although there is no clear consensus on how to draw the line between use and misuse in these cases (which involves political or moral judgements).
- **Structural.** The structural perspective is distinctive from the other two in that it does not focus on the immediate harms from an AI system misfiring or being misused, but instead looks at the (often more long-run) impact of AI on social dynamics.[26] This

[24] Morgan R. Frank et al., 'Toward Understanding the Impact of Artificial Intelligence on Labor', *Proceedings of the National Academy of Sciences* 116, no. 14 (2 April 2019): 6531–39, https://doi.org/10.1073/pnas.1900949116.

[25] Remco Zwetsloot and Allan Dafoe, 'Thinking About Risks From AI: Accidents, Misuse and Structure', Lawfare, 11 February 2019, https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.

[26] Zwetsloot and Dafoe; Paul Christiano, 'What Failure Looks Like', AI Alignment Forum, 17 March 2019, https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

would include possible impacts of AI on: employment, nuclear stability, and the resilience of authoritarian regimes.[27] This category also includes the concern that humans gradually lose control over the future by delegating more and more decision-making to AI systems.[28]

Finally, another way of mapping the space of risks is to look at harms *already* occurring today and harms which would require stronger AI capabilities than we have today. A much-discussed cluster of risks in the latter category would be catastrophic or existential risks from AGI (artificial *general* intelligence: an AI system that has the wide-ranging cognitive capabilities of a human, or greater). The Oxford philosopher Nick Bostrom argued in his 2014 book, *Superintelligence*, that AI systems that exceed the cognitive capabilities of humans would pose an existential risk to humanity, either through seizing control of the world or through wielding weapons of mass destruction.[29] The idea that AI could in future pose catastrophic or existential risks to humanity has some support among AI researchers: in a 2022 survey, 36% of the AI researchers surveyed agreed that it was plausible that AI could cause a catastrophe on the level of 'all-out nuclear war'.[30]

4. The case study

My case study goes from 2019 to 2021. Fortuitously, 2019 was both when I began collecting empirical data but also the natural starting-point to the story: I am not aware of any previous

[27] Ben Garfinkel, 'Is Democracy a Fad?', *The Best That Can Happen* (blog), 26 February 2021, https://benmgarfinkel.wordpress.com/2021/02/26/is-democracy-a-fad/; Keir A. Lieber and Daryl G. Press, 'The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence', *International Security* 41, no. 4 (1 April 2017): 9–49, https://doi.org/10.1162/ISEC_a_00273.

[28] See Stuart Russell, *Human Compatible: AI and the Problem of Control*, 1st edition (London: Allen Lane, 2019).

[29] Nick Bostrom, *Superintelligence : Paths, Dangers, Strategies* (Oxford, UK: Oxford University Press, 2014); See also Russell, *Human Compatible*.

[30] Julian Michael et al., 'What Do NLP Researchers Believe? Results of the NLP Community Metasurvey' (arXiv, 26 August 2022), https://doi.org/10.48550/arXiv.2208.12852.

attempts to reconfigure AI research KCRs around risk. My empirical material can be broken down into three episodes, i.e. three attempts to change how AI research outputs are shared. My thesis has a chapter devoted to each.

First, there was the "staged release" of GPT-2, announced in February 2019. GPT-2 is the name of a large, pretrained language model.[31] This means the model is fed a very large quantity of text data (scraped from the internet), but with certain words hidden, and the training task is to predict the missing words. In order to succeed in this task, the model must learn generally applicable skills and knowledge: initially, this involves basic things like knowledge of syntax, but to squeeze out further performance, the model must learn something about how the world works, and about the underlying concepts expressed in language. After that pretraining process, the model can then be adapted to perform a range of "downstream" tasks, such as answering reading comprehension questions or classifying text into categories. GPT-2 was also well-suited to generating text, because, given an initial chunk of text, the user could repeatedly sample from the model's predictions about what the next word would be.

GPT-2 was surprisingly capable of generating coherent stories and articles. Members of the research lab that made the model, OpenAI, were concerned that it could be misused to flood the internet with fake text. They decided to only open source a small version of the model, which was less capable. The plan was to gradually release larger versions of the model, while watching to see if people were misusing the models that had been released. This was partially successful, and a few other research groups followed OpenAI's lead, withholding similar

[31] Alec Radford et al., 'Language Models Are Unsupervised Multitask Learners', *OpenAI Blog*, 2019, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

models. However, at the same time, the effort was undermined by a succession of GPT-2-like models that other research groups released.

The second episode concerns GPT-3, the successor to GPT-2, which was much larger and more powerful. The GPT-3 paper came out in May 2020. By this time, OpenAI had transitioned from being a non-profit to a for-profit company. GPT-3 became the company's first product: users could interact with the model through a web interface or an application programming interface (API). This was an attempt to make money from the model, but simultaneously an opportunity to experiment with a new governance regime. OpenAI set rules on how the model could be used. For example, GPT-3 cannot be used for targeted political messaging. OpenAI set up a system for monitoring how people are using the model, allowing the company to step in and prevent misuse. This was not common for commercial APIs in AI, and not possible under the standard scientific practice for sharing models, i.e. uploading open source models. Through this new regime, the lab took on responsibility for shaping how GPT-3 would be used in the world. AI researchers are ordinarily quite ignorant of how their work gets used in the real world, often viewing this as falling outside of their purview. For technical reasons, GPT-3 was harder to replicate than GPT-2, and at the time of writing there exists no similar open source model. A few other companies do offer API access to similar models, though likely without the same level of resources devoted to monitoring for misuse.

Finally, I look at changes to the way conference papers are published. In 2020, NeurIPS — the most prestigious deep learning conference — began reviewing papers on ethical grounds. They also required that authors submit a "broader impact" statement at the end of their papers, discussing the possible societal consequences of the work. For a conference of nearly 2000 accepted papers, 13 were flagged for ethical review, and four of those were rejected on ethical

grounds. As with the staged release of GPT-2, the initiative was met with some protestation from within the research community, especially the broader impact statement requirement. Researchers argued that the vast majority of papers had no ethical or societal significance – or at least, nothing that was at all unique to specific papers. The broader impact statements were generally low quality, and the 2021 conference organisers watered down the requirement.

Throughout the project, my analysis is focussed on the level of the *field*.[32] Although two of my substantive chapters begin with the actions of a single AI lab (OpenAI), my project is not an organisational analysis, and I do not devote much attention to the internal decision-making or politics of that lab. Not only did I lack unfettered access to the within-organisational dynamics, but more importantly, a single lab cannot make important changes to the field's KCRs by acting alone. In both those chapters, the most interesting thing was how the experimental KCRs attempted to reorganise, and depended upon, the wider research field.

It is difficult to define, in a clean way, the boundaries of the field that I am studying. The conference publication regime helps with this task: one starting point would be 'the English speaking researchers who regularly publish papers at the NeurIPS conference'. However, in many cases this definition will be artificially narrow, leaving out, for example: (a) researchers who do similar work but prefer to publish at other conferences or do not speak English; and (b) people working in non-research roles within important AI labs, and cloud computing

[32] See Paul J. DiMaggio and Walter W. Powell, 'The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields', *American Sociological Review* 48, no. 2 (1983): 147–60, https://doi.org/10.2307/2095101; Paul DiMaggio and Walter W Powell, *The New Institutionalism in Organizational Analysis* (Chicago: University of Chicago Press, 1991); Huseyin Leblebici et al., 'Institutional Change and the Transformation of Interorganizational Fields: An Organizational History of the US Radio Broadcasting Industry', *Administrative Science Quarterly*, 1991, 333–63; W. Richard Scott, *Institutions and Organizations*, vol. 2 (Sage Thousand Oaks, CA, 1995); Neil Fligstein and Doug McAdam, *A Theory of Fields* (New York: Oxford University Press, 2012). Fligstein and McAdam define 'strategic action fields' as 'mesolevel social orders, as the basic structural building block of modern political/organizational life in the economy, civil society, and the state'.

providers, who help to shape the events in my case study. Also, the boundaries of the relevant field are partly endogenous to the events I am studying – see, for example, my discussion in chapter 2 about the range of actors that have the ability to replicate GPT-2. I prefer to expand my view of the field to accommodate these various actors, even if practically speaking it is difficult for me to fully capture all the different actors within the field.

The research field has a mix of university-based academic researchers and industry labs. While top conferences still have a higher number of papers from academia, some of the biggest contributors come from industry. Google AI Research, for example, publishes more papers at top AI conferences than any university group, with 178 papers accepted to NeurIPS 2020 (Microsoft was also at 95 – slightly less than Stanford and MIT).[33] Industry labs also train an especially high proportion of the largest models (i.e. models with a high number of parameters), which requires greater computational resources and engineering effort than is normally within reach for academic researchers.[34] These industry labs have been surprisingly academic in their culture and behaviour, with researchers focussing heavily on contributing to the scientific literature, often free from the pressure to work on commercial products (although this varies from place to place). I do not have a strong answer for what motivates the large technology companies to house and fund these labs, but some possibilities are: (1) the labs offer prestige and marketing for the companies as leaders in AI; (2) it is helpful to have leading in-house AI experts who can be called upon for more product-led AI projects; (3) sometimes the models trained by researchers will be used in production; (4) the companies are waiting for a later moment when AI becomes more commercially valuable, at which point they will move to a less academic approach; and/or (5) the leaders of the technology

[33] See Sergei Ivanov, 'NeurIPS 2020. Comprehensive Analysis of Authors, Organizations, and Countries.', *Criteo R&D Blog* (blog), 15 October 2020, https://medium.com/criteo-engineering/neurips-2020-comprehensive-analysis-of-authors-organizations-and-countries-a1b55a08132e.

[34] Deep Ganguli et al., 'Predictability and Surprise in Large Generative Models', in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, 1747–64, https://doi.org/10.1145/3531146.3533229.

companies are genuinely interested in advancing AI. It is therefore difficult to draw neat distinctions between different categories of actor, although I do highlight where there are important differences between industry and academic labs, especially where these differences are relevant to governance.[35]

In building the case study, my primary data source was interviews. I interviewed both people designing and running the new KCRs (e.g. members of OpenAI), and with the AI researchers who reacted to, and interacted with, these new KCRs. I conducted 42 interviews with 37 different people between 2019 and 2021, most of which lasted around an hour. My interviewees have been a mix of university-based researchers and researchers based at industry AI labs. My interviewees are not a random sample of the population of AI researchers. Some of them I met organically at conferences: in 2019 I attended ICML, ICLR and NeurIPS, and in 2020, I attended AAAI. Often I would go through conference proceedings (again, NeurIPS, ICLR and ICML) and take authors' email addresses from their papers, sending out bulk emails. The vast majority did not get back to me; those that did were probably disproportionately interested in AI risk or issues around publication. Indeed, a few researchers have declined my interview requests by saying that they do not think about my topics at all and just prefer to get on with their research. I did not take everything my interviewees told me at face value. I felt a number of them probably saw me as an outsider, almost like a journalist, and therefore naturally adopted a defensive position (which normally meant defending the existing publication regime). Nonetheless, this still made for useful data. I have given most of my interviewees pseudonyms, even though many of them consented to be named. I have tried to find gender-neutral pseudonyms to help protect the identities of those who wanted to remain anonymous.

[35] In chapter 1, I highlight the bureaucratic environment of industry labs as conductive to more top-down rules on what artefacts should not be made openly available. In chapter 2 I highlight how the large compute resources of certain industry labs affords them a kind of governance power.

I also collected thousands of tweets during 2019-2021. I tried to follow as much of the Twitter-active, English-speaking AI research community as possible; in 2019, I made a social network graph of this community, which allowed me to find members who I was not following (i.e. AI researchers who were followed by many other AI researchers). With both my interview and Twitter data, I have targeted AI researchers, i.e. members of the technical, scientific community developing and studying novel forms of AI, especially deep learning. This community is split between universities and industry labs (above), which is reflected in my sample. I have *not* been studying: software developers who use AI inside their companies; humanities and social science academics who work on AI; or policy-makers interested in AI. I have interviewed people with social science or humanities backgrounds who are now closely integrated into the AI research community — for example, members of the OpenAI Policy team, or staff at the Partnership on AI (a non-profit that features in my case study).

I have focussed on the scientific community because this is where the KCRs governing scientific research outputs are made and applied. Also, most of the world's novel, advanced AI techniques and systems come from this community, and so it is an important source of AI risk. An alternative focus for my thesis could have been policymakers grappling with AI: for example, those contributing to the EU AI Act. There are three reasons why I did not cover this. (1) Considerations of focus and time: focussing specifically on the AI research community has allowed me to go into greater depth on that world. (2) During the time period I studied (2019-2021), policymakers were not openly contemplating controls on the proliferation of AI. For example, the EU AI Act, in its current form, focuses primarily on the provision of AI services, rather than the underlying proliferation of AI capabilities. My research focuses on this more *upstream* issue of proliferation, which is still an important site for governance, not least because the nature and magnitude of AI proliferation sets the stage

for the downstream use of AI (and how easily those uses can be governed). (3) Relatedly, it comes down to an interesting question of power and influence. The AI research community has the de facto power to decide how much AI proliferation occurs, because it is they who produce the insights and artefacts, and there do not exist go-to policy levers for governments to modulate this. During the "crypto wars", the US government struggled to effectively regulate (using export controls) the proliferation of cryptographic research from the 1970s onwards, and this became a freedom of speech issue. Government efforts to regulate the dissemination of AI research could be met with a similar backlash, or framed as an attack on scientific progress. Similarly, in one of my interviews with an employee of the Partnership of AI interested in the risks of AI proliferation, a major theme was their lack of power over what AI researchers choose to do, and their unwillingness to adopt a strong position before AI researchers had been "brought along". (As a different interviewee put it: "Leverage in this space is being attached to a big computer.")

As additional sources of data, I rely upon publicly available online sources, such as company blogs, personal blogs, reports, job adverts, research papers, API documentation, application forms for accessing models, discussions on forums like Reddit and Hacker News, podcast interviews, talks, and conference web pages.

With all my data sources, I have qualitatively coded the data into thematic categories using Roam Research. These categories are relevant across all three of the experimental KCRs that I examine in the case study. I developed the categories continuously throughout the research, going between my data and my research question, and adjusting the categories as I collected more data.[36] The high-level categories are now reflected in the six central themes that I

[36] See Barney G. Glaser and Anselm L. Strauss, *The Discovery of Grounded Theory : Strategies for Qualitative Research*, 1967; Matthew B Miles and A. M Huberman, *Qualitative Data Analysis : A Sourcebook of New Methods* (Beverly Hills: Sage, 1984).

introduce in section 4, below. For example, one category was ‘large models and the relationship between industry and academia’. I also had more specific categories that were useful for organising my data, such as ‘GPT-2 decision-making’ and ‘GPT-2 replications’.

I have tried to resist my project becoming a study in how AI researchers *talk about* knowledge-control regimes and AI risks, despite collecting a wealth of such data. My study is more institutionalist in nature: I use the qualitative data to gain a deeper understanding of the new KCRs and how they rubbed against the existing field.[37] This follows Hilgartner, whose study was likewise quietly informed by institutional theory. This is a natural position given that knowledge-control regimes are institutional in nature: they are persisting social structures that can be maintained or disrupted by social actors.[38] In fact, the case study could be alternatively framed as a study in institutional change, in that the relevant KCRs have undergone a process of being problematised and adapted in light of potential risks from AI. (Nonetheless, the aim of the thesis is not to contribute to the literature on institutional change, at least when pitched at such a high level of abstraction; see section 7 for the intended contributions.)

Despite studying three different episodes, I view the thesis as a single case study project. The different episodes overlap in terms of the people involved and the time period, and (at least with GPT-2 and GPT-3) lessons from one KCR endogenously inform the next. Nonetheless, I still draw comparisons between these three different KCR experiments. My aim with such comparison, which I will use frequently throughout the dissertation, is to add depth to my

[37] Similar to Hilgartner, I adopt new institutionalist theory as a background theoretical paradigm. This literature brings a focus on the ‘field’ as a level of analysis, and institutions and institutional change as key objects of interest. See DiMaggio and Powell, *The New Institutionalism in Organizational Analysis*.

[38] See Charlene Zietsma and Thomas B. Lawrence, ‘Institutional Work in the Transformation of an Organizational Field: The Interplay of Boundary Work and Practice Work’, *Administrative Science Quarterly* 55, no. 2 (1 June 2010): 189–221, https://doi.org/10.2189/asqu.2010.55.2.189.

description of the new KCRs, bringing out aspects that I might otherwise take for granted. Also, there are commonalities between the different KCRs (at least in the challenges they must address), and this gives me some added confidence in the robustness of my arguments, even though I do not claim that my findings will necessarily generalise to other fields of science and technology.

5. What is special about AI?

Focussing on AI as a single case study allows for deeper exploration, but comes at the cost of missing potentially interesting connections with historical cases. In this section, I briefly touch upon some relevant parallels to other fields of science and technology. This is not intended to substitute for the benefits of a multiple case study design, but to help situate AI as a case study.

AI is not the first time that KCRs within a scientific field have had to adapt to concerns about societal risks from the technology being produced. I would highlight a few other examples: nuclear science in the 1930s and 1940s; computer security (in particular, the disclosure of software vulnerabilities) in the late 1980s through to the 2000s; and gain-of-function virology research in the early 2010s. All these examples, alongside AI, involve risks from sharing research knowledge and artefacts, and "upstream" (i.e. addressed to the researchers) norms or interventions in an attempt to govern these risks. In comparing them to AI, we can look at both (a) the nature of the field and the actors involved, and (b) the nature of the technology and its risks.

The field and actors. All these examples have involved academic research communities, and often with private sector actors also doing research in the area. What stands out about AI is the absence of state involvement in publication decisions, which we can see in each of the other examples. For nuclear research, there was a fast transition, at least in the US, from an

academic field to total state involvement and control over the research via the Manhattan Project. The US Department of Defense thereby imposed a very strict regime of secrecy on nuclear research.[39]

The geopolitical context for such state intervention in nuclear research was World War II. In virology, the geopolitical context (in the 2000s and 2010s) was the post-9/11 "war on terror". In the US, a 2004 report from the National Academy of Sciences led to the creation of the National Science Advisory Board for Biosecurity (NSABB), a federal-level panel of experts reporting into the US government. (The 2004 report was titled: "Biotechnology Research in an Age of Terrorism".) The NSABB then played a central role in a 2011 controversy over a couple of papers describing how to increase the transmissibility (between mammals) of H5N1 influenza. The NSABB recommended against publication of these papers, and the US government supported this decision.[40] The issue was even discussed between states at the international level, at the WHO and the Australia Group.[41]

In the computer security case, too, the US government has played some role in governing how software vulnerabilities are published. In 1988, there was the spread of the first computer worm (the "Morris worm"), created by a graduate student, which caused much disruption on the early internet. When researchers at UC Berkeley recovered the code for the worm, the US Department of Defense reportedly encouraged them not to publish this code.[42] Then DARPA, a federal defence funding agency, funded the creation of a new centre for coordinating the defence against computer viruses.[43]

---

[39] Alex Wellerstein, *Restricted Data* (University of Chicago Press, 2021).

[40] Ian Sample, 'Bird Flu: How Two Mutant Strains Led to an International Controversy', *The Guardian*, 28 March 2012, sec. World news, https://www.theguardian.com/world/2012/mar/28/bird-flu-mutant-strains.

[41] Jade Leung, 'Who Will Govern Artificial Intelligence? Learning from the History of Strategic Politics in Emerging Technologies' (PhD Thesis, University of Oxford, 2019).

[42] Katie Hafner and John Markoff, *CYBERPUNK: Outlaws and Hackers on the Computer Frontier, Revised*, Updated edition (New York: Simon & Schuster, 1995).

[43] E. H. Spafford, 'Crisis and Aftermath', *Communications of the ACM* 32, no. 6 (1 June 1989): 678–87, https://doi.org/10.1145/63526.63527.

The geopolitical context for AI is currently the tensions between the US and a rising China. This has not yet led the US or other Western governments to intervene in the AI research field to regulate publication, although it is possible that this may occur in future (the AI case study is far from closed). With AI, economic and strategic competition between states has so far played out in different fora. We have seen national AI strategies; state sponsorship of AI research projects;[44] and there has been some discussion of export controls on AI technologies, which has not yet led to controls on AI research publications.[45] The EU's proposed AI Act does not focus on AI research (see chapter 2). All this means that the case study in this thesis is notably isolated from state involvement, relying much more heavily on private actors. With time, we could learn that this was a function of the case study covering the *early stages* of AI's rise as a risky (and strategically important) technology.

***The technology and its risks*.** In all the examples discussed, there has been some disagreement or uncertainty over the nature and magnitude of the risks, although perhaps none more so than AI.

With the nuclear example, some people had impressive foresight about the nature of the risk. In 1914, nearly three decades before the Manhattan project, H.G. Wells wrote *The World Set Free*, a science fiction novel in which atomic bombs are invented, leading to a devastating nuclear war. The physicist Leo Szilard read the book in 1932, a year before he had the idea for a neutron chain reaction. Szilard foresaw the destructive potential of his discovery, especially given the rise of Nazism in Germany, and encouraged other nuclear scientists against publishing on the chain reaction.[46] The main disagreement between scientists during

[44] For example, the French government has provided the computational resources for BigScience's effort to replicate GPT-3: see chapter 2.
[45] Jade Leung, Sophie-Charlotte Fischer, and Allan Dafoe, 'Export Controls in the Age of AI', War on the Rocks, 28 August 2019, https://warontherocks.com/2019/08/export-controls-in-the-age-of-ai/.
[46] Richard Rhodes, *The Making of the Atomic Bomb* (New York: Simon & Schuster, 1986), 24.

this period was about the feasibility of creating nuclear weapons, with some thinking it was impossible or many years away.

In the H5N1 controversy in 2011, the main disagreement was over the risks and benefits of sharing the research, rather than (for example) whether the virus was dangerous or possible to produce. Some scientists argued that it was important to publish the research so that researchers could better understand the virus and what mutations to watch for in the wild; and that this benefit outweighed the risk of bioterrorism. Likewise in computer security, the main focus of the debate was on the risks and benefits of publication.[47] Many computer security experts argued that open publication of software vulnerabilities incentivised software-makers to patch the software more quickly.[48]

AI is characterised by deep ambiguity about the nature of the technology and its risks. For example, there is variation across different expert communities in how we should think about AI. Some experts view AI as a (sometimes defective) tool that human decision-makers put too much faith in, leading to unfair and non-transparent decisions – and question whether AI is a form of "intelligence". Other experts view AI in terms of a set of capabilities that are increasing and becoming more general over time, trending towards human-level and then super-human level intelligence. Even among experts with the latter view, where many worry about existential risks to humanity's future,[49] there is a diversity of opinion over how these risks come about (what technical failures might occur; the role of human misuse; whether the risk would be sudden or gradual; and how close in time we are from dangerous AI being developed). These risks are often pinned to the future technology of "artificial general intelligence" (AGI), but it is difficult to define or precisely envision what AGI will look like.

[47] Sample, 'Bird Flu'.

[48] Bruce Schneier, 'Crypto-Gram: November 15, 2001', *Schneier on Security* (blog), 15 November 2001, https://www.schneier.com/crypto-gram/archives/2001/1115.html.

[49] See Toby Ord, *The Precipice : Existential Risk and the Future of Humanity* (London: Bloomsbury, 2020).

As we shall see in the case study, it is also not settled which kinds of AI insights or artefacts are most risky, whether from the perspective of AGI risks or present-day risks. Finally, AI is a very general-purpose technology, with a very wide range of potential applications. This adds to the difficulty of thinking about what the impact of AI will be in society.

Overall, then, AI is an interesting case study of an emerging technology, where the roles of different actors have not yet been settled – but where the state is not yet heavily involved – and where the risks of the technology are not settled either. This is the context for the development of new, risk-based KCRs, where private actors are forced to rely on their own powers of institutional entrepreneurship, in a context of deep ambiguity about what the relevant risks are and how they should best be tackled.

6. Introducing the central topics of analysis

Here, I introduce six central topics for the dissertation. The first five are all important dimensions along which the different KCRs will be analysed. The sixth topic, in contrast, is about understanding the forces that shape the KCR development process. Everything here will be further substantiated in the rest of the dissertation.

First, *what source of power does the new KCR rely upon*? For GPT-2 and GPT-3, an important factor was that, for a limited time period, OpenAI was the only actor that had those models. The company could therefore pick and choose what capabilities to make available and when. Replication by other research groups threatened to undermine the temporary control over proliferation that OpenAI had established, although this threat was moderated by the size of the models. Especially in the case of GPT-3, there was a limited number of research groups who had the ability to train such a large model, including paying the cost of

compute (estimated to be around $5 million).[50] The scientific trend towards "bigger is better" creates new opportunities for controlling AI research models.

In addition to establishing a bottleneck, the KCR governing GPT-3 also relies upon strong surveillance capabilities. In March 2021, GPT-3 was generating 4.5 billion words per day.[51] Effective monitoring of such activity can only be possible using natural language processing AI systems – it is likely that language models are being used to monitor the use of language models. The KCR therefore relies upon a technical (as well as social) machinery of governance, at a level that has only recently become technically possible. Finally, ethical review of conference publications relies upon an existing bottleneck in the conference publication regime, namely that papers must be approved by reviewers before being published by the conference.

Second, *how does the KCR construct a pipeline for expert knowledge about risk to feed into decision-making?* KCRs normally include decision-making processes – for example, peer reviewers assess a paper for scientific knowledge, or a patent office assesses the novelty of an invention. In my case study, decision-making processes are constructed within the new KCRs, and these serve as a means for controlling risk. What do these decision-making processes look like — who are the experts, and how are they built into the KCR?

In the case of ethical review of conference papers, if a paper is flagged by reviewers as raising ethical issues, dedicated "ethics reviewers" are called in to make a judgement. These reviewers are drawn from the broader fields of AI governance and AI ethics. In the case of the GPT-3 API, OpenAI's in-house policy team crafts the rules on what constitutes safe and unsafe uses of the model. Partly, this was informed by early, in-house experiments with GPT-

[50] Chuan Li, 'OpenAI's GPT-3 Language Model: A Technical Overview', Lambda Blog, 3 June 2020, https://lambdalabs.com/blog/demystifying-gpt-3/.
[51] OpenAI and Ashley Pilipiszyn, 'GPT-3 Powers the Next Generation of Apps', OpenAI Blog, 25 March 2021, https://openai.com/blog/gpt-3-apps/.

3: for example, can humans distinguish GPT-3-generated news articles from real ones? Then, on an ongoing basis, OpenAI's policy team sought to understand misuse of the model, both by (a) encouraging external researchers to do further studies on the capabilities and limitations of the model via the API, but also (b) monitoring how users are interacting with the model. They can ask, for example: are many users generating political content, and in what context? The API helps to produce knowledge about misuse, and this knowledge then feeds directly back into the rules governing use of the API, which the company updates over time. A lawyer might say that the API evolves as a "living tree"; in the same way, a cybernetician might say that the API sets up feedback loops between the governance of the model and the outside world.

As the case study developed, I realised that it was more about *production* of knowledge (about models and their risks) than it was about the *suppression* of knowledge. At the outset, I expected this to be a project about secrecy, and I read literature on, for example, the organisation of secrecy at Bletchley Park during World War II. To some extent, this was right: most obviously, the KCRs used for GPT-2 and GPT-3 both required that the weights of the models be withheld, and this could be considered a form of secrecy. And more generally, the industry labs regularly use non-disclosure agreements to protect the secrecy of ongoing research – even if the same research will often be published openly in papers soon after. Nonetheless, the bigger story was the endless drive to produce more knowledge. Actors were asking themselves questions like: "Can this model actually be misused?" "Instead of withholding the model, shouldn't we produce another model that counteracts the misuse?" "Is the model being misused in the real world?" And "How can we inform policymakers and users about the risks of today's AI systems?" The KCRs I have studied all seem more devoted to producing knowledge, in answer to such questions, than concealing knowledge. My explanation for this is one half cynical: secrecy is often not in the interests of researchers, and so it is systematically ignored as a tool for preventing the proliferation of dangerous

technology, with researchers preferring to do what they do best (knowledge production). Nevertheless, in equal measure, it is also a function of some inherent connection between knowledge and governance. Governing new technologies does require that decision-makers are informed about risks.

Third, *how does the KCR position the relationship between science and society?* As I have noted, the prevailing KCR landscape is built around the principle that scientists should be able to communicate with each other effectively. Forums for scientific publication, such as conference proceedings, were not originally built as channels for disseminating knowledge and technology to "the public" or governments and companies. Indeed, many early scientists were reluctant to have their work read by the masses.[52] Roger Bacon, in his Opus Majus, written in the 13th Century, wrote: "the secrets of the sciences are not written on the skins of goats and sheep so that they may be discovered by the multitude."[53] Even today, when I ask AI researchers why scientific openness is important, they usually foreground the within-science benefits, i.e. researchers can validate and build upon each other's work. Especially since the internet, it has become very easy for actors outside the scientific community to eavesdrop on this process, but AI researchers often seem to view this technological diffusion as a welcome bonus rather than the reason they share their work. "Society" is somewhere "out there" – an object that might get transformed by AI, or a jungle where the uses of AI will be thrashed out. It is not a force that actively shapes the way AI research is shared.

In the case study, this basic premise is challenged, and actors have sought ways of putting society in the driving seat. There are two main ways in which this has occurred. We have already seen the first: feedback loops. With GPT-2, OpenAI was watching how actors would

[52] William Eamon, 'From the Secrets of Nature to Public Knowledge: The Origins of the Concept of Openness in Science', *Minerva* 23, no. 3 (1 September 1985): 321–47, https://doi.org/10.1007/BF01096442.
[53] Roger Bacon and Robert Burke, *The Opus Majus of Roger Bacon* (New York: Russell & Russell, Inc., 1962), 11–12. Bacon attributes this view to Aristotle and Socrates.

use the smaller versions of the model, and using that to inform the release schedule for the larger, more capable versions. For example, they monitored "anonymous forums with a history of spreading disinformation and organising hate movements" to see if these people were planning on using GPT-2.[54] With GPT-3, as I have mentioned, the API setup allowed the company to monitor misuse of the model directly.

The other mechanism for bringing society into the KCR is "representational": representatives of society are built into the decision-making process. This is analogous how, in a representative democracy, the legislature is supposed to represent the population. Of course, the link between the representation and the represented is not always very strong. One example of the representational method is how the ethical reviewers for conference papers are selected; as the organisers for one conference put it, "we wanted to ensure that potential societal impacts of work published at NAACL were considered from multiple different cultural perspectives."[55] Also, a research team that replicated and open sourced a GPT-2-like model claimed that their decision to do so had drawn upon "emerging norms and governance processes that incorporate a broad set of values from across society"[56] (a tenuous claim). In other words, people, organisations, and institutional processes are constructed as representing some notion of society, and this is then built into the functioning of the KCR. This has occurred in parallel with a related drive (outside of my scope) to engineer AI models with one eye on society – for example, removing societal prejudice from datasets used to train the models, or the general effort to align AI systems with human values (a.k.a. alignment

[54] Irene Solaiman et al., 'Release Strategies and the Social Impacts of Language Models', *ArXiv:1908.09203 [Cs.CL]*, 2019, 5.
[55] Emily M. Bender and Karën Fort, 'NAACL Ethics Review Process Report-Back', NAACL 2021, 20 May 2021, https://iuliaturc-google.github.io/naacl-org/naacl-2021-website/blog/ethics-review-process-report-back/.
[56] Nitish Shirish Keskar et al., 'CTRL: A Conditional Transformer Language Model for Controllable Generation', *ArXiv:1909.05858 [Cs.CL]*, 2019, 12.

research). The difference is that, in my case study, it is KCRs that are putatively aligned with human values, rather than the models themselves.

As well as making the scientific process more responsive to society, there has also been some effort to make society more responsive to the science. In the case of the NeurIPS changes, the broader impact statements partly functioned to position scientists as public educators, the idea being that policy makers and civil society actors might read the statements and learn something about the technical realities of AI risk. That said, the requirement was also imagined as a way of forcing self-reflection upon AI researchers, the idea being that if they contemplated the social impact of their work, this might lead them to produce more socially responsible research.

Fourth, *how does the KCR position different actors inside the scientific community*? One of the things that upset many AI researchers about GPT-2 was that the research community had no special access to GPT-2 (there was an academic access programme but it was limited in scope and did not give full model access). Researchers got access to the models on the same schedule as everyone else. The KCR did not really honour the boundary between scientists and non-scientists. GPT-3 was similar: there is not one API for researchers and another for software developers, for example; although OpenAI appeared quite keen to grant scientists API access when they applied. GPT-3 and related developments have contributed to an anxiety among some university-based researchers that they are being left out of certain areas of research, partly due to the high computational costs of training large models, but partly because they do not perceive that these models are being shared in a fully "open" way.[57]

[57] In the Stanford HAI Virtual Workshop on Foundation Models (23-24 August, 2021), one of the leaders of the initiative said: "Sadly, it's becoming harder and harder for the community to engage. A big but often underappreciated factor driving the deep learning AI revolution is the incredibly open culture and open ecosystem. It's the norm that researchers release their datasets and code and tutorials so others in academia and industry can learn and build on top of them, and this has led to astonishing progress in the field. But with foundation models, this openness is being eroded, with datasets and code increasingly not being released. For example, while some models like GPT-2 and T5 are public, the code for training them is not. And with GPT-3,

Fifth, *what types of research outputs count as risky?* One risk philosophy that is common in my case study is a concern that dangerous AI systems will proliferate and be used in a careless or malicious way. This is the key focus of the KCRs used for GPT-2 and GPT-3, which both aim to keep the models out of the hands of "bad actors". The idea was that the models themselves were the hardest thing for bad actors to recreate, given the high compute costs required to produce the models. This was assuming that the papers, in contrast, did not contain truly novel insights. However, this focus on models has been questioned, with some risk-conscious AI researchers arguing that the GPT-3 *paper* was actually the risky thing. The paper, alongside other papers that OpenAI published in 2020, demonstrated to many onlookers the benefits of scale: if you throw a large amount of compute and data at a model with a very high number of parameters, you can get very impressive capabilities. Some people viewed this as dangerous in that it accelerates the field's progress towards advanced AI, thus giving the world less time to prepare. This concern was actually discussed before the publication of these papers within OpenAI, and they even delayed publication for a number of months. However, we will see that, in the AI research field, insights can be harder to govern than artefacts; and moreover, there are stronger incentives for researchers and labs to publish papers. In fact, a general finding of the case study might be the robustness of the research paper as an institution: the paper is still going strong, having successfully deflected or absorbed concerns around risk.

In the case of ethical review, the concern is not about proliferation of risky technology; often, the concern is that the paper contributes negatively to the discourse around AI – for example, authors claiming they have trained an AI system that predicts criminality from a photo of someone's face. The mechanism is to deny such papers the prestige of conference publication. Overall, I will argue that the three new KCRs each govern different objects: the

even though a few people could get API access, the models are not available." See: *Workshop on Foundation Models: Day 1* (Stanford HAI, 2021), https://www.youtube.com/watch?v=dG628PEN1fY.

staged release of GPT-2 governed the model as information (i.e. weights that could be uploaded to online repositories); the GPT-3 API governed the model as technology (i.e. a piece of machinery with which users could have arm's length interactions); and the NeurIPS ethical review governed papers as contributions to discourse.

Moreover, I have found that actors generally start with the question: what kinds of research outputs can we actually, practically govern? They work backwards from there, identifying the most governable objects as the relevant source of risk. This contrasts with a naive theory where the process of deeming risky certain kinds of research outputs (papers, models, code) would begin with an assessment of the risks of sharing that kind of output. For example, one might claim that models are the most risky thing, because they are the easiest thing for bad actors to weaponise. Or alternatively, one could claim that sharing the code carries greater risks, because it allows bad actors greater flexibility to customise models. However, if we ask, *why – in chapters 1 and 2 – do we find such an emphasis on* models *as a source of risk?* the answer, I will argue, has a lot to do with the *governability* of models. Equally, in chapter 3, if we ask why papers became subject to ethical review, I would argue that the primary reason is that the conference publication system (including peer review) conveniently provides an existing bottleneck through which papers must pass. In other words, the social construction of risk is driven not only by deeper risk philosophies, but largely by the opportunities and barriers to governance (which themselves have technical and institutional roots). The new KCRs I study are all *opportunistic* in their approach to governance. As a mirror image, when I interview researchers who clearly do not want new, risk-focussed KCRs to enter the field, they usually focus very little on arguing that AI is not risky per se. Instead, they focus heavily on the futility of trying to govern AI research outputs. They say AI research is too easy to replicate, including large models. Or they say that it is impossible to extricate the negative applications of AI from the positive ones. Such discussions about risk

must be put in context: they might look like they are arguing against the riskiness of AI, but often they are actually arguing about the structural inability of their field to take action.

Finally, stepping back, I argue that the AI research field is an environment that is hostile to governance, due to its anarchic nature. My study could be usefully read alongside Thomas Gieryn's work on how scientists use rhetoric to defend science against political interference.[58] My empirical material contains many examples of this, especially if you expand "political interference" to include any actor, including AI researchers themselves, who wish to govern AI research. However, the governance-hostile environment is based on much more than rhetoric. As an analogy, I would draw upon James Scott's 2009 book, *The Art of Not Being Governed.*[59] Scott studies the people living in Zomia, a very large, mountainous region that crosses several Asian countries. He argues that their way of life can be explained in the context that, for thousands of years, they have been trying to avoid being integrated into states. Scott writes: 'Virtually everything about these people's livelihoods, social organization, ideologies, and (more controversially) even their largely oral cultures, can be read as strategic positioning designed to keep the state at arm's length.'[60] For example, Scott argues that rice is the perfect crop for state control: it grows above ground, so its value can be assessed by tax collectors; it matures at a specific time of year, and so the state knows when to show up and appropriate the harvest; it travels well, so the farmers can be integrated into national trade; and it requires that a large group of farmers live and work in a fixed location. Scott argues that state-fleeing populations avoid rice in favour of crops like sweet potatoes that are difficult to appropriate (for example, they grow below ground, and are heavy

[58] Gieryn, 'Boundary-Work'; Thomas F Gieryn, *Cultural Boundaries of Science: Credibility on the Line* (Chicago: University of Chicago Press, 1999).
[59] James C Scott, *The Art of Not Being Governed : An Anarchist History of Upland Southeast Asia*, Yale Agrarian Studies (New Haven, Conn.: Yale University Press, 2009).
[60] Scott, x.

to transport) and do not require farmers to settle in fixed, high-density settlements. Scott argues that the people of Zomia are 'barbarians by design'.[61]

We can analyse the AI research field in a similar way. This involves looking at the institutional setup of the AI research field – including its existing KCRs – and asking whether the institutions are conductive or resistant to governance. Such governance might come from states but also might be driven by other actors, or might arise organically within the field. At first glance, the conference publication regime might look like the equivalent of rice: something that makes governance easy. It is transparent, with authors listing their names on papers, and with the papers openly describing the relevant innovation; the conferences occur at a regular schedule; and there is an existing bottleneck, i.e. peer review, through which the paper must travel. These features help to explain the ease with which NeurIPS established their broader impact statement requirement and ethical review.

However, upon closer inspection, the conference publication system – and the broader framework surrounding it – make governance very difficult. The transparency of conference publications and open source code might look like they help accountability, but in a way, this openness helps scientists avoid responsibility for the consequences of their work. As I noted above, the researchers deny ownership of their work, claiming that it belongs to the literature. The techniques and artefacts wind their way into many different applications, but if they are open source, the researchers can maintain complete ignorance of this. If instead the researchers sought greater ownership over their work – and, say, built it into products themselves, or sold access to the work to specific companies and governments – they would quickly run into moral and policy questions. In other words, the openness of scientific research supports researchers in the challenge of – as Hilgartner, Prainsack, and Hurlbut put it – "positioning the creation of new knowledge (the domain of science) as institutionally

[61] Scott, 8.

separate from making societal choices about how to use knowledge (the domain of markets, law, policy, and ethics)."[62] This positioning is especially necessary (and tenuous) in AI research, where — in the case of models and code — the research outputs are usable, technical artefacts.

In addition, the research process is very piecemeal: it unfolds as a very long chain of 8-page papers. As Michael Polanyi pointed out, this can be a fruitful way of coordinating scientific development across time and space[63] (although one of my interviewees would disagree, arguing that the field needs more large-scale, long projects, which are not incentivised under the current system).[64] But scientific efficiency aside, the piecemeal nature of the current system makes it very hard to pinpoint a particular paper as responsible for a particular AI capability. AI capabilities accumulate over time, thanks to thousands of tightly focussed contributions. (This also combines with the fact that AI capabilities are often multi-purpose, useful in many different domains and by actors with various intentions.) As an illustration, many researchers criticised the NeurIPS broader impact statement requirement by arguing either that (a) much AI research is too theoretical or abstract in nature to be linked to specific, real world effects; or (b) papers only have "broader impact" when aggregated into large sub-fields, or even the whole AI field, and so the different statements will be very repetitive. I believe these criticisms point to genuine weaknesses of the initiative, rather than being purely a rhetorical strategy to avoid governance. The crucial point is that these weaknesses come from the underlying foundations of the KCR: it is built as a layer on top of the existing publication regime. The researchers, therefore, do not need to rhetorically invent their own lack of accountability (pace Gieryn) — the existing institutional framework has already done the work for them.

[62] Hilgartner, Prainsack, and Hurlbut, 'Ethics as Governance in Genomics and Beyond', 837.
[63] Polanyi, 'The Republic of Science'.
[64] See Chapter 1.

Another source of anarchy is that the ability to develop a given AI system is distributed across a wide range of research groups who generally hold no obligations to one another. This becomes very important when one research group wants to prevent others from not releasing a particular type of AI system – a recurring issue in my case study.[65] A few structural factors contribute to the fragmented and disorganised nature of the research field. (1) Nearly all the knowledge needed to build the relevant AI system will already be publicly available. This is thanks to the piecemeal nature of the current publication regime (and its high degree of openness). (2) Researchers are quite thinly spread across a wide range of universities and companies. Again, the existing publication regime contributes to this situation, allowing researchers to passively collaborate and learn from one another even when they are geographically and organisationally disparate. (3) Researchers do not sit within overbearing, hierarchical organisations that could coerce them into not pursuing or publishing some line of research. This is most obviously the case in university settings, and less true in corporate labs; although these labs are generally quite academic environments. The field does not have many strong, local power structures that can be knitted together into a field-wide governance regime. In my case study, all these different factors combine to make the anti-proliferation attempts (as with GPT-2 and GPT-3) difficult to maintain.

7. Contribution to literature

As well as contributing to longstanding discussions within the field of science and technology studies (below), my dissertation contributes to the growing literature on the governance of AI.[66] The AI governance literature could benefit from more in-depth, qualitative research on

[65] See Nick Bostrom, Thomas Douglas, and Anders Sandberg, 'The Unilateralist's Curse and the Case for a Principle of Conformity', *Social Epistemology* 30, no. 4 (3 July 2016): 350–71, https://doi.org/10.1080/02691728.2015.1108373.

[66] Allan Dafoe, 'AI Governance: A Research Agenda', July 2017, https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf.

contemporary attempts to govern AI. Such research can help us to better understand the AI field and the social dynamics that operate there. The present time is ripe for this kind of research, now that new AI capabilities are coming online and various actors are considering how the risks of these capabilities should be addressed. My research was fortuitously timed, given that the staged release of GPT-2 (see chapter 1) occurred during the first year of my DPhil studies. Looking back, this episode symbolised the start of a new era in AI research: the rise of large language models, trained by industry labs with large computing resources, and in parallel the concerns over the risks of proliferating these models. This era in AI research is still ongoing, and when trying to understand the latest AI developments and how they might be governed, it will be useful to draw upon the reservoir of experience and insight that comes from a detailed, three-year case study.

In addition, my thesis raises a normative question for AI governance: is academic science the right institutional home for the development of advanced AI? I will return to this question at the conclusion of the dissertation. I argue that the current academic system, where stronger AI capabilities are developed and shared paper by paper, provides a very insecure footing for AI governance. The current system precludes meaningful control over the AI development and deployment, leaving governance efforts with no foothold. This could become a serious problem if the risks from AI – and even AGI – are as grave as some fear. I do not offer concrete solutions to this structural problem, although I highlight some of the possible strengths of corporations as focal points for governance. There are many who are naturally and healthily sceptical of corporations playing this role, including many working on open source AI. One contribution of my thesis could be to highlight that the status quo regime – which revolves around the academic paper at its centre – has its own problems too, and greater adherence to the ideal of academic openness is far from a cure-all for AI governance.

The dissertation also contributes to the STS literature, by building links between Hilgartner's work on KCRs and the broader literature on the governance of science and technology. I especially have in mind three related strands of literature.

First, there is the work of Sheila Jasanoff and others on the role of scientific knowledge in making regulation[67] and (in reverse) the role of regulation in shaping how scientific knowledge is produced.[68] In other words, at a high level, Jasanoff studies the two-way relationship between knowledge and governance.[69] Hilgartner's work on KCRs in genomics speaks to this literature, because he studies the relationship between knowledge (namely, scientific research knowledge) and governance (namely, KCRs). By adding a focus on risk, as I do, we can give the study of KCRs an even stronger link back to Jasanoff's work. The KCRs I study shape, and are shaped by, the ongoing production of knowledge about AI risks. This gives a new dimension to the operation of KCRs, which can play a central role in commissioning and channelling risk knowledge. The thesis therefore contributes to the Jasanoffian project of understanding the relationships between science, risk, and regulation and governance.

Second, there is the literature on the social construction of risks from science and technology;[70] and third, as I have mentioned, there is Thomas Gieryn's work on rhetorical

---

[67] Sheila S. Jasanoff, 'Contested Boundaries in Policy-Relevant Science', *Social Studies of Science* 17, no. 2 (1987): 195–230; Sheila Jasanoff, *Science at the Bar : Law, Science, and Technology in America* (Cambridge, Mass: Harvard University Press, 1995).

[68] Herbert Gottweis, *Governing Molecules: The Discursive Politics of Genetic Engineering in Europe and the United States*, Inside Technology (Cambridge, Mass.: MIT Press, 1998); Sheila Jasanoff, *States of Knowledge : The Co-Production of Science and the Social Order* (New York: Routledge, 2004); Natalie Hannah Porter, 'Ferreting Things out: Biosecurity, Pandemic Flu and the Transformation of Experimental Systems', *BioSocieties* 11, no. 1 (March 2016): 22–45.

[69] On this high-level theme, see also Foucault's later works: Foucault. Michel Foucault, Michel Senellart, and Graham Burchell, *Security, Territory, Population : Lectures at the Collège de France, 1977-78* (Basingstoke: Palgrave Macmillan, 2007); Michel Foucault, *The History of Sexuality, Volume I : The Will to Knowledge* (1976; repr., UK: Penguin, 2020).

[70] Mary Douglas and Aaron B Wildavsky, *Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers*, Ebook Central (Berkeley: University of California Press, 1982); Stephen Hilgartner,

attempts by scientists to draw a boundary between scientific and non-scientific arenas. I would argue that both these two literatures can be enriched by an institutionalist lens (even if I am not the first to attempt this). Much of the literature on risk construction focuses on how different ways of measuring risk, and different ways of thinking and talking about risk, contribute to actors' understanding of risk. While valuable, this must be complemented by research into the role played by the institutional context.[71] My research pushes against the idea that expert communities, relying on their toolbox of epistemic devices (such as the precautionary principle, cost-benefit analysis, or scenario-planning exercises), construct ideas about risk that are then simply plugged into governance regimes. Rather, the governance regimes I study – and the structure of the AI research field more broadly – continuously help to identify, study, and package risks, in tandem with governing those risks. As Herbert Gottweis put the point, governance 'is never simply a reaction to a problem and is always co-productive of the problems to which it seems to react'.[72] The social construction of risks from AI cannot be separated from the (shifting) institutional and technical context of AI research; and in particular I would argue that the tools we have for sharing research (conference publications; open source repositories; and now APIs) have significant and unnoticed flow-through effects on the construction of risks from AI. This is interesting because the AI research field is currently going through a period of institutional change – with research becoming much more resource-intensive, and with industry labs taking an

'The Social Construction of Risk Objects: Or, How to Pry Open Networks of Risk', in *Organizations, Uncertainties, and Risk*, ed. James F. Short and Lee Clarke (Boulder, CO: Westview Press, 1992), 39–53; Andrew Lakoff, 'Preparing for the Next Emergency', *Public Culture* 19, no. 2 (1 May 2007): 247–71, https://doi.org/10.1215/08992363-2006-035; Peter Andrée, *Genetically Modified Diplomacy : The Global Politics of Agricultural Biotechnology and the Environment* (Vancouver, Canada: UBC Press, 2007), http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=3412552; Steve Maguire and Cynthia Hardy, 'Riskwork: Three Scenarios from a Study of Industrial Chemicals in Canada', in *Riskwork* (Oxford: Oxford University Press, 2016), https://doi.org/10.1093/acprof:oso/9780198753223.003.0007.

[71] See, for example: Benjamin Hurlbut, *Experiments in Democracy : Human Embryo Research and the Politics of Bioethics* (New York, US: Columbia University Press, 2017), http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=4771915.

[72] Gottweis, *Governing Molecules*, Location 1942. The quotation refers to 'policymaking' rather than 'governance' more generally.

increasingly prominent role – which provides new opportunities for governing (and therefore socially constructing) AI risks.

Similarly, Gieryn's focus is on scientists' *rhetorical* boundary drawing. I would argue that the boundary between scientists and non-scientists, as it relates to AI risks, is made of something firmer and more practical than rhetoric. The regime of scientific publication functions exactly as Gieryn describes the role of rhetorical boundary-drawing. Gieryn says that rhetorical boundary-drawing works to protect scientific autonomy, in that scientists 'construct a boundary between the production of scientific knowledge and its consumption by non-scientists (engineers, technicians, people in business and government)'.[73] He continues: 'The goal is immunity from blame for undesirable consequences of non-scientists' consumption of scientific knowledge.' The same description is apt for the scientific publication regime, which allows AI researchers to play a central role in building the technology whilst holding its societal impact as beyond their control. It should not be surprising that, of all the three newer KCRs I study, the one that gives the lab greatest ownership over the social impacts of the technology is the GPT-3 API, which is also the most radical departure from the existing publication regime. The upshot is that the institutions themselves (i.e. the KCRs) play a central role in positioning the boundary between science and society.[74]

[73] Gieryn, 'Boundary-Work', 789.

[74] In the terms of institutionalist theory, publication practices help to shape the boundaries of the field. See Zietsma and Lawrence, 'Institutional Work in the Transformation of an Organizational Field: The Interplay of Boundary Work and Practice Work'.

CHAPTER 1: STAGED RELEASE

Abstract: This chapter examines the staged release of GPT-2, a large language model trained by OpenAI. I start by describing the background to OpenAI's decision to release GPT-2 in a staged manner. The rest of the chapter is devoted to analysing the new, 'staged release' regime and how it interacted with the AI research field. The regime relied upon coordination between different AI research labs that would have been able to reproduce GPT-2 themselves, and I describe the challenges of such coordination. The 'staged release' regime also positioned different actors inside and outside the AI research community in a new way. I look at how the regime organised the production of knowledge about GPT-2's risks, and attempted to build that knowledge into ongoing decisions about GPT-2's release schedule. The 'staged release' regime sets up a feedback loop between the lab and the outside world: the lab monitors how the model is being used in the world, which informs how the lab shares larger versions of the model. In this way, the 'staged release' regime was an innovation in the relationship between science and society, making them more closely interconnected. This was a break from prevailing ideas within AI research about the proper role of the scientific research community, and I discuss four visions of the science-society relationship that conflicted with OpenAI's 'staged release' regime.

1. Introduction

In February 2019, OpenAI – an AI research lab based in San Francisco – made a blog post announcing their new model, GPT-2:

> Our model, called GPT-2 (a successor to GPT), was trained simply to predict the next word in 40GB of Internet text. Due to our concerns about malicious applications of the technology, we are not releasing the trained model. As an experiment in

> responsible disclosure, we are instead releasing a much smaller model for researchers to experiment with, as well as a technical paper.[75]

This was the latest in a series of impressive results in the area of natural language processing (NLP). In 2018, OpenAI had released the original GPT model, and later that year, Google researchers had released the BERT model.[76] These models all relied on the Transformer architecture, a recently developed method for building neural networks.[77] They were also all trained in an "unsupervised" manner. The model is essentially trained on a game of *guess the missing word*, known as language modelling. The model is fed some text from the internet and must predict what the next word was. This means the researchers can use very large datasets – e.g. including all English-language Wikipedia articles – scraped from the internet, without needing to build the dataset themselves. The model is a neural network with millions of different connections. During the training process, the weights of each connection (the "parameters") are automatically updated based on the model's successful or unsuccessful predictions. The language modelling task is used as a "pretraining" stage, which teaches the model generally applicable knowledge and skills. A pretrained model like GPT-2 can subsequently be trained again, on a more specific task (such as text classification, using human-labelled datasets), and it will often outperform models that did not go through the initial pretraining stage.

GPT-2 was a scaled-up version of the models that had gone before it. The text dataset was larger, the model itself was larger (the largest GPT-2 model had 1.5 billion parameters,

---

[75] Alec Radford et al., 'Better Language Models and Their Implications', *OpenAI Blog* (blog), 24 February 2019, https://openai.com/blog/better-language-models/.

[76] Jacob Devlin et al., 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding', *ArXiv:1810.04805 [Cs]*, 24 May 2019, http://arxiv.org/abs/1810.04805.

[77] Ashish Vaswani et al., 'Attention Is All You Need' (arXiv:1706.03762 [cs.CL], 2017), https://arxiv.org/abs/1706.03762.

compared to 117 million for GPT-1 and 345 million for BERT), and it was trained using more computational power. GPT-2 had very impressive capabilities. It was able to perform tasks like reading comprehension even without explicitly being trained to perform such tasks. Also, the model had an unprecedented ability to generate coherent passage of text. You give the model a few lines of text and successively sample from its predictions about what words are most likely to follow. The blog post contained an impressive example of this, where GPT-2 was fed the prompt:

> In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2 continued this text as follows:

> The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.
>
> Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.
>
> Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

And so on. The model has learnt not only syntax, but also is able to predict that, for example, that scientists working in the Andes might have names like "Dr. Jorge Pérez" and might be based at the University of La Paz. The generations were not always entirely coherent, and it would often take several tries to get high-quality outputs.

After introducing GPT-2, the blog post announced that OpenAI was experimenting with a new approach to sharing the model:

> Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter: "we expect that safety and security concerns will reduce our traditional publishing in the future, while increasing the importance of sharing safety, policy, and standards research," and we see this current work as potentially representing the early beginnings of such concerns, which we expect may grow over time. This decision, as well as our discussion of it, is an experiment: while we are not sure that it is the right decision today, we believe that the AI community will eventually need to tackle the issue of publication norms in a thoughtful way in certain research areas.[78]

In other words, the lab was experimenting with a new knowledge-control regime. Initially, this looked like it was going to be a regime of "partial release", but the post hinted that OpenAI might return to this question.[79] A few months later, in May 2019, OpenAI updated the blog post to clarify that the regime was one of "staged release":

[78] Radford et al., 'Better Language Models and Their Implications'.
[79] The post states: "We will further publicly discuss this strategy in six months."

> Staged release involves the gradual release of a family of models over time. The purpose of our staged release of GPT-2 is to give people time to assess the properties of these models, discuss their societal implications, and evaluate the impacts of release after each stage.

The release was "experimental" on two levels. On one level, the hope was that releasing smaller versions of GPT-2 would allow OpenAI to better understand the risks of these models. They could see how one of these models was used in practice, and more generally, OpenAI and the broader research community could further study the capabilities of the model. The release was also experimental in the sense that the lab was trying out a new knowledge-control regime. As the post mentioned, OpenAI staff were not confident about the risks of GPT-2. But they predicted that, with continued progress in AI capabilities, it would become increasingly important that the AI community had a well-developed process for dealing with potentially risky AI models.

1.1 Overview of the chapter

In this chapter, I focus on the story of GPT-2: how the release decision was reached, how the AI research community reacted, and the success or failure of the proposed regime. The GPT-2 episode speaks to the core themes of my thesis. We can ask: did the AI research community act in a coordinated way, or did other research groups undermine OpenAI's efforts? Were the technical features of GPT-2 conducive to governance? How was knowledge about the risks of GPT-2 produced, and how did that feed back into the way GPT-2 was shared? How did the "staged release" regime position the relationship between industry and academia? Was the new regime compatible with the existing regime of scientific publication?

Section 2 explains the lead-up to OpenAI's decision to withhold release of GPT-2. This serves as a good introduction to the topic, and I tell the story without much theoretical analysis. The other chapters all analyse some aspect of the 'staged release' regime and how it interacted with the AI research field. Section 3 looks at the problem of identifying novel (and risky) AI capabilities that should be put through the 'staged release' process. Section 4 asks whether coordination within the AI research community, over the staged release of GPT-2, was possible. Section 5 analyses how the 'staged release' regime distributed – or failed to distribute – the responsibility for studying the risks of GPT-2 across the broader research community. Section 5 looks at how GPT-2 positioned the relationship between science and society, and how this was controversial within the AI research community.

The GPT-2 episode **speaks directly to the core themes of the thesis** in three ways. First, we see the difficulties inherent in a research lab seeking to control the proliferation of AI capabilities. The new KCR relied on many other researchers following OpenAI and not releasing their own versions of the new model. This was very challenging, given the uncoordinated nature of the AI research community, the incremental nature of AI research, and the fact that GPT-2 was not prohibitively expensive to train for many actors.

Second, the 'staged release' regime – as a new, risk-based KCR – had to confront the question: how to organise the production of knowledge about GPT-2's risks? The core idea was that OpenAI would learn over time about the risks of GPT-2 and integrate that into ongoing decisions about release. In this case, the research community was not given a very central role in this process, which relied more heavily on OpenAI monitoring for real-world misuse. Insofar as there was a conflict between granting researcher access and preventing wider proliferation of GPT-2, OpenAI chose to limit proliferation. Many researchers objected

to this, viewing themselves as well-placed to contribute to the production of knowledge about GPT-2 – in particular, by studying the model.

Third, the stage release regime was a break from how AI researchers normally conceptualised the relationship between their research community and the wider world. Often, researchers view societal impacts as simply downstream of scientific research, with researchers having little ability to actually steer these impacts via how they share their work. The staged release regime presented a different picture: the research lab should actively seek to shape the social impact of its work by moderating the timing and content of ongoing releases.

2. The lead-up to the GPT-2 release decision

2.1. The malicious use of AI

The GPT-2 blog post foregrounded the concern of malicious use. This would include, for example, the production of "misleading news articles" and abusive or fake social media posts. Before GPT-2, some of the conceptual groundwork around "malicious use" had already been laid. The first mention of the phrase "malicious use of AI" I have found is by Ian Goodfellow, a well-respected AI researcher. In 2016 he posted an answer on Quora:

> I think the most important issue in AI safety is malicious use of AI by human beings. A lot of AI safety discussions revolve around the idea that someday AI might be superintelligent and pursue goals that we didn't intend for it to pursue. Such discussions seem to imply that the right way to defend against malicious AI is to make sure that AI's goals never depart from ours. I think that point of view is dangerously limited. There will definitely be AI with goals that are opposed to at least some

> people's goals, because different people have different goals and different people build and control AI.
>
> Already, we see malicious use of simple AI, for things like breaking CAPTCHAs. These are relatively minor issues compared to using AI to do targeted assassinations. But it's a difference of degree, rather than kind. As AI gradually becomes more capable, people will use it to do greater harm as well as to do greater good.[80]

At this point, there was already a loose community of AI researchers and others who were concerned about long-term risks from advanced AI. A seminal text is Nick Bostrom's (2014) *Superintelligence*, which argued that superintelligent AI systems might in future present an existential risk to humanity.[81] Stuart Russell, an AI researcher at UC Berkeley, made a similar argument in his 2019 book, *Human Compatible: Artificial Intelligence and the Problem of Control*.[82] Both Bostrom and Russell give reasons for why it might be difficult to design a superintelligent AI system that can reliably act in the interests of humans. The "classic" story, then, is that agent-like AI systems cause a major catastrophe contrary to the intentions of its human creators. As Ian Goodfellow explains above, the category of "malicious use" risk is different, because the intention to cause harm need not arise within the AI system, but could be supplied by the human operator.

The category of "malicious use" became more concrete in 2018 with the publication of a report titled 'The Malicious Use of Artificial Intelligence'.[83] There was a long list of authors,

---

[80] Ian Goodfellow, 'When Do You Expect AI Safety to Become a Serious Issue?', *Quora*, 11 August 2016, https://www.quora.com/When-do-you-expect-AI-safety-to-become-a-serious-issue.
[81] Nick Bostrom, *Superintelligence : Paths, Dangers, Strategies* (Oxford, UK: Oxford University Press, 2014).
[82] Stuart Russell, *Human Compatible: AI and the Problem of Control*, 1st edition (London: Allen Lane, 2019).
[83] Miles Brundage et al., 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', *ArXiv:1802.07228 [Cs.AI]*, 2018.

largely academics at Oxford and Cambridge universities, as well as staff at OpenAI. The authors were largely a mix of social scientists and computer scientists. Malicious use was defined "loosely, to include all practices that are intended to compromise the security of individuals, groups, or a society."[84] The report argued that AI is a dual use technology, with specific AI systems carrying having both beneficial and harmful applications:

> For example, systems that examine software for vulnerabilities have both offensive and defensive applications, and the difference between the capabilities of an autonomous drone used to deliver packages and the capabilities of an autonomous drone used to deliver explosives need not be very great.[85]

The report also emphasised the ease with which novel AI systems proliferate, referring to the openness of the scientific community:

> AI developments lend themselves to rapid diffusion. While attackers may find it costly to obtain or reproduce the hardware associated with AI systems, such as powerful computers or drones, it is generally much easier to gain access to software and relevant scientific findings. Indeed, many new AI algorithms are reproduced in a matter of days or weeks. In addition, the culture of AI research is characterized by a high degree of openness, with many papers being accompanied by source code. If it proved desirable to limit the diffusion of certain developments, this would likely be difficult to achieve . . .[86]

[84] Ibid, p.9
[85] Ibid, p.16
[86] Ibid, p.17

The report therefore identified "exploring different openness modes" as a priority area for future research. It pointed out that, despite the benefits of open publication, it 'increases the power of tools available to malicious actors.'[87] The report continued: 'This raises an important research question: might it be appropriate to abstain from or merely delay publishing some findings related to AI for security reasons?' The report left this question open, and advocated for debate and careful thought.

2.2 GPT-2 inside the lab

As one of the authors of the Malicious Use Report, and a member of OpenAI, told me: "ultimately for things like publication norms, talk is cheap; and it's easy to say *well like people should figure out the right publication norms* or whatever and harder to actually do something about it."

OpenAI is a private research company based in San Francisco, which began as a non-profit. OpenAI's stated mission is to build 'safe and beneficial AGI', where AGI stands for artificial general intelligence, which they define as 'highly autonomous systems that outperform humans at most economically valuable work'.[88] OpenAI's researchers are mainly technical, but they also have a policy team, which studies the social impact of AI and how the technology should be governed.

At the time of the Malicious Use Report, some of OpenAI's members had already been thinking about the problems raised by open publishing of their research. Soon after the report, OpenAI created an organisational charter, which included the line: 'we expect that safety and

[87] Ibid, p.54
[88] OpenAI, 'OpenAI Charter', 2018, https://openai.com/charter/

security concerns will reduce our traditional publishing in the future' (OpenAI, 2018). OpenAI was the only prominent AI lab that was publicly making this argument at this time.

I interviewed CAMERON, a senior member of OpenAI, who told me the story of how the staged release of GPT-2 came about. He told me:

> 2018 was the year in which GPT-2 was being developed. And so we were playing around with it internally at OpenAI. The way that almost all release decisions are made at OpenAI stems from a load of people in the organisation playing with something that we have developed, and GPT-2 was no different. And the reason why I'm telling you this is I think this sort of experiential thing is important – having a load of people in the org who are just sort of prodding your technical artefact. And what kind of happened is we started talking to people about it, and you know, we'd sort of say, "Well, we've got like, deep fakes for text. Is that something to worry about or not?" And generally, I think people were pretty confused, including us, and including me, frankly. Like, lots of people reacted with quite a lot of worry to the idea of it. Especially when we saw some samples, because you know, we've all normalised GPT-2, but you've got to remember at the time, it was completely wacky stuff; it didn't really make sense relative to previous NLP things, it seemed like a very significant qualitative advance.

CAMERON remembers people telling him things like "don't release this before the midterms", which were in November 2018. Another of my OpenAI interviewees mentioned the political context, including misinformation on Facebook during the 2016 and 2018 elections; this combined with the fact that GPT-2 would write stories about things that were

completely untrue. The concerns increased as, during training, the model became more capable. CAMERON recalls playing with GPT-2 internally around December or January:

> I was playing with GPT-2 on my laptop in the airport, generating samples of I think fake news about Theresa May and Brexit at the time, along with fake news about Miley Cyrus shoplifting . . . And I just remember being like, "holy shit, this has got really good". And I remember talking to a few people on Slack and just posting samples that I had generated, and being like, "this sample was generated on the first roll; this one was generated after that one" — you know, we're very careful about like, not cherry picking internally on this kind of thing. And I think that got people to freak out a lot more. (Not in the sense of I was trying to freak them out.)

So, GPT-2 was a problem chasing a solution. At the same time, greater caution over releasing models was, in reverse, a solution chasing a problem: GPT-2 was an opportunity to get the ball rolling on experimenting with new KCRs. CAMERON told me that OpenAI:

> wanted to withhold release, because we thought it was both important from a potential malicious use thing, but also a symbolic thing. A thing that drove a lot of this thinking along with potential for misuse was like, What if we take a stance and do something different to the default?[89]

ROBIN, another member of the OpenAI policy team, also told me that there was "some symbolism" involved. The OpenAI staff had one eye on the future. CAMERON explained:

[89] He continued: "Which is going to suck for everyone, it's gonna be awful — we weren't naive about how badly this would pan out."

> a lot of thinking about GPT-2 was: progress is rapid, it's going to continue to get more rapid; if we don't start changing publication norms now, when it will seem like a big joke, we're not going to be ready to change them in five years when it's very important. And I am feeling confident in that; like, I think this year [2020], some weird things exist, and some weird things are being released. I think in two years from now, it's gonna look really, really odd that everything was being released as open source forever.

The final factor contributing to the GPT-2 release was the prospect of other labs training similar models to GPT-2 and releasing them in the normal way. In January 2019, before the GPT-2 release, a team of Google researchers released a blog post about Transformer-XL, an architecture related to that used for GPT-2.[90] CAMERON and others at OpenAI read this, and saw that the post mentioned, as part of a list of 'exciting potential applications', the prospect of 'generating realistic, long articles'. CAMERON told me:

> And that drove more of our release thinking because we were like, *Shit, Google's doing this as well. Google's definitely not going to do a crazy release philosophy here. They're trying to bury this stuff.* So another thing that drove the things we did was being like, *Well, shit, other people are just about to start talking about this stuff. We should try and frame this.*

The experiment with 'staged release' relied on being the first to the punch, both because otherwise the capability would already be widely available, but also because it would give OpenAI the chance to set the discourse around text generation.

[90] Zhilin Yang and Quoc Le, 'Transformer-XL: Unleashing the Potential of Attention Models', *Google AI Blog* (blog), 29 January 2019, https://ai.googleblog.com/2019/01/transformer-xl-unleashing-potential-of.html.

2.3 The release

The initial GPT-2 release was on 14 February 2019. The blog post linked to a preprint of the scientific paper, which described how the model was built and recorded the results of various tests of GPT-2's performance. The post also linked to a GitHub repository. This contained some of the code necessary for training GPT-2, as well as a 117 million parameter model. The paper refers to models of other sizes – 345m, 762m, and 1.5B – and these were not released.

The GitHub repository also contained a licence. It was an MIT Licence, common for open source software on GitHub, which allows the user to "deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software". They made one modification to the standard MIT Licence, with the following line inserted:

> We don't claim ownership of the content you create with GPT-2, so it is yours to do with as you please. We only ask that you use GPT-2 responsibly and clearly indicate your content was created using GPT-2.

It is not clear that this language creates a legally enforceable obligation, and using GPT-2 "responsibly" is not defined. I am not aware of any attempts by OpenAI to enforce this licence. The licence was therefore not a core part of the new, "staged release" knowledge-control regime that OpenAI was putting forward. ROBIN told me in 2021 that OpenAI had considered the option of using licences to tackle misuse of models, and mentioned the downside that licences are not practically enforceable around the world (for example, against authoritarian governments).

3. The identification problem

The ‘staged release’ regime attempts to control the rate at which a new capability proliferates. This raises what we can call the *identification problem*: the lab must pinpoint specific capabilities of its AI system that are both risky and not already widely available. GPT-2 was an unusually good candidate for staged release because the ability to generate such coherent text was only just becoming within reach for the research community. The quality of text generation was novel, even if the fundamental methods were not. However, the requirement – inherent in the logic of staged release – to identify novel capabilities of GPT-2 comes with a couple of difficulties.

First, research progress in AI research is usually very incremental. Incremental progress is a natural result of the existing, conference publication KCR. If, say, a team of 50 researchers spent three years working on a big project, the payoff would be low: each would have made a proportionally small contribution to just a single paper. JAMIE, a researcher at DeepMind, made this point well. He complained that the existing system gives so much of the credit to the first and last author on a paper. He continued:

> It disincentivises people to work together. Versus like if you work on Gmail, it's not like there's one day going to be a list of names and you're going to be ranked in terms of how valuable you were considered by the Gmail team on Gmail. Because if you did that, everyone is basically not valuable. Except, the first three people feel extra, extra valued, and then everyone is worthless. And that's terrible, because it's such a big piece of complex software. It's obvious that many, many people uniformly did lots of different things, and if they didn't do that then it wouldn't be as good. Now, research projects. Ideally you would have more grand vision research projects. Where

> that's the case, it's just hard with the academic system to pull people together to do that. . . .[E]ven like big projects where there's a lot of glory to be had. You don't end up having that many people working on them, really.[91]

Hilgartner found the same phenomenon in the Human Genome Project, where 'Long-term mapping projects would take too long and involve too many contributors to help early career researchers to build a distinctive scientific identity.'[92]

The upshot in AI research is that would-be 'breakthroughs' accrete over time, across many different papers. This makes it difficult to point to a specific paper or model as extending the research frontier enough to independently constitute a novel source of risk.[93] For example, in response to the initial GPT-2 release, one AI research criticised OpenAI's approach by tweeting:

> But their perplexity on wikitext-103 is 0.8 lower than previous sota. So it's dangerous now. 🙃

This is referring to a measure of how accurately the model can predict missing words from Wikipedia articles. (Lower is a better score, and "SOTA" stands for "state of the art".) On this measure, GPT-2 was the strongest model available, but still an incremental advance. Similarly, another researcher blogged:

---

[91] See also Yoshua Bengio, 'Time to Rethink the Publication Process in Machine Learning', 27 February 2020, https://yoshuabengio.org/2020/02/26/time-to-rethink-the-publication-process-in-machine-learning/.

[92] Stephen Hilgartner, *Reordering Life : Knowledge and Control in the Genomics Revolution*, Inside Technology (Cambridge, Massachusetts: The MIT Press, 2017), 108.

[93] An interviewee working at the Partnership on AI told me: "I think one of the challenges is because progress is so incremental, and so lots of people are working on like roughly the same things just like slightly, like making slight improvements - That's one of the challenges of this idea of deciding which piece of research is the point at which you don't publish it. Like, at what point does it become risky if it's just incremental improvements? So yeah, I don't really have a good answer for this yet."

> ...what makes OpenAI's decision puzzling is that it seems to presume that OpenAI is somehow special—that their technology is somehow different than what everyone else in the entire NLP community is doing—otherwise, what is achieved by withholding it? However, from reading the paper, it appears that this work is straight down the middle of the mainstream NLP research.

This highlights the difficulty of cleanly identifying the properties of the model that are novel and risky. At the same time, the requirement to focus on novel capabilities skews the discussion towards a subset of the risks of a given model. OpenAI's blog post focussed on the risks of text generation, because this was the area where other similar models (like BERT) were clearly weaker. Nonetheless, these large language models pose other risks, too – for example, they could be used for automated surveillance of any kind of text. Indeed, many of the capabilities of large language models have nothing to do with text generation, such as text classification, search, translation, and summarisation. However, withholding GPT-2 on the grounds that (for example) all these capabilities are surveillance-relevant would not have had a cleanly identifiable impact, because GPT-2 would only marginally (if at all) outperform existing models, like BERT, on these tasks.

Second, if a lab draws attention to the novelness of its own work, it will very easily be read by the research community as trying to "hype up" its own research out of academic or commercial self-interest. To add to this difficulty, the AI research community is especially sensitive to industry labs taking too much scientific credit for their research. The industry labs are widely viewed as more effective at marketing their papers, much to the chagrin of university-based researchers working on similar topics. A common criticism is that this interferes with the allocation of credit for advances in AI, with the papers from industry labs

getting an outsized number of citations. From the perspective of many researchers, OpenAI's claim that GPT-2 had novel capabilities became, rather than a justification to trial the staged release regime, an attempt to take credit for that as a breakthrough. One might wonder, why would a lab try to take credit for something that is a risk to society? The explanation is that the creation of a novel, powerful capability will often be a sufficient condition for both credit and risk. It is difficult, then, for a lab to make a claim about the risks of its work without also being read as making an implied claim for credit. Again, we see that the existing publication regime provides a hostile environment for the task of identifying novel and potentially dangerous capabilities. Not only does the current system favour incremental advances, but it also encourages a preoccupation with scientific credit allocation that spills over into discussions about risk.

4. Coordination within the AI community

4.1 Staged release depends upon coordination

The "staged release" KCR can only be successful insofar as it successfully delays proliferation of the relevant models. The main threat is that other research groups will replicate and open source those models. Some researchers criticised OpenAI's decision on these grounds: replication would be so swift that the whole enterprise was futile. And more generally, AI researchers often present the field as incapable of doing anything about malicious use (see section 6, below). On the other hand, those concerned about the long-term risks of AI – including some of the OpenAI staff – often emphasise the importance of *coordination* between AI labs, as necessary for avoiding a race-to-the-bottom on safety standards.[94] The staged release regime clearly relied upon coordination between research

[94] Bostrom, *Superintelligence : Paths, Dangers, Strategies*; Nick Bostrom, Thomas Douglas, and Anders Sandberg, 'The Unilateralist's Curse and the Case for a Principle of Conformity', *Social Epistemology* 30, no. 4 (3 July 2016): 350–71, https://doi.org/10.1080/02691728.2015.1108373; Amanda Askell, Miles Brundage,

groups, because many research groups would have the ability to reproduce GPT-2. The rest of 2019 was therefore a test for the new "staged release" KCR. Was it a viable regime? Was the AI research community able to coordinate around delaying proliferation, or is the field too anarchic in nature? In other words, the governability of the research field was at stake.

I would argue that the institutional structure of a large scientific field – including the prevailing publication regime – is ill-suited to coordinating around delayed release. Firstly, the research community is fragmented across many different organisations. For example, at ICML 2020, there were over 1,000 papers accepted, coming from nearly 500 different organisations.[95] In analyses of conference publications, much is often made about how many papers are published from the top industry labs. However, there is a long tail of smaller research groups who, as we will see, will sometimes have the capacity to replicate the work of top labs. Second, many of the research groups are within universities – normally, over 50% of the papers come from universities.[96] University departments and labs, compared with industry labs, have little top-down bureaucratic power to prevent researchers from publishing a particular piece of work. Third, as we saw in the previous section, progress in AI is very incremental. Therefore, most of the necessary techniques for replicating a model will already be available in the literature. Finally, researchers are heavily incentivised to publish papers. This is especially true for "hot" topics like large language models. Even with the claim that GPT-2 was risky, researchers will naturally be inclined to turn this into a research question that could furnish a paper (e.g. how risky is GPT-2, really?). All these institutional factors

and Gillian Hadfield, 'The Role of Cooperation in Responsible AI Development', *ArXiv:1907.04534 [Cs]*, 10 July 2019, http://arxiv.org/abs/1907.04534; Miles Brundage et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims* (arXiv:2004.07213 [cs.CY], 2020).

[95] Khiem Pham, Duong Le, and VinAI Research, 'An Overview of ICML 2020's Publications', 7 July 2020, https://www.vinai.io/an-overview-of-icml-2020s-publications.

[96] For example, at ICML 2020, 54% of papers had academic-only affiliations, with the rest either coming from industry-only (8%) or mixed (38%).

make it very difficult to orchestrate a coordinated delay on the release of a model like GPT-2.

At the same time, the coordination problem depended upon technical factors that affected the ease with which GPT-2 could be reproduced. If reproducing GPT-2 was extremely easy, then OpenAI would not just have to coordinate with the research field, but a much wider range of actors. An assistant professor at NYU was quoted in the media: “I’m confident that a single person working alone with enough compute resources could reproduce these results within a month or two.” I had interviewees who made the same argument. CAMPBELL, a PhD student in Europe, told me: “In general. . . you can’t really hold technological progress back. . . . It’s the same with the atomic bomb and nuclear energy. So usually, it’s more like we’re providing tools, and what people do with these tools is very much up to them. . . . Just because it’s possible to misuse certain tools does not mean we should forbid them or prevent them from being published at all.” SAM, a PhD student at UC Berkeley, told me: “I think that the solution is not going to be any kind of, like, constraint on the technology, because people will just go around that. So I don’t think we should be looking for solutions like “you can’t make facial recognition systems that are this good” or “you can’t make language generation that’s this good." I think that’s going to fail.”

A live issue was whether the unprecedented size of GPT-2 would limit the number of research groups who could replicate the model. Crudely, a trained model is the result of: code + data + compute. Assuming a researcher has a paper and accompanying code, their ability to reproduce the model will depend upon the ease of accessing enough data and compute. Now that the research field is moving towards larger-scale models (as symbolised by GPT-2), this distance between (on the one hand) the paper and code, and (on the other hand) the trained model is getting larger and larger. In the case of GPT-2, the dataset was scraped from publicly

available text on the internet, and so access to the data was less of a barrier. The computational cost, purchased from a cloud computing provider, would be on the order of 25 to 50 thousand dollars (see below). These technical factors bounded the range of actors who could feasibly reproduce GPT-2 – although, as we will see, a team at Google's cloud computing service was keen to assist with GPT-2 reproduction projects for free.

Overall, the staged release of GPT-2 was a partial success. On the one hand, it was about 4 to 7 months before a malicious actor would have been able to download a model similar to the 1.5 billion parameter GPT-2. And there were several research groups who decided to follow OpenAI, not releasing larger models at least until OpenAI had. On the other hand, not all research groups played along, with some researchers knowingly undermining the staged release regime. In this section, I will go through different occasions where GPT-2-like models were created and how they were shared. I address two key questions. Given the technical ease or difficulty of reproducing GPT-2, how large was the pool of actors that needed to coordinate around the 'staged release' of GPT-2? Second, within that pool of actors, was coordination possible?

4.2 A GPT-2 model trained specifically for fake news

Almost four months after the initial GPT-2 release, in late May 2019, a team of researchers published the *Grover* paper.[97] The team was from the University of Washington and the Allen Institute for AI, a non-profit research lab. The Grover models were similar to the GPT-2 models, except they were specifically designed to generate fabricated news stories. The training data was a large corpus of news articles.

[97] Rowan Zellers et al., 'Defending Against Neural Fake News', *ArXiv:1905.12616 [Cs.CL]*, 2019.

The team had additionally trained a set of "discriminator" models, whose function was to look at an article and detect whether it had been generated by a Grover model or by humans. The team of researchers held back the larger Grover models, equivalent to the large, unreleased GPT-2 models. Nonetheless, other researchers were given access: researchers just had to email the lead author of the Grover paper. In addition, the team uploaded to GitHub: the training code (going beyond what OpenAI had released); smaller models for generating news; and models of all sizes for detecting machine-generated news (the "discriminator" models).

The project was fast: it started in April and finished in late May. The results were impressive. Human participants rated Grover-generated propaganda-style articles as more trustworthy than human-written articles of the same style. Grover was adept at generating 'fake news', having been trained specifically to do so. Also, the training data included metadata like the headline of an article and the URL it came from. This made Grover's outputs easier to guide, because the operator specifies the headline and URL as an input, which give strong clues about what the content should look like.[98] And (likely for the same reasons) Grover's outputs were more consistent than GPT-2, requiring fewer attempts to generate a realistic article.[99]

[98] An extract from my interview with the lead author:

> Toby Shevlane:
> So, in a way, it's more directable. It's more like a gun that you can point in a particular direction, rather than a shotgun or something.
>
> Rowan:
>
> Exactly. Yeah, that's a really good metaphor. And that's why we built it for this, threat modelling, of trying to figure out what an adversary might do with these generators, because as we know, an adversary isn't going to want to just generate random news. They'll want to generate news that fits their agenda, or that is lucrative.

[99] Source: interview with the lead author of the Grover paper.

The team mirrored OpenAI by holding back the larger models, but the overall tone of the Grover release was critical of OpenAI's "staged release" regime. The most acute criticism – as Rowan, the lead author, explained during our interview – was that OpenAI did not release the larger versions of GPT-2 to the research community. Rowan's strategy was to give the model weights to researchers who emailed him. He saw OpenAI's approach as harmful because it did not allow researchers to study the larger models and find defensive solutions. He thought this was the overriding imperative. He told me:

> I think the biggest danger is not releasing it. Because if we don't release it, and we can't study it, then when adversaries do scale this up and have better model generators then we'll have no answer to this. That to me is the biggest worry.

I got the impression from talking to Rowan that he would have been happy to publicly release the largest model, because he thought the discriminator models were a sufficient defence. He told me, 'I'm not worried, and I don't think you should be worried either, about this model', because 'we have a good solution for right now' (i.e the discriminator models).

Rowan told me that he did not publicly release the largest model immediately because that might "rub people the wrong way". He had spoken to some other NLP researchers who did see these models as dangerous, and had heard about a workshop at a recent NLP conference where some participants were saying the same thing.[100] He saw his approach as striking a middle-ground, where he was accommodating some of OpenAI's concerns around proliferation whilst still allowing researchers full access.

[100] Rowan told me: "a couple of people are like, well if you release this specific model right now, before everyone gets a chance to implement the safeguards, like Grover Mega as a discriminator, then it might cause some issues. And so I think because of that, we've kind of tried to stay right in the middle of that."

Did the Grover release undermine the 'staged release' regime? The picture is mixed. On the one hand, the research team was supporting the regime by delaying release of the larger models. The fact that they disagreed with OpenAI's assessment of the risks of GPT-2 makes this even more significant. On the other hand, they released all the code necessary for training a GPT-2-like model. Their process for giving out the largest Grover model was also very insecure: Rowan would Google search the name of the person emailing him; he told me it 'maybe also helps if you have a name that's like real and maybe even an educational address'. There is also the question of whether an attacker could recover the generator model from the open source discriminator model, which is just an adapted version. Further, Grover was better at producing fake news articles than GPT-2, and so it is unclear that matching OpenAI's release by model size is the right approach; and regardless, they subsequently released the largest Grover model in September 2019, before OpenAI released the largest GPT-2 model, citing the availability of other GPT-2-like models.[101] Finally, the research team did not communicate with OpenAI before the initial Grover release.

4.3 An undergraduate student reproduces GPT-2

A different threat to the "staged release" regime came from an undergraduate student in Germany called Connor. In June 2019, Connor uploaded a blog post where he announced that he had replicated the largest GPT-2 model, and was planning on open sourcing the model unless anyone could change his mind.[102] The staged release regime, in theory, relies upon some kind of duty upon other researchers to play along. While this might seem enforceable among colleagues in a research field, it looks harder if the pool of people who need to play

[101] The GitHub page for Grover explains: 'Now that several months have passed since we put the paper on arxiv, and since several other large-scale language models have been publicly released, we figured that there is little harm in fully releasing Grover-Mega.' Available at: https://github.com/rowanz/grover

[102] Connor Leahy, 'GPT2, Counting Consciousness and the Curious Hacker', Blog, *Towards Data Science* (blog), 6 June 2019, https://towardsdatascience.com/gpt2-counting-consciousness-and-the-curious-hacker-323c6639a3a8.

along encompasses, say, all computer science students across the world. Can the regime stretch to cover such people?

Connor assumed that OpenAI (which he had previously admired) was trying to restrict information out of corporate greed. He read it as a case of "a big entity that's hoarding information." Connor was influenced by hacker culture, and told me that he distrusted authority. He described himself as a "curious hacker", by which he means someone who is technically proficient, enjoys pushing the boundaries of what is technically possible, and intrigued by the off-limits.

He was also motivated by the mystique that OpenAI had created around GPT-2 with their partial non-release:

> I mean, OpenAI made GPT-2 a myth, they made a thing, something magical, something scary . . . The important thing was the mystique, the mystery, the social… How these things have been set up in the public consciousness.

Connor thought that this would have been a disincentive for replication among "well established people, people who are in the academic, in the industry community". But for him, it meant that replicating GPT-2 would be cool and attention-grabbing. "It was much *cooler* to do something scary, you know?" He attributed this to being a hacker type, where if "you make something off limits, you make it more enticing". The other main factor for Connor was that Google was giving him free access to cloud computing for his project (see below).

Connor put hundreds of hours into the project during his semester vacation. He used the code that OpenAI had released, which he said was "basically 80% of what you need": OpenAI

released the “structure of the actual model”, and you “just have to write the code around it to read the data, to train it,” etc. I asked Connor how many people around the world could have done the same thing under the same conditions. He said:

> I mean, seeing as I'm just a dude, I would say a lot of people. Going into this, I had basically never used TensorFlow, I had never used a transformer model before. I never used a TPU in my life before. And it took me like 200ish hours. And there's a lot of people that are a lot more capable than I am.

He thought that the majority of people who had studied AI to an undergraduate level could have done the same. He thought that the main bottleneck was accessing the necessary compute, rather than the technical difficulty. Even though Connor’s model turned out to be worse than Connor had thought (it wasn’t nearly as good as the 1.5B parameter GPT-2), Rowan told me that Connor had got nearly everything right.[103] Rowan retweeted Connor’s blog post, arguing that it reflected badly on OpenAI’s ‘staged release’ regime that even an undergraduate student could reproduce GPT-2.

Connor’s blog post blew up online. He thought the majority of people responding to him agreed with releasing the model, but not everyone. He was getting responses from not just the AI research community but all sorts of people, such as the cryptocurrency community. One negative comment came from “a Spanish motherhood website about pregnancies”. Clearly, the issue of GPT-2 had moved far beyond the boundaries of the AI research community.

[103] Rowan said: “Conner could do it probably, just given more money and more time. To me the problem that . . . Connor messed up on . . . is that he didn't give it enough data. And he probably didn't use some optimization parameters that were probably important. But I bet if you give him another month, if you gave him more money to download more news, just articles from the web, then you could have gotten way better results.”

In tandem with releasing the blog post, Connor emailed OpenAI. By the next day he was having a Skype call with Jack Clark (the head of the policy team at OpenAI) and a couple of the authors of the GPT-2 paper. Connor said that Jack's tone was understanding. "He just said: *hey, we just want to give you information to help you update your beliefs to help you understand the situation.*" The conversation did not fit with Connor's prior image of OpenAI as a behemoth, as he realised that the people at OpenAI were "just dudes, they're just like me", even if it was clear that Jack was more experienced and knowledgeable than Connor. Connor was happy that Jack had taken the time to "teach" him. Jack gave Connor a deeper level of detail about the potential misuses of GPT-2 than had been in the GPT-2 blog post – for example, with reference to the Russian disinformation campaign on social media during Russia's annexation of Crimea. This added information did make Connor "update [his] beliefs" about the harms of releasing GPT-2. But nevertheless, he still was not fully convinced, and went away from the conversation minded to release. The call demonstrates that OpenAI has some, but far from total, influence over people like Connor.

After talking to Jack, Connor had a call with Buck, a researcher at the Machine Intelligence Research Institute (MIRI) in Berkeley, California. MIRI does theoretical research about how to build a superintelligent AI system in a safe way. Connor was a big admirer of MIRI's work. Buck's argument to Connor was very different to Jack's. Buck's argument was that the GPT-2 model itself is not the important thing. Buck told me, "I started by saying that I didn't think that the direct effects of releasing the model would be very bad." He said: "the potential harm caused by something that makes it easier to generate babbling text seems pretty small compared to the risks associated with way more powerful AI systems in the future." Instead, Buck read the staged release of GPT-2 as an early attempt to improve the norms of the AI research community ahead of more dangerous, future AI systems being built. For people

worried about unsafe, superintelligent AI systems, a central concern is that AI labs will not have enough time to make their AI systems safe before deploying them. On this view, it is important that AI labs are not incentivised to rush their AI systems out of the door before they are confident that the AI system is safe. Buck's argument was that the staged release of GPT-2 was a step in the right direction from this perspective: setting norms in favour of caution over the release of AI systems.

Upon hearing this argument, Connor's view immediately did a 180 degree turn. He completely agreed with Buck. He worried that there was a small chance that the whole GPT-2 episode could become a case study about setting norms in the AI research community, and he did not want to be the "famous case" of the first person to go against the norm or to be "non-cooperative". He said:

> In a perfect world, what I would like is that we can then create norms in the AI research community, [where if] a researcher says, *Hey, guys, I'd made this thing but I'm kind of not sure how dangerous it is, I'd like to take my time,* then no one even thinks of making fun of them. . . . Everyone just accepts that: *Yeah, good on you. Take your time. If you need help, give me a call.* And that is currently not the default. . . . The default norm is still: publish everything or we will make fun of you. And in a way I was part of the problem.

Connor thought we still do not have a good solution to the problem, but continued: "I think the idea of just dumping onto the internet, as soon as it's ready, and making fun of people that don't do that is probably one of the worst systems that I can imagine. And I don't want to be implicitly supporting that by my actions, which I think I would have done." Therefore, Connor decided against open sourcing his model.

4.4 Eventual replication

In August 2019, two computer science graduate students at Brown University announced that they had replicated GPT-2, and fully shared the 1.5B parameter model online.[104] By this point, OpenAI had released their second-largest model (774m parameters) but not the 1.5B model. This was the first time that the weights of a model roughly equivalent in performance to the largest GPT-2 model could be easily downloaded online.[105] The students did not speak to the OpenAI policy team about their plans to release the model, although they had emailed some of the authors of the GPT-2 paper asking for technical advice on building the model (and got no reply).

The final nail in the coffin came in September 2019, when a research team at Salesforce open sourced a 1.6 billion parameter model. It was slightly larger than GPT-2 and, like Grover, it was built so as to better allow the user to specify the overall content of the generated text. After the original GPT-2 release, this was one of the main areas of weakness that AI researchers had pointed to with GPT-2, claiming that the model's uncontrollability would make it ill-suited to malicious use. With the Salesforce team's model, "CTRL", the user can specify, for example, that the text be written in the style of posts on the Reddit page dedicated to discussing conspiracy theories (r/conspiracy).

The model was uploaded to GitHub with a very permissive licence. Alongside the licence on the GitHub page, there is a statement: "but we also ask that users respect the following",

[104] Aaron Gokaslan and Vanya Cohen, 'OpenGPT-2: We Replicated GPT-2 Because You Can Too', *Medium* (blog), 22 August 2019, https://blog.usejournal.com/opengpt-2-we-replicated-gpt-2-because-you-can-too-45e34e6d36dc.

[105] The reported perplexity results (a measure of language modelling capability) across different datasets are generally slightly worse than the largest GPT-2 model, but better than the second largest.

giving a list of generic societal ills to be avoided.[106] As they write in the scientific paper, this "has no legal force".[107]

In the scientific paper and the accompanying blog post, much ink was spilled on the ethics of releasing CTRL. Several arguments for releasing the model were offered: (1) releasing the model is good for scientific reproducibility; (2) (à la Grover) releasing the model helps the "broader research community" to "combat the potential negative use cases" by providing good actors with "needed resources"[108]; (3) scientific "self-moratoriums" and "self-regulation" are not adequate to govern AI misuse.[109] In addition, the authors had spoken to staff at the Partnership on AI, a multi-stakeholder non-profit, to discuss the release. The researchers were defecting from OpenAI's staged release of GPT-2, but were keen to emphasise that they were doing so in an ethical way. Again, the researchers did not reach out to OpenAI's policy team before releasing CTRL.

In November 2019, OpenAI released the largest GPT-2 model.[110] CAMERON admitted that the prevalence of GPT-2-like models did contribute to the eventual decision to release. OpenAI also released a report reflecting on what they had learned, both in terms of the risks

[106] "This software should not be used to promote or profit from: violence, hate, and division, environmental destruction, abuse of human rights, or the destruction of people's physical and mental health." Nitish Shirish Keskar et al., *CTRL - A Conditional Transformer Language Model for Controllable Generation*, Python (2019; Salesforce), accessed 30 December 2021, https://github.com/salesforce/ctrl.

[107] Nitish Shirish Keskar et al., 'CTRL: A Conditional Transformer Language Model for Controllable Generation', *ArXiv:1909.05858 [Cs.CL]*, 2019.

[108] Richard Socher, 'Introducing a Conditional Transformer Language Model for Controllable Generation', *Einstein Blog* (blog), 2019, https://blog.einstein.ai/introducing-a-conditional-transformer-language-model-for-controllable-generation/.The post continues: 'With language models, it is critical that we promote awareness and understanding of these artificial generation processes. Similar to information security research, it is necessary for these tools to be accessible, so that researchers have the resources that expose and guard against potentially malicious use cases. We hope that research into detecting model-generated content of all kinds will be pushed forward by CTRL.'

[109] Keskar et al., 'CTRL', 11.

[110] Irene Solaiman, Jack Clark, and Miles Brundage, 'GPT-2: 1.5B Release', *OpenAI Blog* (blog), 5 November 2019, https://openai.com/blog/gpt-2-1-5b-release/.

of GPT-2 and in terms of how to conduct a 'staged release' regime.[111] The report struck a cautiously optimistic tone about the risks of GPT-2, finding that the concerns around machine-written disinformation on public platforms had not yet materialised.

4.5 Free compute for replicating GPT-2

All the GPT-2 replication projects discussed above have something in common: they were trained using Google's cloud computing service. This was not a coincidence. Companies selling cloud computing services have an incentive to have models like GPT-2 trained on their own hardware. The code for training a large model will be tailored to a specific type of hardware. Google has its own brand of hardware (TPUs), and the company seemed keen to have GPT-2 replications trained on Google TPUs. This would mean that anyone doing follow-up work will find it much easier to continue using Google's TPUs, rather than changing the code (e.g. to run on NVIDIA's GPUs).

Connor's project started when he was given beta access to Google's TensorFlow Research Cloud (TFRC), a Google programme that gives free access to compute to researchers and startups. He had originally not given any information about his project. During his project, Connor realised that the hardware was not going to be powerful enough: the TPUs, which were not the latest version, did not have enough memory capacity for training a 1.5B GPT-2 model. He told me that he had "basically given up at this point". Google's TFRC team emailed him asking for feedback on using the hardware. Connor gave detailed feedback, and they asked whether he was working on any blog posts or papers with the hardware. Connor told them that he was trying to make GPT-2, but unfortunately he would have needed a full TPU pod. "Lo and behold, they send me an email like, *oh we can get you one of those."*

[111] Irene Solaiman et al., 'Release Strategies and the Social Impacts of Language Models', *ArXiv:1908.09203 [Cs.CL]*, 2019.

Connor told me he received the email in a lecture and almost fell off his chair. After Connor's series of blog posts about his project, the TFRC team gave him additional access to compute.

Connor estimated that the computing power he used for his project was worth around $43,000. Connor's theory for why he was given this was that there is someone on the TFRC team who thought that this was "all really funny" and "just wants to see what happens". A more compelling explanation came from Aaron, one of the Brown graduate students who replicated GPT-2, also relying on compute from Google's TFRC team. Aaron told me that Google had a commercial incentive to have many papers published using large, GPT-2-like models, to showcase the fact that, for a window of time, it was much easier to train such models using Google's TPUs than anything else. This was until NVIDIA, another hardware company, published code for training GPT-2-like models on their machines.

The two Brown graduate students, Aaron and Vanya, also used free compute from TFRC for their project. They had initially applied for cloud computing resources through the TFRC website, but that had not worked. Instead, they found someone at their university who knew the head of the TFRC team. Aaron told me that they were happy to help, but emphasised that when the students released the model, they should do it quietly. They did not want a repeat of the Connor situation, which got too much attention, and he did not want the press to have the story, "*Google giving out lots of dangerous compute*." I asked Aaron what value the Google's TFRC team would have seen in the project:

> Aaron:
>
> To test the systems, and just like...They just want to get more papers out there on it that can only be done on TPUs, because if you're a company and you want to iterate on that for your own personal thing then you have to pay them to use TPUs. That's

basically it - that's basically the value in general. GPT-2 was really good marketing for their TPUs. You could only do it with their TPUs, initially at least.

Toby:

I didn't realise that. So why could you only do it with TPUs originally?

Aaron:

Well, you could do it with GPUs, but it's a lot more code to write. . . . Otherwise you can just use some really basic Tensorflow APIs, and just open your wallet, and it will just work.

This is because the code for training the model, especially for large models that are expensive to train, must be tailored to the specific hardware being used. This might help to explain, then, NVIDIA, a market leader in selling GPUs, also built their own GPT-2-like model, "Megatron", and open sourced the code.[112] Megatron was published in August 2019, and was even larger than GPT-2 (8.3 billion parameters). NVIDIA did not release the model itself, which CAMERON (at OpenAI) told me was to avoid negative PR in the wake of OpenAI's concerns around GPT-2. The paper and code is focussed on the problem of training such a large model, by splitting the training process across many different GPU machines that work in parallel.

As ROBIN (of OpenAI) told me, the incentive of a company like NVIDIA is to make it easy for people to train these large models on the machines that they sell; the incentive to release the model weights is not as strong. The Megatron release made it very easy for anyone to

[112] Mohammad Shoeybi et al., 'Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism', *ArXiv:1909.08053 [Cs]*, 2019, http://arxiv.org/abs/1909.08053.

train GPT-2-like models if they had access to enough GPUs; Aaron told me that with the Megatron code and enough money, a high school student could do it. At NVIDIA, AI research falls under the 'demand' side of NVIDIA's research arm, which according to their R&D chief 'tries to drive demand for Nvidia products by developing software systems and techniques that need GPUs to run well.'[113]

Overall, then, it seems that the TFRC team was hoping to benefit from students like Aaron and Vanya, and Connor too, as part of this competition with NVIDIA. Both Google and NVIDIA wanted to highlight the advantages of their own hardware, whilst also trying to avoid negative press where possible. Even the acknowledgements to the Grover paper thanks the head of the TFRC team and "the Google Cloud TPU team for help with the computing infrastructure" and "credits from Google Cloud";[114] and the Salesforce CTRL paper includes a similar reference.[115]

Aaron and Vanya's blog post was titled, *OpenGPT-2: We Replicated GPT-2 Because You Can Too*.[116] At this point, the Grover paper had been released, and Aaron and Vanya's post captures a similar sentiment. They present themselves as stress-testing OpenAI's staged release approach, with the conclusion that it is too easy for an attacker to replicate GPT-2. They saw this as undermining the staged release approach, because attackers get access to the largest model before anyone else. The project had used $500,000 worth of free compute, although Aaron said that most of it was wasted. They trained 10 different models, trying a slightly different approach each time (e.g. changing the dataset), and the best model was

[113] John Russell, 'Nvidia R&D Chief on How AI is Improving Chip Design', *HPC Wire*, 18 April 2022, https://www.hpcwire.com/2022/04/18/nvidia-rd-chief-on-how-ai-is-improving-chip-design/
[114] Zellers et al., 'Defending Against Neural Fake News', 9.
[115] Keskar et al., 'CTRL', 12. The paper thanks: 'Zak Stone and his team at Google for assistance with TPU infrastructure and code'.
[116] Gokaslan and Cohen, 'OpenGPT-2'.

actually the first. The blog post originally quoted the $500,000 figure. One comment on the post was: 'You mean "you can too" if you have a spare $500k laying around?' Aaron and Vanya soon removed the $500,000 figure from the post, because they thought it gave a misleading impression. Aaron told me that (a) they could have got access to the compute in other ways: for example, they had a friend working at a startup that had spare, free credits from Google; and (b) their project could have been executed much more frugally. He thought somebody could train a model for under $10,000 if they were willing to wait for over a month of training (and without experimenting across different training runs). He thought, then, that the "barrier to entry was realistically too low", making GPT-2 poorly suited to staged release.

This is a good example of how technical factors mediate the ease with which models can be governed. The size of GPT-2 helped slow down replication. It was a large model, by 2019 standards – and indeed, GPT-2 symbolised a move in the research community towards scaling up models (more parameters, more data, more compute) as a method for achieving greater performance. The scale of GPT-2 meant that students like Aaron and Connor still relied, at least to some extent, on the sponsorship of companies like Google. But it was not large enough to put GPT-2 out of reach for them, given that such sponsorship was forthcoming. This provides a useful point of comparison against GPT-3, which I will cover in chapter 2, which was an order of magnitude more computationally expensive to train.

4.6 Coordination between companies

A number of industry labs played along with OpenAI's staged release. One example was a company called Hugging Face. At the time, their product was a conversational AI system, and they did NLP research to support that. Shortly after the initial GPT-2 release, they trained their own GPT-2 model. Again, Hugging Face is publicly committed to supporting open source AI, and was very open with their research. Nonetheless, they wrote in a blog post in

May 2019: ‘We are aligning ourselves with OpenAI in not releasing a bigger model until they do.’[117] One Hugging Face interviewee told me that they did this because it would be unfair to OpenAI to release one of the largest models, given how much they had borrowed from GPT-2; and that they assumed OpenAI would release the model soon anyway. It also seems likely that OpenAI’s stance on the possible risks of GPT-2 meant that companies like Hugging Face would have wanted to avoid the controversy of going against the staged release. CAMERON (of OpenAI) told me:

> 10 years ago, it's very easy to be like: "AI thing, it's just sunshine and daisies, what could be better than a new AI thing?" And now that's a harder argument for many entities to make. And you see this. We've changed the press dynamics. Because we've just created a loud enough noise that went on long enough about bad shit to do with stuff, and we willingly talked about it, that that's entered the incentive structure.

NVIDIA did not release the weights for their Megatron model, which CAMERON told me was for PR reasons. An Israeli AI research company called AI21 trained a GPT-2-like model and bowed to OpenAI’s release schedule, even though they did not agree that the risks of GPT-2 warranted this.[118] This was out of ‘respect for our colleagues who have thought hard about these issues’.[119] Researchers at Facebook and Harvard University, in June 2019,

---

[117] Clément Delangue, ‘Ethical Analysis of the Open Sourcing of a State-of-the-Art Conversational AI’, Medium, 9 May 2019, https://medium.com/huggingface/ethical-analysis-of-the-open-sourcing-of-a-state-of-the-art-conversational-ai-852113c324b2.

[118] AI21 Labs, ‘HAIM: A Modest Step Towards Controllable Text Generation’, 2019, https://web.archive.org/web/20201111223535/https://www.ai21.com/haim-post.

[119] The post states: ‘We applaud the caution; fake news is a real issue which we as a community should take seriously, and we're glad it is taken seriously. We do feel compelled to mention that, in our view, the "fake news risk" is overemphasized relative to the benefits of facilitating better and more efficient writing. Furthermore, insofar as one is concerned with fake news, we’re not sure that suppressing text generation technology is either doable or the most relevant factor. All that said, we have much respect for our colleagues who have thought hard about these issues, and we are following their lead regarding release policy for now.’

published a paper on detecting machine-generated text in which they had trained a 1.4B parameter model; they did not open source this model.

All three of the largest corporate AI research labs (by publication count at major conferences) – Google, Microsoft, and Facebook – quietly decided against releasing various models on the grounds of risk. In January 2020, researchers at Google trained a 2.6 billion parameter language model on social media conversations, which could act as a chatbot.[120] They did not release the weights of the model.[121] Microsoft researchers also trained a similar chatbot on Reddit conversations, and withheld from the release a small amount of code necessary for generating outputs from the model.[122] (It is easy to obtain replacement versions of this code, and one of the authors links to these third party implementations on the GitHub page.) In early 2020, Microsoft also announced that they had trained a 17B parameter language model, TuringNLG.[123] They did not share the model weights, and CAMERON told me this was due to the concerns around the risks of text generation. With Facebook, in addition to the example I gave above, Facebook researchers also declined to open source the model for their superhuman poker playing system, Pluribus.[124] According to a media article, this was because

[120] Daniel Adiwardana et al., 'Towards a Human-like Open-Domain Chatbot', *ArXiv:2001.09977 [Cs, Stat]*, 2020, http://arxiv.org/abs/2001.09977.

[121] See Daniel Adiwardana and Thang Luong, 'Towards a Conversational Agent That Can Chat About…Anything', *Google AI Blog* (blog), 28 January 2020, http://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html.http://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html They write: 'Also, tackling safety and bias in the models is a key focus area for us, and given the challenges related to this, we are not currently releasing an external research demo. We are evaluating the risks and benefits associated with externalizing the model checkpoint, however, and may choose to make it available in the coming months to help advance research in this area.'

[122] Yizhe Zhang et al., 'DialoGPT: Large-Scale Generative Pre-Training for Conversational Response Generation', *ArXiv:1911.00536 [Cs]*, 2019, http://arxiv.org/abs/1911.00536. The paper states: 'DIALOGPT is released as a model only; the onus of decoder implementation resides with the user. Despite our efforts to minimize the amount of overtly offensive data prior to training, DIALOGPT retains the potential to generate output that may trigger offense. Output may reflect gender and other historical biases implicit in the data. Responses generated using this model may exhibit a propensity to express agreement with propositions that are unethical, biased or offensive (or the reverse, disagreeing with otherwise ethical statements).'

[123] 'Turing-NLG: A 17-Billion-Parameter Language Model by Microsoft', Blog, *Microsoft Research Blog* (blog), 13 February 2020, https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/.

[124] Noam Brown and Tuomas Sandholm, 'Superhuman AI for Multiplayer Poker', *Science* 365, no. 6456 (30 August 2019): 885, https://doi.org/10.1126/science.aay2400.

the system could allow online poker players to cheat.[125] These decisions were all taken quietly, and I was unable to obtain interviews with the authors of these different projects.

I would argue that the bureaucratic structure of these companies helped with enforcing the staged release regime, in two ways. First, although the industry labs are largely academic in their culture, the company still retains top-down, bureaucratic power over researchers. For example, papers go through an internal review process in these industry labs, and the company has the authority (even if rarely exercised) to block publication. This issue came to the fore in 2020, when Google's internal reviewers requested that one of their researchers, Timnit Gebru, hold off from publishing a paper on the risks of large language models.[126] The company claimed this was because it did not yet meet their academic standards, but plenty of onlookers were suspicious that it was trying to suppress discussion of the risks of large language models. Gebru resigned over the controversy. For many people, the incident highlighted a potential dark side to AI labs' top-down bureaucratic power. At the same time, it seems likely that the existence of these internal, pre-publication review processes also explains some of the examples we saw above, where labs refrained from open sourcing language models on grounds of risk. We saw how companies like Microsoft and NVIDIA were concerned about the potential PR fallout, and this made its way into how a number of GPT-2-like models were shared. CAMPBELL, a PhD student at a European university, explained why the industry labs, compared with university labs, have greater ability to set top-down norms:

[125] Will Knight, 'Facebook's New Poker-Playing AI Could Wreck the Online Poker Industry—so It's Not Being Released', *Technology Review*, 11 July 2019, https://www.technologyreview.com/2019/07/11/134224/facebooks-new-poker-playing-ai-could-wreck-the-online-poker-industryso-its-not-being/.

[126] Emily M. Bender et al., 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.

> [Industry labs] have the structure for imposing these norms, because they still have this more corporate structure. . . . They can say: "okay, I'm Google or OpenAI, I want to follow these norms". And then, because I have this corporate structure, I can say "I'm paying you $200,000 a year, this is the least you can do – follow these norms". So they definitely have an easier way of implementing these norms, if they decide on them. The rest of academia is very heterogeneous in the sense that you have people sitting at different institutions and different countries, and it's very hard to impose any norms.

Second, the industry labs employ people to work specifically on issues of AI governance, policy, and ethics. This gives the industry labs more organisational capacity to consider the possible risks of the technologies they build; and means that there is someone in the organisation who are more likely to be concerned about such issues. The most obvious example from the GPT-2 story is OpenAI's internal policy team, which played a leading role in the staged release of GPT-2. Moreover, from my conversation with CAMERON, it is apparent that he was in contact with similar teams at other labs – for example, Microsoft's Office of Responsible AI. Industry AI labs also employ lawyers and public relations staff, who would naturally be more anxious about releasing potentially dangerous technologies.

A similar illustration of coordination between industry labs is how OpenAI and labs synced up via the Partnership on AI soon after the GPT-2 release. Although PAI does not have the power to push through a particular policy that gets applied to these companies, it has the potential to act as a forum through which companies can voluntarily converge towards collective standards. Together with staff at OpenAI, staff at PAI organised a dinner in March 2019, bringing together different groups of AI researchers and corporate managers to discuss

the issue of malicious use and publication norms. The dinner took place in early March 2019, and there was a summary blog post from PAI titled: '*When is it appropriate to publish high-stakes research?*'[127] The discussion was about whether corporate labs should include in their pre-publication review process an analysis of the potential for malicious applications. The attendees did not reach agreement on this point. However, LENNON (working at PAI) told me that one of the attendees, who had a leadership role in one of the major industry labs, had recently designed their pre-publication review process. This person went away from the dinner keen to look into whether this process should be updated to include a check for malicious applications.

5. Distributing the responsibility to identify and address risks

The 'staged release' regime sets up a feedback loop between the lab and the outside world: the lab monitors how the model is being used in the world, which informs how the lab shares larger versions of the model. An important question is: what is the role of the AI research community in this process? Many AI researchers would naturally see themselves as central to the process of identifying the risks of an AI system: the research community would collectively study the system, as part of the normal scientific process, and uncover its capabilities, limitations, and pathologies. Does the 'staged release' regime facilitate this, or does it decentre the AI research community from the epistemic process of evaluating the risks of a model? In this section, I will examine the tension between (on the one hand) the 'staged release' regime adopted for GPT-2 and (on the other hand) the existing publication regime, which gives the research community a more central role.

[127] Claire Leibowicz, Steven Adler, and Peter Eckersley, 'When Is It Appropriate to Publish High-Stakes AI Research?', *Partnership on AI* (blog), 2 April 2019, https://www.partnershiponai.org/when-is-it-appropriate-to-publish-high-stakes-ai-research/.

As we saw above, CAMERON emphasised the value of 'having a load of people in the org who are just sort of prodding your technical artefact'. This process, of different OpenAI researchers playing with GPT-2, clearly informed the lab's internal assessment of GPT-2's potential risks. Was there any additional process for having the wider AI research community test GPT-2 and evaluate its risks? I would argue that the regime included only a limited capacity for facilitating such research, and this limited its acceptability within the scientific community.

In May 2019, OpenAI announced that they were forming "partnerships" with outside researchers to study the possible social impacts of GPT-2.[128] This was on OpenAI's terms: OpenAI staff had to approve the research, and only four partner research groups were mentioned in OpenAI's 6-month follow-up post.[129] There was one research group studying each of the following four topics: (1) "human susceptibility to digital disinformation"; (2) "how GPT-2 could be misused by terrorists and extremists online"; (3) biases in GPT-2's outputs; and (4) detection of GPT-2-written text. These researchers could access the larger versions of GPT-2 under a licence that prevented them from using it beyond the agreed project or sharing it with others.[130] These external researchers were not all drawn from the technical AI research community: it included social scientists too.

[128] An update (May 2019) to the original GPT-2 blog post included this information: "We are currently forming research partnerships with academic institutions, non-profits, and industry labs focused on increasing societal preparedness for large language models. In particular, we are sharing the 762M and 1.5B parameter versions of GPT-2 to facilitate research on language model output detection, language model bias analysis and mitigation, and analysis of misuse potential. In addition to observing the impacts of language models in the wild, engaging in dialogue with stakeholders, and conducting in-house analysis, these research partnerships will be a key input to our decision-making on larger models. See below for details on how to get involved."

[129] Jack Clark, Miles Brundage, and Irene Solaiman, 'GPT-2: 6-Month Follow-Up', *OpenAI Blog* (blog), 20 August 2019, https://openai.com/blog/gpt-2-6-month-follow-up/.

[130] OpenAI shared a draft contractual agreement in their 6-month follow-up post.

Therefore, the vast majority of AI researchers had to wait for OpenAI to release larger versions of GPT-2 publicly – they had the same access as anyone else in the world. Normally, "sharing" AI research means sharing it with the research community, and it just so happens that anyone with an internet connection can also then access it. In this case, the emphasis was flipped: it was a proliferation-centric regime, and the research community would get access whenever wider proliferation was deemed safe. On a very basic level, many researchers seemed to find this offensive – they were being deprioritised. This is evident from the social media reaction to the initial GPT-2 release, where some researchers, for example, complained that journalists were able to play with GPT-2 before the research community.[131]

More fundamentally, many researchers saw the issue of GPT-2's safety as a research question falling squarely within their jurisdiction. Especially in recent years, models have become sufficiently complicated that the exploration of a model's capabilities, limitations, and quirks cannot be comprehensively done in a single paper. Releasing the model to the research community allows for that exploration to be done collectively, in a distributed fashion. BELLAMY, a postdoc at a US university, told me:

> One argument in favour of more openness would be that if you don't make the model public, if you don't make it open, then understanding the model would be hindered. So people would be less clear about what the capabilities of the system are, what are potential good uses of this model? What are potential bad uses of this model? And so a better understanding of that will come about only if I can make it more widely available, so different people can try out different things to see, like what are potential risks of it.

[131] For example, one researcher tweeted 'Let's also stop and reflect on how reporters knew about this work longer than other scientists', and 'Making a big deal about withholding the model to the press before starting the conversation in academic circles rubs me the wrong way'.

In my interview with ALEXIS (a PhD student at UC Berkeley) he also emphasised the importance of this process of collection exploration:

> ALEXIS:
>
> The question I usually face is not like, *have I done something that could be used for evil*? It's more like, *have I even done something at all*? Right? [Something] that is at the frontier. And so, I guess just because of the salience of that question in my mind, I tend to be sceptical of other people's advances too. Unless I can probe it and test it myself and build on it. *Is this even a problem? Is this even worth paying attention to?* Is always a question in the back of my mind.
>
> Toby Shevlane:
>
> Yeah. And so then, in that case, open sourcing AI research is a way of the field collectively exploring capabilities and limitations and getting more knowledge or something like that?
>
> ALEXIS:
>
> Exactly. It just clarifies, like capabilities. Yeah, I'd put it that way. The only problem is: as you clarify capabilities you clarify if this [model] can be misused or not; and at that point, it's too late, if you've clarified that it *can* be. [Laughter]

This is an issue that risk-based KCRs must confront. How does the KCR delegate responsibility for answering the epistemic question of the technology's riskiness? (And how can this be squared with any concerns around proliferation?)

In answering this question, the KCR must also adjudicate between the jurisdictions of different expert communities.[132] As Hilgartner (2017) argues, KCRs do not just control the flow of knowledge objects. At the same time, KCRs position different actors. This issue comes out strongly with the KCRs that I study, which all must confront epistemic questions about risk, and therefore jurisdictional questions about who should assess that risk. Many AI researchers naturally saw their own expertise as central to the question of GPT-2's risks. In the social media reaction to GPT-2, a number of AI researchers argued that GPT-2's outputs were too unreliable to be useful to malicious actors. The Grover project (above) is another good example, where the authors argued that models like GPT-2 are useful for defending against misuse risks, because they can be fine-tuned to detect machine-written text. On the other hand, ROBIN told me:

> I think it's important to distinguish between, say, technical detection questions, which AI people can help a lot with, from like, "how will this be used in East Africa to topple governments" or something like that. Like no-one in the AI community has a clue how to think about that, but there are actual disinformation experts who do. So, we're talking to a range of experts.

This raises the question of whether the most important unknowns about GPT-2's risks were social in nature, i.e. about how social actors would use the model and what knock-on effects that could have, or more technical in nature. Scholars of other scientific fields have argued that scientists have a tendency to replace complicated social questions with technical questions, in a way that cements the scientists' control over how risks should be handled.[133]

---

[132] Hilgartner, *Reordering Life*, 7; See more generally: Andrew D. Abbott, *The System of Professions: An Essay on the Division of Expert Labor* (Chicago: University of Chicago Press, 1988).

[133] Hurlbut studies the 1975 Asilomar Conference on Recombinant DNA. He argues that "questions of whether and how technology serves the public good are rendered subsidiary to narrowly technical assessments of the potential for harm". He argues: "By confining problems to the lab setting and solutions to the competencies of

In the case of GPT-2, interests of the research community would favour a distributed system for assessing GPT-2's risks, where scientists were free to study the model. Ultimately, researchers want to be able to write papers about AI systems, especially those that seem cutting edge and widely talked about. This leads researchers to emphasise the technical side of GPT-2's risks, and the possible technical solutions.

This issue came to the fore with the Grover project, described above. The lead author of the Grover paper was critical of OpenAI for withholding the model from researchers, arguing that researchers needed to study GPT-2 in order to find defences against misuse. In his mind, the Grover project showcased the value that the research community could bring: their finding was that GPT models could be modified to act as effective detectors of GPT-generated fake news. After the Grover paper, this logic quickly became one of the main justifications for openness that my interviewees would rely upon.

SAMWELL (a researcher at a large academic lab) disagreed. He told me in an interview in 2021:

> I don't lean the nuke metaphor too hard, but like, "oh, if everyone had access to early nuclear technology, we would have great defences against nukes." It's like, no, we wouldn't; no, that's not how it works, right? And I think that that's likely to be the case with AI in general. People say, "Oh, you can tell maybe whether the text was from GPT-3," it's like, well, maybe sometimes, but certainly you need a certain amount; it's probably easy to work around these defences. I think that's motivated reasoning."

scientists, the Asilomar guidelines in effect excluded other forms of input, expert and nonexpert alike. Given that the objects of concern were circumscribed to the laboratory, input from other experts (e.g. ecologists) and even potentially affected publics was not seen as necessary." J. Benjamin Hurlbut, 'Remembering the Future: Science, Law, and the Legacy of Asilomar', in *Dreamscapes of Modernity*, ed. Sheila Jasanoff and Sang-Hyun Kim (University of Chicago Press, 2015), 130–31, https://doi.org/10.7208/chicago/9780226276663.003.0006.

As I have argued elsewhere, it is far from clear that openly sharing AI research will always make a bigger contribution to preparing defences against misuse than the contribution to the misuse itself.[134] Fighting AI misuse with better AI surveillance (which is the logic behind Grover) could run into a number of problems. For example, will everything written on the internet have to be fed through a Grover discriminator model? Will websites be happy to pay for this, and will users accept it? There is also the possibility that eventually AI systems will be able to mimic human text so closely that other AI systems are unable to tell the difference. In our interview, Rowan accepted that discriminators were not necessarily the "be-all and end-all", because we also need platforms to implement them, and we also need a public that actively wants to tell apart real and fake news. But other than this concession, Rowan (and plenty of other AI researchers I have spoken to) was keen to emphasise the power of researchers studying potentially dangerous models. Researcher access to models was seen as the absolute priority (even when the story of how this would prevent misuse seemed underdeveloped).[135] If granting research access conflicts with the other priority of stemming wider proliferation – and during the GPT-2 episode the assumption was that these priorities *were* in tension – then the claim is that research access should win out.

---

[134] Toby Shevlane and Allan Dafoe, 'The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?', in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20 (New York, NY, USA: Association for Computing Machinery, 2020), 173–79, https://doi.org/10.1145/3375627.3375815.

[135] For example, I presented Rowan with a hypothetical case where we release to the public 1,000 machine guns and 1,000 bulletproof vests, and asked why this would make the world more dangerous whereas releasing the Grover generator and discriminator models would make the world safer. He replied:

> Yeah, I totally agree. [...] And I'm not gonna advocate giving everyone machine guns obviously. I think that maybe where that analogy breaks down to me, is that what we should do as people who are interested in being safe against these things, is we should open up machine guns to researchers. Right? Like if it was a world where everywhere like every other person has a machine gun then probably you need fundamental research in how to make better defences against it, right?

The assumption is that there is something technical inside the gun, or inside the AI system, that hints at possible defences. As the example of machine guns demonstrates, this will not always be the case.

The Grover project is a good example of the research community's endless drive to produce more knowledge and artefacts. Naturally, the community's preferred answer to potentially risky research is not to suppress research, but to make more research – that is what the field is already set up to do. OpenAI's claim that large language models pose a risk to society was quickly turned into a research question, and one that could be answered using the community's existing methods: building and evaluating models. As one of my other interviewees said, Grover was a smart move from an academic perspective, because it made for a good paper. The GPT-2 blog post involved existential reflection, lamenting that it would not be "possible to control research in [generating text, images, and video] without slowing down the progress of AI as a whole". The discourse around Grover circumvents these searching questions. Instead, the solution to problems created by AI research is *more AI research*.[136] The concerns about proliferation become relegated, because the downsides of proliferation are seen as wiped out by further research effort.

Although one might be sceptical about this position, I would still argue that the GPT-2 'staged release' regime was made weaker by its low capacity for delegating risk-related research to the broader community. There are three reasons for this. First, many of the researchers who replicated GPT-2 did so because they did not have access to the model themselves. This is what happened with the Grover project, for example. Rowan (the lead author) recalled that OpenAI only gave him and his colleagues very restricted access to GPT-2 through OpenAI's "quote, unquote, partnership programme". At least at that point in time, OpenAI was offering: not the largest model (instead, the second or third largest); not the weights of the model (just

[136] Ironically, this was the same claim that the founders of OpenAI used to justify setting up a new AI lab to rival others (like DeepMind). When OpenAI was founded in 2015, Elon Musk said in an interview: "I think the best defense against the misuse of AI is to empower as many people as possible to have AI. If everyone has AI powers, then there's not any one person or a small set of individuals who can have AI superpower." Steven Levy, 'How Elon Musk and Y Combinator Plan to Stop Computers From Taking Over', Blog, *Backchannel* (blog), 11 December 2015, https://medium.com/backchannel/how-elon-musk-and-y-combinator-plan-to-stop-computers-from-taking-over-17e0e27dd02a#.fto7v7a1r.

access to a "demo", where the user interacts with the model at arm's length); and without being able to change the algorithm by which GPT-2's outputs were sampled.[137] [138] Rowan told me:

> . . . once the people at UW [who] were applying for [GPT-2 access] realised, *oh, wait a second, we're not going to be able to do research on this*, then we independently had these ideas about how to actually study this problem of neural disinformation and what adversaries might do, and we kind of just did everything ourselves.

The graduate students from Brown who replicated and fully open sourced their GPT-2 model also originally came into the project wanting to study GPT-2, but did not get access to the model from OpenAI. They were additionally motivated by the idea of having an open version of GPT-2 against which the research community could compare the performance of subsequent models. Therefore, there is a direct connection between the limited researcher access to GPT-2 and the fact that it was replicated and open sourced by other researchers – which (as we saw above) partially undermined the 'staged release' regime.

Second, the AI research community is undoubtedly well-placed to study the capabilities and pathologies of AI systems. Even if this is not a panacea, it will often be relevant to addressing risks from AI. In the case of GPT-2, OpenAI staff seemed to think they had a better

[137] On sampling: these language models do not just output the next word, but a probability distribution over many different possible next words. To use them to generate text, you sample from this distribution. OpenAI was using a very simple way of doing so: "Top-K" sampling, where you only sample from the top, say, 100 most likely next words. It is easy for a discriminator to detect text that is generated in this way, because the distribution looks very unnatural – all the words in the article are quite predictable from the previous context, with none beyond some cut-off point of unpredictability. The team wanted to use a sampling method that would be harder to detect.

[138] OpenAI's final report on GPT-2 gave some detail on what level of access to GPT-2 the external researchers got. It reads: 'When forming partnerships, we signed a non-commercial legal agreement with a partner organization to provide our model for their research use, and/or we provided a partner organization with a secure sampling interface to the larger models.' Solaiman et al., 'Release Strategies and the Social Impacts of Language Models', 3.

understanding of the model's text generation capabilities than the likely social impacts of those capabilities. In other cases, though, the lab might be uncertain about a model's capabilities, including – for example – situations where the model might respond in a surprising and unwanted way. There is much more research capacity for the study of models outside any single lab than inside it.

Third, even where the lab's biggest unknowns are 'social' in nature, again, research into these questions might sometimes benefit from external social scientists having access to the model. OpenAI tried to achieve this by coordinating with a small handful of external social scientists, but again, the lab has a limited capacity for coordinating this work, compared to the vast number of social scientists out there in the world. As we will see in Chapter 2, the KCR used for GPT-3 (the successor to GPT-2) more effortlessly brought in outside researchers to study the model.

In the absence of large-scale delegation, the in-house researchers at OpenAI had their work cut out for them. CAMERON told me that in April 2019 one of his colleagues in the policy team "successfully freaked me out" by pointing out that "we are nowhere near where we need to be in terms of having a substantive amount of research into this to let us make good decisions." Getting into a more solid epistemic position to make decisions about when the larger models should be released required a lot of hard work from multiple researchers on the policy team.

6. The relationship between science and society

The GPT-2 release positioned the relationship between science and society in a way that was unusual for the AI research field. From the GPT-2 blog post, and looking at the 'staged release' regime, the following vision of the science-society relationship emerges. (1)

Scientific research can have a direct, negative impact upon society. Society, like an ecosystem, is vulnerable. (2) Scientists can manage the impact of their work on society, by choosing the timing and content of the research they release. (3) Scientists should warn society about upcoming risks from scientific research. (4) Scientists should monitor how their research is used in society, and use that to inform decisions about what to release.

In other words, the new KCR presented science and society as interconnected. It presented the scientific community as capable of *acting*, as an agent, in order to steer the impact of its creations on society. It presented scientists as responsible for playing this role.

This overall vision of the science-society relationship was controversial within the AI research community. In this section I will go through four visions of the science-society relationship that conflicted with OpenAI's 'staged release' regime, which were used by AI researchers to attack OpenAI's decision-making. I rely on my interview data, the social media reaction to GPT-2, and the discourse that came with Salesforce's release of their CTRL model (which, as we saw above, undercut the staged release of GPT-2).

Vision 1: Scientists as knowledge-producers, society as the beneficiary

A recurring theme in my interviews was the special role of scientists as knowledge-producers. A common view goes as follows: *The function of the scientific community is to produce knowledge. AI research outputs, like trained models, count as knowledge. Knowledge is usually, or even always, beneficial for society.* The same discourse has been observed in other fields, including genomics.[139] A good illustration comes from my conversation with

[139] Hurlbut, 'Remembering the Future'; Stephen Hilgartner, Barbara Prainsack, and J. Benjamin Hurlbut, 'Ethics as Governance in Genomics and Beyond', in *The Handbook of Science and Technology Studies*, ed. Ulrike Felt et al., Fourth, Ebook Central (Cambridge, Massachusetts: The MIT Press, 2017), 823–51. See also Gieryn's description of a 1982 report from the NAS Panel on Scientific Communication and National Security: 'The Report isolates a "core" of science by demarcating the production of scientific knowledge from its

HEYDAN, the most senior researcher at a large industry lab, who I presented with a hypothetical scenario:

> Toby Shevlane:
>
> Say if you came up with a model, and trained it, and for whatever reason, you thought, this is harmful. …[T]his is actually something that's going to be net negative for the world. What do you think you would do in that situation?
>
> HEYDAN:
>
> I mean, I can't think of a scenario like that. Because even if...You know, I hate bringing up nuclear technology, because AI is nothing like nuclear technology, right? But look at Albert Einstein, or Richard Feynman, all these people who are pioneers there, they still feel that it was net good for the world,[140] because knowledge is in the end, I think, a net positive aspect. . . .
>
> Toby Shevlane:
>
> Okay. I see. And so is part of the argument that technology is always beneficial on net, in some way?
>
> HEYDAN:

consumption. Selected characteristics are attributed to science in order to distinguish from technological applications: . . . the goal of science is the creation, dissemination and evaluation of knowledge as its own end, not as a means for material production; open scientific communication transmits theoretical and empirical knowledge about nature, not "know-how" or "recipes" immediately transferable to production of hardware'. Thomas F. Gieryn, 'Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists', *American Sociological Review* 48, no. 6 (1983): 790, https://doi.org/10.2307/2095325.

[140] This claim about Einstein and Feynman appears incorrect. Einstein did not work on the Manhattan Project, although he did sign a letter to President Truman recommending that the US should build nuclear weapons. Einstein later described this as the "one great mistake in my life". Feynman fell into a deep depression after Hiroshima and Nagasaki. He is quoted as saying: "Maybe from just the bomb itself and maybe for some other psychological reasons I had just lost my wife, I was really in a depressive condition."

Knowledge. Knowledge, I would say.

It might seem strange to describe AI systems as simply "knowledge", with no mention of technical artefacts. One response is that the quest of AI research is to uncover the nature of intelligence; AI researchers seek understanding of intelligence by building it. As one AI researcher put it (quoting Richard Feynman), "what I cannot create, I do not understand".[141] However, even if the ability to create something is a necessary condition for understanding, it is not sufficient. AI researchers generally admit that they have little clue what is going on inside their models. As BELLAMY put it, "usually, it's just gonna be like this huge array of numbers that's not interpretable".[142] Understanding what algorithms are instantiated in this huge array of numbers is a subfield within AI research, but not a very large one.[143] At the same time, these models often have many practical applications. There is a strong claim, then, that models should be classed as useable technical artefacts, even if they simultaneously carry epistemic value.

[141] One prominent AI researcher said the following on a podcast: "There is this wonderful quote from Richard Feynman, which is: "What I cannot create, I do not understand". And I feel like you understand best by building, and so I wanted to switch from just talking about intelligence to trying to build it as a way to better understand it."

[142] With deep learning, *building* a neural network and *understanding* a neural network are quite separate activities, because the features learnt by the network are not hand-crafted by the researcher, but emerge through the training process. In an influential blog post, Andrej Karpathy, an AI researcher, argues that deep learning is an example of "Software 2.0", which is a less hands-on process for building software. He writes that in comparison to traditional software, "Software 2.0 is written in much more abstract, human unfriendly language, such as the weights of a neural network. No human is involved in writing this code because there are a lot of weights (typical networks might have millions), and coding directly in weights is kind of hard (I tried)." He continues: "Instead, our approach is to specify some goal on the behavior of a desirable program (e.g., "satisfy a dataset of input output pairs of examples", or "win a game of Go"), write a rough skeleton of the code (i.e. a neural net architecture) that identifies a subset of program space to search, and use the computational resources at our disposal to search this space for a program that works. In the case of neural networks, we restrict the search to a continuous subset of the program space where the search process can be made (somewhat surprisingly) efficient with backpropagation and stochastic gradient descent." Andrej Karpathy, 'Software 2.0', *Andrej Karpathy* (blog), 11 November 2017, https://karpathy.medium.com/software-2-0-a64152b37c35.

[143] From an analysis of the 1449 papers submitted to ICLR 2019, around 10 were deemed to be on the topic of 'interpretability'. Sergey Ivanov, 'ICLR 2019: Stats, Trends, and Best Papers.', Criteo AI Lab, 5 December 2018, https://ailab.criteo.com/iclr-2019-stats-trends-and-best-papers/.

A similar example comes from my interview with JESSIE, an AI researcher at Google. We were talking about the use of AI systems for tracking people, which they brought up as an example of a potentially harmful AI application. JESSIE was arguing that China is going to work on this regardless, so there's nothing Western researchers can do – they might as well continue working in surveillance-relevant fields. They said it just comes down to how much investment a country like China puts into AI, and China is putting a lot of money into AI research. In passing, they added:

> And I mean, actually, maybe, if you force these governments to go down that route, that will be good, because it educates the people and education is usually good and usually destructive to totalitarianism.

Again, this comes down to the question: what are AI research outputs? No doubt, certain types of knowledge could be a thorn in the side of totalitarian governments – perhaps knowledge of government cover-ups, or knowledge that refutes state propaganda. But does the category of "knowledge" include all technology, too? It is much less clear that all technology has a counter-totalitarian bias – even technologies that are directly useful for propaganda, surveillance, censorship, imprisonment, and so on. With GPT-2, part of the concern was that authoritarian governments could use this kind of model for disinformation and propaganda. And more generally, a large proportion of AI research is surveillance-relevant (including language models like GPT-2, which could be used for the surveillance of text). I have asked a few of my interviewees (all AI researchers): what proportion of AI research do you think could, even if remotely, be relevant for surveillance? These are the three answers I got:

> "Yeah, in principle, all of it, probably. [Laughter]"

> "I'm struggling to think of something that could not be used for surveillance. […] So if I had to put a number on it, probably over 50%."
>
> "It's part and parcel, right?"

Conceiving of AI models purely as "knowledge" obscures these kinds of applications. A final example comes from Salesforce's release of the CTRL model. Their blog post claimed that, in releasing the CTRL model, the lab aimed to "foster transparency".[144] What kind of transparency is this – what is the object of transparency? Sharing the model weights does very little to shed light on the activities or decision-making of Salesforce as a company. Ironically, the company's decision-making process for releasing CTRL was not particularly transparent (see below). Again, we have the ontological question: what is an AI model – is it knowledge, or a piece of machinery? If we see models as machinery, then being "transparent" about the model weights is the same thing as widely proliferating the machinery. The unsettled ontology of AI models helps the lab to obscure this. The model is presented as a window through which something can be seen, instead of a tool that can be used to achieve particular effects in the world.

Vision 2: Science as powerless; society as powerful in determining the technology's impact

Most of my interviewees who criticised 'staged release' were not content with balancing the whole argument on a fragile distinction between knowledge and technology. The conversation often shifted to the comparative advantage of scientists and governments to tackle risks. The argument is that governments are better equipped to deal with problems like

[144] Socher, Controllable Generation.

misuse of AI. Often this is a question of power: governments have more power over individuals and companies who might misuse AI. And sometimes it is about expertise: AI researchers are not experts in societal issues, but – as experts in the technology – they have a responsibility to educate governments and the public. For example, SAM (a PhD student at a US university) told me:

> I'm not sure it's your responsibility, as the person who made it, to figure out the solution to it. I think it's all of us, as a society, who has the responsibility.
>
> I think the best thing that we can do is recognise the risks that we create and communicate them effectively. Because we're experts in the technology; we're not experts in economics or regulation.[145]

FRANKIE, a US researcher, emphasised the research community's lack of power to address misuse:

> What's the [solution] for models being used in ways that we don't like, whatever that way is? It's to stop the use. That's not something that the research community can possibly address. That's something that's external, like that's where the thing gets used. That's not something that researchers choose or do anything about. It's a policy problem.

[145] Similarly, JESSIE told me: 'So personally, I think that the right way to deal with big societal changes, or new technologies is regulation and legislation. [...] And you can still use them but you run the risk of breaking the law. But for politicians or the legislature to like to do the right things, they need to be educated and they need information. [...] Yeah, and that's I think the biggest responsibility of people who are doing the research is to educate other people, like lawmakers about the benefits and pitfalls so that they can make the right decision.'

Similarly, NOEL is a professor at a US university and a researcher at a large tech company. They were critical of OpenAI's decision to withhold the larger GPT-2 models. I asked if they thought that GPT-2 could be used for harm. They said that you could find a way to use it to harm someone, but that this is the case for all technologies. They said that even a pen can be used to stab someone in the eye. I asked if some technologies are more weighted towards harm, for example, a bomb being better-suited to misuse than a pen. They said no, because the impact of the pen will depend upon societal variables such as literacy rates. They continued:

> So, it's actually largely about the social agreement and the context that we need to bring into account. . . . Now, if and when these so-called AI technologies start to be deployed in society on a much wider and much deeper level, what kind of society would we want to have, so that these technologies are going to be more harmless than harmful?

I asked if they thought it would be a good thing if human-level AI were developed. They clarified that "we have to always strive to exceed human intelligence", and continued:

> But is it going to be good? I think so. Not because that kind of technology is inherently assistive or helpful for us. But because I believe we will be able to change or drive our society so that we can benefit from these kinds of technologies.

This position is the opposite of technological determinism: it holds that social forces wholly determine the impact of a technology, regardless of the technology's inherent properties. It is a certain kind of *social* determinism in which society is imagined as a coherent agent,

capable of imposing its values on any technological shift.[146] Interestingly, while technological determinism is sometimes accused of absolving technology-makers of responsibility, in this case the reverse is true. It no longer matters what the technologists do, because all the action happens outside of their domain. This kind of social determinism is therefore a convenient way for AI researchers to disown the responsibility to address AI risks.[147]

In contrast to governments and social forces, the research community is presented as powerless to make a difference. ELLIOT, A PhD student at a large Canadian lab, told me:

> I mean, [GPT-2] could be harmful. But then again most research is potentially harmful. I mean most research we do, like . . . A lot of computer vision is like state surveillance. It's already happening. . . . Massive state surveillance or autonomous weapons. And speaking for myself, I know that everything I do is going to be pretty harmful.

[146] CHARLIE, a PhD student at a large North American lab, disagreed strongly with this position. He told me: "I think the view that technology is a tool and humans decide what to do with the tool is like, I don't even think it's coherent. Because I think it places agency in either individual people or humanity as a collective in a way that I don't think makes sense. I think the reality is that we're part of a very complicated socio-technological system. And that system includes a lot of complicated incentive structures, a lot of complicated organisations, composed of many humans that can be modelled as agents, but like, maybe sometimes should and sometimes shouldn't be. A lot of what happened is driven by forces that are in some sense beyond our control. Like there's nobody running the world economy, but the economic forces have this massive influence over what happens, and they create really strong incentives that we're all heavily influenced by. [...] I'm not sure I can even like pass the ideological Turing Test or something of somebody who doesn't understand that, and I'm not sure what it would mean to really expand upon and fully articulate this view that technology is just a tool and we decided how to use it. Like, I don't actually know what people are saying; I feel like it's just constructing this very simple model that if you actually try and map it onto reality, you'll just notice that it doesn't, there's no way of doing it."

[147] See also Stephen Hilgartner, 'The Social Construction of Risk Objects: Or, How to Pry Open Networks of Risk', in *Organizations, Uncertainties, and Risk*, ed. James F. Short and Lee Clarke (Boulder, CO: Westview Press, 1992), 39–53.

> . . . And why am I still doing it? . . . I am not that extraordinarily smart or important that somebody else can't replace me. I'm not irreplaceable. If I don't do it, someone else would.

In the reaction to GPT-2, this kind of discourse was often concentrated on the issue of replication. Critics of OpenAI argued that independent replication of GPT-2 was so easy that "staged release" would be completely undermined. For example, an assistant professor at NYU was quoted in the media: "I'm confident that a single person working alone with enough compute resources could reproduce these results within a month or two". A professor at a large US university tweeted: "I see no natural moats for this tech, and if @OpenAI can do it, so can others around the world. The Pandora's Box is open, and we have to learn to live with #AI fake reality.."[148] This view allows researchers to accept the risks of AI, and accept that AI research directly contributes to those risks. Instead, the issue becomes a very practical one, about the uncoordinated nature of the AI field.

A good illustration of this discourse comes from Salesforce's release of their CTRL model. In tandem with the CTRL paper, a subset of the same researchers uploaded a preprint about AI governance.[149] The paper is written as a Science and Technology Studies (STS) paper, and the lead author minored in STS during his graduate studies. The main argument of the

[148] I had interviewees who made the same argument. CAMPBELL, a PhD student in Europe, told me: 'In general… you can't really hold technological progress back...It's the same with the atomic bomb and nuclear energy. So usually, it's more like we're providing tools, and what people do with these tools is very much up to them. ... Just because it's possible to misuse certain tools does not mean we should forbid them or prevent them from being published at all." SAM said: "I think that the solution is not going to be any kind of, like, constraint on the technology, because people will just go around that. So I don't think we should be looking for solutions like "you can't make facial recognition systems that are this good" or "you can't make language generation that's this good." I think that's going to fail. I feel like to get in front of it, what we really need to do is reach out to the broader population and make them aware of how good video fakes and language fakes are. Maybe a weak analogy is nuclear weapons. We all know that they are a terrible idea but we keep producing them…'

[149] Lav R. Varshney, Nitish Shirish Keskar, and Richard Socher, 'Pretrained AI Models: Performativity, Mobility, and Change' (arXiv:1909.03290 [cs.CY], 2019).

paper is that the researchers who train a large language model should not be the focal point for governing the model, because the social impact will depend upon dynamics external to the lab, occurring subsequent to release.

The authors (Varshney et al) correctly highlight an important aspect of the challenge of governing models like GPT-2. These large, pretrained models are often adapted by subsequent users. Two important types of adaptation are: (1) fine-tuning, i.e. the model goes through a further training phase, more specific to a particular application. For example, GPT-2 could be fine-tuned on a reading comprehension dataset. (2) The models can be combined with other software, including other AI systems. I would add an example from 2021: OpenAI's CLIP model, which was pretrained to predict the captions of images scraped from the internet.[150] In doing so, the model learns generic skills for understanding images, and can be applied to specific tasks. Like GPT-2, CLIP was put through the staged release regime. When I asked ROBIN why the staged release regime was used for CLIP, they said:

> Basically the same reasons as GPT-2, which is to say that we're uncertain about the ways in which it could potentially be misused. And in fact, we've learned some things as a result of releasing it in a staged manner that we didn't necessarily know, prior. We flagged in the paper that we were concerned about things like surveillance and stuff like that, but we weren't necessarily thinking of other uses that have subsequently come more into prominence. Like using it to steer generative models was not something that was as high on our radar as it is now. . . . And that was not originally how we were thinking about it, and that in turn now is informing our thinking. . .

[150] Alec Radford et al., 'Learning Transferable Visual Models From Natural Language Supervision' (arXiv:2103.00020 [cs.CV], 2021).

ROBIN was referring to the fact that CLIP, which was initially thought of as a model for classifying images, has been used to *generate* images. CLIP is paired with a model that generates images, and iteratively steers the generated image towards some target description (such as "a painting of Barack Obama, by van Gogh"). There are popular, open source tools that make this process very easy. It is a good example of how the possible applications of a large, pretrained model will depend on how it is adapted.

Varshney et al argue, therefore, that the producers of models are part of a broader ecosystem, and it is that ecosystem that determines how the model will be used. They connect this point to STS literature on how the users of technology are part of the process of innovation. Varshney et al turn this into a theory of responsibility on the producers of large models. The producers should be decentred, essentially because the important action happens elsewhere. There is a call to expand governance to the broader ecosystem of actors ("responsible innovation should be expanded to include the role of users"), although the paper does not explain how that should be done. If you read the paper alongside the CTRL release, the conclusion seems to be that labs do not have an obligation to refrain from sharing their work on grounds of risk. Indeed, Varshney et al also argue that scientific "self-moratoriums" do not work, because scientists "cannot be expected to step outside the momentum of their own work to self-regulate".

As such, the paper reproduces, in more academic language, the discourse we saw above about why AI research labs should not be expected to limit publication of their work. On the one hand, it is about the location at which risks from AI arise, and accordingly, where governance should be located. The claim is that AI risk arises external to the lab's actions, in the broader community and society. Theories about the social shaping of technology are wheeled out in support – for example, Varshney et al emphasise that 'the social world acts to fundamentally

shape technical development at every level'. This is the same kind of social determinism we saw above, and it is again used in service of insulating scientific researchers from the demands of governance. And on the other hand, the Salesforce researchers are making an argument about the governance *capacity* of researchers. Researchers are presented as being incapable of restraining themselves. This is then used, in an arguably circular way, to justify the release of the CTRL model.

The paper provides an interesting contrast with my conversations with members of the OpenAI policy team. In a way, the researchers from Salesforce and OpenAI have a similar starting point. Consider this line, for example, taken from the Varshney et al paper, but which looks like it came straight out of my transcripts from my conversations with ROBIN:

> Our estimates suggest that the cost to train . . . XLNet was $50,000, to train RoBERTa was $60,000, and to train GPT-2 was $250,000. On the other hand, the cost to fine-tune BERT on the SQuAD dataset is estimated to cost only $3.

In the Varshney et al paper, this is evidence of how easily these large models evolve subsequent to release, thus escaping the reach of the labs that produce them. ROBIN takes this same starting point but goes in a different direction. If you spend millions of dollars training a large model and then just upload the weights for anyone to download, then you have given away your only opportunity to shape how the model is used. The high training costs narrows the number of actors that can build the model themselves, which puts the lab in a stronger position to influence how the model is deployed. It is precisely because the costs of training models is so high – relative to the costs of fine-tuning or running a model – that elevates the lab that produced the model as an important actor in the governance process.

Vision 3: Society as a threat to science, requiring management of the public perception of AI

From the "staged release" perspective, a key reason for delaying release is to give the world advanced warning of the risk. Hence, in the GPT-2 blog, OpenAI was not just addressing AI researchers but a broader audience. For example, they argued that governments should start efforts to 'more systematically monitor the societal impact and discussion of AI technologies'.[151] The logic is that AI misuse is not something that can be solved quietly within the AI research community, but requires engaging other actors. I asked ROBIN, given a key aim behind the GPT-2 release was to 'start a debate', who needed to be part of that debate – AI researchers or the public? They said that AI researchers are a 'key part of the solution to the publishing question', but 'it is not clear that the AI community is actually the people that need to solve' the overall problem of misuse. In other words, compliance from AI researchers would be a necessary but not sufficient component of the governance regime's success.

However, in raising the alarm to a general audience, OpenAI was accused of polluting a common-pool resource: the public's perception of AI research. Onlookers presumed that OpenAI was doing this for selfish reasons: to raise corporate investment (and in 2019 OpenAI was transitioning away from being a non-profit and raising capital). The concern was that the public would perceive AI research as moving at an exciting or scary pace. The motivation for avoiding too much public excitement is often that, if AI progress does not live up to the hype, there will be a backlash against AI, leading to an "AI winter", where funding dries

[151] Radford et al, Better Language Models and Their Implications.

up.[152] The motivation for avoiding public fears is rarely spelled out so explicitly, but might also have something to do with funding.[153]

This means that many AI researchers are suspicious of journalists. They prefer journalists not to write alarmist news stories about AI. OpenAI had briefed a handful of journalists in advance of the GPT-2 release, and some AI researchers reacted with hostility to this. For example, here are three different tweets from AI researchers:

> You are using media to hype up your language model. There are so many others doing research on same topic. You claim yours is far better by giving limited access to only journalists
>
> —
>
> "AI is scary" is a narrative that reporters cannot pass up. Spoon fed to reporters this time. [...] Making a big deal about withholding the model to the press before starting the conversation in academic circles rubs me the wrong way
>
> —
>
> My overall stance is that OpenAI made the right decision not to release the larger trained model but communicated the reason why this decision was made the wrong way and to the wrong audience (the media).

[152] One AI researcher tweeted: "Every time you overstate the capabilities of an AI system or the speed of AI progress, you are doing the public trust equivalent of taking on credit card debt. Which at some point the industry will have to repay…" Another AI researcher, in 2020 (i.e. not specifically about GPT-2), blogged: "To begin with, something about crying wolf. If we (AI researchers) keep bringing up the specter of Strong AI or Artificial General Intelligence every time we have a new breakthrough, people will just stop taking us seriously."

[153] For example, one AI researcher, responding to the suggestion that work on AI lip-reading could be used for surveillance, complained on Twitter: "Researchers working on lipreading to help speech impaired patients will struggle to raise funds."

More generally, there are many examples of AI researchers policing journalists' language on Twitter. Journalists often get criticised for describing AI systems as agents, for example writing that "an AI did…", or comparing the learning process to how human children learn.[154] This might seem hypocritical, because AI researchers themselves often describe AI systems as agents, and aspire towards building more human-like learning processes. But they would say that the public is prone to misinterpreting this kind of language. Van Lente found similar behaviour among the early pioneers of electricity; he writes: "the experts gave themselves the task to teach the public the 'proper' promises."[155] TYLER, a PhD student at the University of Cambridge, told me that there are two conversations about AI risk and the future of AI: one when AI researchers are quietly talking to one another, and another when it is public-facing.

This is important context for interpreting the reaction to GPT-2. OpenAI was departing from the public relations strategy informally adopted by most AI researchers, which is to downplay the pace of AI progress. AI researchers do (as we saw above) sometimes see themselves as having a duty to educate the public, but they normally have a certain kind of 'education' in mind. For example, it is very uncontroversial, among the AI research community, to say that the public should be educated to be mistrustful of current AI systems, which will often fail in unexpected ways. It is much more controversial to tell the public that the pace of progress in AI capabilities is a source of risk.

[154] Andrey Kurenkov, 'AI Coverage Best Practices, According to AI Researchers', *SKYNET TODAY* (blog), 11 November 2019, https://www.skynettoday.com/editorials/ai-coverage-best-practices.
[155] Harro Van Lente, 'Forceful Futures: From Promise to Requirement', in *Contested Futures: A Sociology of Prospective Techno-Science*, ed. Nik Brown and Brian Rappert (Ashgate Burlington, VT, 2000), 43–64.

Vision 4: Society as a decision-maker within the KCR

Finally, one vision of the science-society relationship is that society should be somehow brought into decision-making within the KCR.

The Salesforce researchers consulted with the Partnership on AI (PAI) before releasing the CTRL model, and described this in their paper as follows:

> Rather than self-governance, we sought to diversify inputs to governance through pre-release review from experts at the Partnership on AI (PAI). These experts, in turn, drew on emerging norms and governance processes that incorporate a broad set of values from across society.[156]

I want to focus on the line about incorporating values *from across society*. It gives the impression that 'society', in some form, was brought into the decision-making process. The quotation could be read as saying: *we have constructed a decision-making process that acts as a stand-in for public discussion*. The implication is that, whereas OpenAI merely coordinated with social scientists (as experts), Salesforce was somehow institutionalising the voices of members of society. In doing so, the Salesforce researchers were claiming to have broken the spell of the staged release regime: something about their decision-making process allowed them to legitimately terminate the duty upon research labs to refrain from release. I read the CTRL paper as a reinterpretation of the staged release regime, under which consent from society can be used to override the existing release schedule.

[156] Keskar et al, CTRL, 12.

However, in this case, the engagement with 'society' appears to have been theatrical rather than substantive. PAI does have a wide range of partner organisations, many of which are non-profits. Nonetheless, these partner organisations were not consulted. I interviewed someone at PAI who explained that, while they would like to find a process of engaging partner organisations in these kinds of decisions, they had not yet established any procedure for doing so when Salesforce came to them. The process was more ad hoc and reactive, relying on PAI's internal staff members. The PAI staff members did not recommend the model be released – they did not see their role as recommending what decision Salesforce should take.

Rather, PAI had been working on a list of questions that researchers should ask themselves when deciding whether to release AI research. These questions are now available on the PAI website,[157] and Salesforce uploaded them to the CTRL GitHub repository. The questions include, for example: *If, in one year, you looked back and regretted publishing the research, why would this have happened?* Another example: *In 20 years, how is society different because of your research? What are possible second and third order effects of your work?* The Salesforce researchers took these questions away and came up with their own answers, working together with an in-house Salesforce ethics team. The PAI staff member I spoke to recalled that Salesforce sent back to PAI a summary of their conclusions, without setting out how they had answered the different questions. Those answers were not subsequently published. When I asked the Salesforce researchers if I could see the answers, I was told they were confidential. The PAI staff members were also given NDAs to sign, which meant that the person I spoke to was initially unsure exactly how much they were allowed to tell me about the process.

[157] Partnership on AI, 'Publication Norms for Responsible AI', The Partnership on AI Website, accessed 22 April 2021, https://www.partnershiponai.org/case-study/publication-norms/.

Moreover, I was told by the PAI staff member that, when reading Salesforce's draft paper, they were concerned that it was not sufficiently clear that when the paper talked about PAI, it really meant internal staff at PAI and not the partner organisations. They asked for that point to be made clearer. In other words, it seems that the draft paper was making the process sound more inclusive and multilateral than it actually was. The CTRL blog post also refers to consultation with Salesforce's "Ethical Use Advisory Council".[158]

What happened, then, would be better described as a process of private, structured self-reflection. The researchers were given a template for thinking about the risks of releasing models, and privately went through their own decision-making process. 'Society' was an object within these discussions, rather than a contributor.

Conclusion

This chapter has examined the emergence of the 'staged release' knowledge-control regime, analysing its fundamental elements and how it interacted with the existing institutional landscape of the AI research field.

On its face, the staged release of GPT-2 might look like only a minor alteration to the existing publication regime. OpenAI wrote a scientific paper, they open sourced much of the code, and they eventually released the whole family of GPT-2 models. But in other ways, the new regime was disruptive. It presented a new view of the relationship between science and society. The regime prioritised the societal effects of the model's proliferation, which

[158] According to the Salesforce website: "Our Ethical Use Advisory Council consists of a diverse group of front-line and executive employees, academics, industry experts, and society leaders." https://www.salesforce.com/company/intentional-innovation/ethical-use-policy/

conditioned the release of the larger models (including to researchers). In turn, this necessitated the production of knowledge about the possible and actual societal impacts of GPT-2. The regime delegated some of this work to the broader research community, but OpenAI did much of the work in-house.

The regime can be read as an attempt to overcome the Collingridge dilemma, which holds that knowledge of the societal impact of a technology only comes *after* widespread diffusion has already made those impacts irreversible.[159] The aim was to speed up the process of learning about the social impact of the technology, whilst slowing down the process of irreversible diffusion. However, OpenAI ran into limits on how much they could control these processes. The lab needed to coordinate with other research groups who had the ability to reproduce GPT-2, and – though they found some success in doing so – ultimately the pool of researchers who could reproduce GPT-2 was too large. A number of institutional factors quickened GPT-2's replication, including the fragmented and disorganised nature of the research field, and the incentive of cloud computing companies to make GPT-2 models easily trainable on their hardware. At the same time, GPT-2 was insufficiently difficult to reproduce, given that the methods were not novel and the compute costs were not prohibitive. These factors combined to make coordination difficult, and the staged release of GPT-2 was undercut by other research groups. The GPT-2 episode was therefore an early encounter with the difficulties of controlling the proliferation of AI technologies in the setting of a scientific research field.

The episode also raised the issue of the scientific community's role in governing AI risks. OpenAI clearly perceived a tension between (on the one hand) granting widespread research

[159] David Collingridge, *The Social Control of Technology* (London: Frances Pinter, 1980).

access for studying GPT-2, and (on the other) limiting the proliferation of the model. This tension was largely an issue of information security, where academic labs were not trusted with a copy of GPT-2's code and weights – and in the next chapter, we will see how this information security problem was later overcome. Many academic AI researchers reacted against the fact that OpenAI's staged release process did not afford the AI research community a more central role, and saw the research process as vital for understanding and mitigating the risks of a model. We saw this dynamic playing out in the Grover case, where the academic researchers sought to highlight the benefits of researcher access, training their own model and repurposing it as a possible mitigation (a detector for AI-generated text). While in some cases researchers might have placed too much faith in greater openness as a solution to the possible risks, it nevertheless seems correct that the staged release regime could have been strengthened by a better mechanism for allowing researchers to study GPT-2.

Finally, the GPT-2 episode revealed how KCRs implicitly or explicitly structure the relationship between the research community and the wider society. The research community was accustomed to an arrangement wherein possible societal impacts of their work did not have a bearing on how that work would be shared. This is an arrangement that lends itself well to the metaphor of 'upstream' (research) and 'downstream' (deployment and societal impact). The staged release regime upended this concept: the release of the GPT-2 research was now *downstream* of the observed effects of the smaller GPT-2 models in the world. There is an analogy here to product development in companies, which often relies upon trying to understand the user and adapting the product to fit the user's needs (in fact, one of my academic researcher interviewees used the analogy of a 'minimal viable product'). This new KCR therefore institutes a more closely intertwined relationship between science and society. This intertwinement is not a *necessary* feature of risk-based KCRs (e.g. an alternative

approach would be blanket secrecy of certain areas of research), but it is at least a very natural feature for a risk-based KCR: it is a means of bringing (societal) risk into the operation of the KCR.

CHAPTER 2: INTERFACE

> *[T]he future could be surprisingly close. And these discontinuous jumps in capability are moments when the future lurches into your present.* (CAMERON, OpenAI staff member)

> *I see the idea of putting things behind an API as following up on some unfinished business from GPT-2.* (ROBIN, OpenAI staff member)

Abstract: This chapter is about GPT-3, the more powerful successor to GPT-2. OpenAI shared GPT-3 via an application programming interface (API), which allows users to interact with the model at arm's length. The API was intended as a form of commercial deployment, but it also doubled up as a way of sharing the model with the research community. The API strategy was more viable due to certain underlying scientific developments, especially the emphasis on training very compute-intensive models like GPT-3. I analyse the API as a platform for governance, arguing that the API gives the lab much greater control over use and misuse of the model than other KCRs we have seen. The lab can monitor how people are using the model, vet developers who want to build applications using the model, and shut down access from users who violate the rules. The API enabled an approach to risk mitigation that was both (a) targeted at risky *applications*, which was a more granular approach than withholding whole models; (b) better suited to *iteration* than the GPT-2 staged release. I argue that these governance properties of the API allowed the lab to reposition itself vis-a-vis society, taking on high levels of responsibility for the minutiae of how the model is used. Finally, I discuss the extent to which the API allows the scientific community to understand GPT-3, in the context that greater scientific understanding of powerful AI systems will arguably be risk-reducing in the long-term.

1. Introduction

GPT-3 was a scaled-up version of GPT-2.[160] The largest GPT-3 model had 175B parameters, over 100 times more than GPT-2. The compute required to train GPT-3 was also over 100 times greater, according to estimates.[161] The dataset of text was around 50 times larger.[162] The model architecture was similar, with some minor improvements. The changes resulted in a significant improvement in GPT-3's capabilities over GPT-2.

The knowledge-control regime used for GPT-3 was very different from GPT-2's staged release. The authors uploaded a preprint of the paper in May 2020, without fanfare from OpenAI. Later, in June 2020, OpenAI announced the *OpenAI API* – their first commercial product. GPT-3 was to be the first model available on the platform. GPT-3 would run in the cloud, and users would be able to interact with it remotely. 'Application programming interface' (API) is a broad term in software engineering, where there is some specified procedure through which an individual or piece of software can query another piece of software or dataset. For example, the Google Maps API allows other companies – such as Uber – to build the navigation services into their products, even without the underlying navigation algorithms being exposed.

It is not surprising that OpenAI would use this setup to sell access to their AI technologies, and other companies had already used APIs to commercialise AI systems. Nonetheless, the GPT-3 API is a very interesting knowledge-control regime. First, GPT-3 was a research artefact: it was created as a cutting-edge research product, and it would become a very well-

[160] Tom B. Brown et al., 'Language Models Are Few-Shot Learners', *ArXiv:2005.14165 [Cs.CL]*, 2020.
[161] 3.1×10^{23} FLOPs compared to 2.5×10^{21} for GPT-3. See Jaime Sevilla et al., 'Compute Trends Across Three Eras of Machine Learning', *ArXiv:2202.05924 [Cs]*, 9 March 2022, http://arxiv.org/abs/2202.05924.
[162] Chuan Li, 'OpenAI's GPT-3 Language Model: A Technical Overview', Lambda Blog, 3 June 2020, https://lambdalabs.com/blog/demystifying-gpt-3/.

studied model within the AI research community. The use of an API was a stark break from the open source paradigm within the AI research community. And second, the API allowed for much greater monitoring and control over how people interacted with the model. OpenAI could monitor what people were using the model for, by scanning the text going into, and coming out of, the model. OpenAI could also vet any third party developers who wanted to build GPT-3 into their products. OpenAI created rules on what uses of GPT-3 were safe and ethical, and the API allowed them to enforce these rules.

In this chapter, I analyse the API as a KCR for reducing the risks of GPT-3. The structure is roughly similar to chapter 1. I begin by briefly telling the story of GPT-3 within OpenAI and how the API emerged. Section 3 then discusses how certain developments within AI research have made the API possible. I argue that the 'scaling' approach to AI research (making everything bigger: more parameters, more compute, and more data) helps to delay independent replication of models like GPT-3. Since GPT-3 is so much more compute-intensive than GPT-2 was, replication has been much slower, meaning users are compelled to go through OpenAI's API. If this creates a bottleneck, then section 4 is about how the API manages that bottleneck with the aim of reducing misuse of GPT-3. I analyse how and why the API gives the lab a high level of control over how the model is used. Section 5 turns to the relationship between science and society. I compare the API against the existing KCR of open source models, and argue that the API allows the lab to take on much greater responsibility – and, by extension, governments that regulate them. Section 6 ends the chapter by examining how the API positions the relationship between industry and academic researchers. I find that the API has become a platform for academic research, although there is potentially some room for improvement in this area. This could become an important issue, if the risks of AI depend upon how well the research community is able to understand the models that it builds.

This chapter addresses the core questions of the thesis, shedding light on the interplay between KCRs and risk. One very important theme of the chapter is about how the resource intensiveness of training GPT-3 granted a greater moat around the model, making it possible to protect against replication for a longer time period. The need for large amounts of compute and data had a kind of centralising effect, providing some momentary shelter against the usual operation of anarchic development and release (which we saw in the previous chapter), and opening up opportunities for governance. This phenomenon speaks to multiple questions that are central to the thesis: (1) What sources of power do risk-based KCRs rely upon? (2) How do developments in the underlying science lead to changes in KCRs? (3) How tractable is it to govern the proliferation of AI research?

The other central issue in this chapter is the API itself, as a platform for governing the risks of the model. As a tool for managing risk, the API allowed more fine-grained control over how people used the model, and – in comparison to staged release – the API allowed for a new (and improved) approach to continually iterating on how GPT-3 was being deployed as information came in. The API is therefore very interesting to analyse as a risk-based KCR. Moreover, the API positions relationships between different actors in new ways. Like we saw in the previous chapter on staged release, the API shrinks the distance between the research lab and the wider society. The API does this in a more high-bandwidth way than staged release, with the lab being able to directly monitor user behaviour. The API also addresses a central difficulty from the GPT-2 episode around squaring researcher access with non-proliferation: researchers can use the API to study GPT-3 without needing to be trusted with the model's underlying code. The analysis of the API as a risk-based KCR helps to answer some important questions for the thesis: (1) How do risk-based KCRs work: practically speaking, how can you make risk a feature of the KCR? (2) How do risk-based KCRs position

different actors inside the scientific community, allocating responsibilities for uncovering and addressing risks? (3) How do they rewrite the social contract between the scientific community and broader society?

2. The origins of the GPT-3 API

GPT-3 had flexible and wide-ranging capabilities compared to previous models. The scientific paper emphasised that GPT-3 had unprecedented "few-shot learning" capabilities.[163] The idea is that GPT-3 could be "programmed" to perform a wide range of natural language processing tasks, without any actual programming. The user simply provides an example of the task in the input to the model. An example from the paper:

> Translate English to French:
> sea otter ==> loutre de mer
> cheese ==>

In other words, the user gives – as a "prompt" – a task description and some examples. Successful performance would be where the model outputs the word "fromage". The hope, in this example, is that (a) the model has acquired some knowledge of French from the dataset (7% of which was non-English), and (b) from processing the prompt, the model is able to correctly identify the nature of the task at hand.

Using these methods, the authors found that GPT-3 could perform quite well on a wide range of tasks, even without the researchers having to train the model on task-specific datasets. For example, GPT-3 was able to:

[163] Brown et al., 'Language Models Are Few-Shot Learners'.

- Answer trivia questions (on some datasets out-performing models trained specifically for this task);
- Translate into English from languages such as French, German, and Romanian (with about 40% accuracy, roughly the same performance as models trained specifically for such translation; while underperforming other models when translating *out of* English);
- Answer common-sense questions about the physical world (to 83% accuracy, outperforming any other model);
- Reading comprehension, with an accuracy of 36% to 85% across different benchmarks (generally underperforming models specifically fine-tuned on reading comprehension);
- Generate fake news articles (which human readers were generally unable to distinguish from real news articles);
- Correct grammatical mistakes.

Because GPT-3 was able to flexibly apply a wide range of knowledge and skills, some people have interpreted it as a step towards AGI. AGI (artificial general intelligence) loosely refers to an AI system that has the kind of wide-ranging cognitive abilities that humans have. When GPT-3 became more widely accessible, someone in a discussion group of AI researchers was recorded as observing: 'GPT-3 is completely alien. . . it's the first thing I've seen where it's not a dumb thing to ask whether it's AGI.'[164] Likewise, Connor Leahy (who was now working full-time as a researcher studying such models) described GPT-3 as a 'proto-AGI'. Another researcher tweeted: 'I used to be fairly confident that language models wouldn't be the fastest route to AGI. I now think that confidence was misplaced. . .'

[164] Alex Tamkin et al., 'Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models', *ArXiv:2102.02503 [Cs]*, 4 February 2021, http://arxiv.org/abs/2102.02503.

It appears the spectre of AGI was also present within OpenAI: CAMERON told me that they thought "a lot about the AGI dynamics stuff". He did not spell out what this meant, but the broad contours of such discussions can be inferred from the wider social context. Among the loose community of people worried about risks from AGI (including existential risks), a common concern is that AGI will be built before researchers have worked out how to do it safely. This is often put crudely in terms of relative progress in AI capabilities (i.e. an increase in what AI systems can do) and AI safety research. For example, the Machine Intelligence Research Institute, which focuses on AI safety, decided in 2018 to begin publishing less of their research. A large part of the justification was that they did not want their research to inadvertently 'spark capability advances'.[165] Proponents of AGI risk will sometimes criticise OpenAI for contributing too much to advances in AI capabilities. GPT-3 was actually a prime example of this. For example, Connor (who is still worried about AGI risk) told me:

> Before GPT-3, it was very widely considered that scaling is a very silly thing to do, it's not what *real* researchers do. And I think that was a great thing to slow down capabilities. And OpenAI just torched all of that. And I... let me just say, I'm not happy about that, to put it rather mildly. I think that was potentially one of the most dangerous things anyone has done, ever.

It appears that these kinds of considerations did inform the way that GPT-3 was shared. CAMERON told me:

[165] Nate Soares, '2018 Update: Our New Research Directions', Machine Intelligence Research Institute, 22 November 2018, https://intelligence.org/2018/11/22/2018-update-our-new-research-directions/.

> GPT-3 existed for a long time before the paper came out. We delayed the paper. That was one of the things we could do for AGI stuff. But it's months, it doesn't really count. And you're sitting there, fucking white knuckling it, because it's really costly if someone releases their paper, and you have fucked this up somehow. So you're under pressure.

In addition to delaying the paper, another strategy was to write the paper in a way that avoids attention-grabbing. The paper was written so as to avoid 'hype' and include discussion of the model's weaknesses. Similarly, CAMERON said:

> [H]ow we felt was responsible, especially with AGI stuff, was release just a really big, dull paper. And by dull, I mean, we stuffed it full of everything. Like all of these graphs, loads of interdisciplinary analysis and stuff.

The paper does not give pride-of-place to actual examples of GPT-3's outputs, which is the most accessible way of seeing GPT-3's capabilities. Moreover, unlike with GPT-2, there was no OpenAI blog post introducing the GPT-3 paper. The paper was announced on Twitter by its authors.[166]

The great irony of GPT-3 is that it became one of the most hyped AI systems ever created. An obvious explanation for this is the API. The API originated as an internal tool within OpenAI, allowing various members of the organisation to easily play with GPT-3. The idea of a commercial API grew from there. CAMERON told me:

[166] The lead author tweeted: 'Language models are few shot learners! We find that larger models can often (but not always) perform NLP tasks given only natural language prompt and a few examples in the context. No fine-tuning.'

> Firstly, it started out as a research API. It probably was . . . early January 2020. Some researchers here OpenAI were like, similar to how we had GPT-2 as a chatbot on Slack and other things, they were like, *why don't we make a proper API service, because we've kind of noticed that the more people of the organisation we can get to interact with our bigger models, the more we discover about it, and the more research it unlocks*. . . . I don't think even started as a commercial thing, I think it started firstly as like, *let's further accelerate our own research agenda*. But then pretty quickly after that, people were like, *shit, like, this might be good enough to be like a product*. And, you know, we had this Microsoft agreement,[167] we had some other things; we have a slight incentive to try and make some money. And it's good for, like your relationship with Microsoft to show that you can.

There was GPT-3 as a research project, and there was GPT-3 as a product. These involved different parts of the organisation. While the researchers building GPT-3 wanted to avoid it getting too much attention, it clearly would not have been a very good product if nobody knew about it. In June 2020, OpenAI began the API as a beta product, i.e. an initial, unfinished version of the product for early customers to try. GPT-3 gained widespread attention and publicity when early users started sharing their experiences online. This was in mid-July, which is when Google Trends shows that worldwide searches for "GPT-3" during the year 2020 spiked:

[167] In 2019, Microsoft invested $1B into OpenAI.

Figure 1: Google Trends showing the relative number of Google searches for the term 'GPT-3' throughout the year 2020.[168]

The spike is during the week of 19-25 July. I found a similar trend when, in late July 2020, I downloaded around 63,000 tweets mentioning 'GPT-3' from Twitter's API, from the period 12th-22nd July 2020. The number of tweets mentioning GPT-3 climbed from close to zero at the start of this period to a spike of about 900 (per 3 hour interval) around July 20th.[169] This spike was nearly two months after the GPT-3 paper was released. The tweets I found with the most engagement (in terms of retweets and likes) were early users of GPT-3 who were demonstrating GPT-3's ability to write functioning software code. This was a much more accessible demonstration of GPT-3's capabilities than the paper had given. On Google Trends, during 2020 the country with the most Google searches for GPT-3, as a fraction of all the country's Google searches, was China.[170]

The API, as a KCR for governing AI risk, was a double-edged sword. On the one hand, it did draw attention to GPT-3's capabilities, which some view as exacerbating AI risk. In late 2020, a number of AI safety researchers left OpenAI to start their own entity. They have not publicly criticised OpenAI, but plenty of onlookers have inferred that the researchers left

[168] My Google Trends search is available here: https://trends.google.com/trends/explore?date=2020-02-01%202020-12-31&q=GPT-3

[169] This was excluding retweets. Including retweets, the same spike was roughly 3000 tweets per 3-hour interval. I do not rely too heavily on these numbers, because the Twitter API did not reveal whether it was giving me 100% of the relevant tweets.

[170] Although note that the sample of Chinese Google users is unlikely to be representative of the population, given that Google has much smaller market share in China than elsewhere.

because the company's commercial incentives were at odds with the goal of reducing AI risk, with the GPT-3 API being the go-to example. However, from another perspective, the API is aligned with the goal of reducing risks. As we will see further below, the API is a natural progression from the strategy behind the staged release of GPT-2. Both involve leveraging the lab's temporary monopoly over the technology to control how it is deployed. The difference is that the API gives the lab greater latitude to monitor and block misuse of the model. This ability to tackle misuse goes hand-in-hand with the commercialisation of the model, which similarly requires that the lab exercise control over access to the technology.

3. Scale as a governance strategy

The misuse-prevention strategy behind GPT-3 – as I interpret it – involves two halves. First, an open source version of the model must not be freely available. This establishes a bottleneck. Users are forced to rely on OpenAI or some other lab that also constrains use of its model. I will discuss this first half of the strategy in the present section. Second, the lab must manage that bottleneck. The lab must have some filter: a way of permitting use and study of the model while preventing misuse and preventing the model from being stolen. That is where the API comes in, which I describe in the subsequent section.

Two clarifications are needed. First, when I describe all this as a 'strategy', I am not making a claim about the decision-making within OpenAI. It might be that many decision-makers within the company did not consider GPT-3 in these terms.[171] Or it might be that the strategy as I describe it was intentional, but primarily emerged as a business strategy, with the protections against misuse coming as a bonus. Second clarification: the lab having a

[171] One of my interviewees, ROBIN, did seem to share my interpretation of the strategy, but I do not know how widely this perspective is shared within the company. Jack Clark, the former head of OpenAI's Policy team, has actively pushed for governments to fund academics to train compute-intensive models, which is contrary to the strategy discussed here (assuming the academic models would be open source).

monopoly on GPT-3-like models is not a necessary part of the strategy. As we will see, the strategy still works where other actors also deploy similar models. What is necessary is that any such actor also installs robust protections against widespread misuse. Making an open source version of the model freely available would be one way of undermining the strategy, as would giving very permissive and unmonitored API access.

What is striking is just how much more successful the strategy has been for GPT-3 than GPT-2. Not only is the API a more powerful governance platform than 'staged release' (see next section), but an open source version of GPT-3 – which would go a long way towards undermining the power of the API – has not yet emerged at the time of writing (February 2022). It has been around two years since GPT-3 was trained. Recall from chapter 1 that OpenAI's efforts to govern GPT-2 were overtaken by open source replications within four to seven months of the initial release. The relative success of GPT-3, in terms of non-proliferation, is thanks to one crucial factor: scale. Building GPT-3 involves access to a lot of computing hardware, coupled with an engineering team that can make efficient use of that hardware. In this section I will explain how the 'scaling' approach to AI research is complementary with the API KCR. At the end I will highlight a couple of additional complementarities between GPT-3 as a scientific artefact and the API.

3.1 Scientific changes

Hilgartner argues that, within a scientific field, the knowledge control-regimes and the underlying scientific research shape one another. GPT-3 is a good example of this – in particular, GPT-3's KCR is made possible by certain scientific changes. To understand the strategy behind GPT-3 we have to understand recent trends in AI research.

To give a flavour of the shift, I will start by briefly comparing GPT-3 against a 2014 paper by Iyyer et al, introducing a model called QANTA.[172] I highlight this paper because: (a) like GPT-3, it was an NLP paper; (b) like GPT-3, it was at the research frontier, winning the Outstanding Demonstration award at NeurIPS 2015; (c) the paper is from 2014, before the shift towards "foundation models" (see below). QANTA is built to answer trivia questions. The model sees a factual description of a person, place, or thing (e.g. a US President) and must name it.

The QANTA paper has five authors, from a few different universities. The GPT-3 paper has 31 authors, all from OpenAI. QANTA was trained and evaluated on a single task. The dataset comprises questions and answers from a series of quiz tournaments; a couple of tournament years were held out from the training dataset and used to evaluate the model's performance. There was no pretraining phase. The authors curated the dataset to make the task achievable, by making sure that for every correct answer in the evaluation set, there were at least six instances where that same answer appeared in the training set. In contrast, GPT-3 was pretrained on a dataset that was over 20,000 times larger.[173] This dataset had been automatically filtered to try and improve the quality of the text. GPT-3 was then evaluated on a wide variety of tasks.

Part of QANTA's performance owed to an algorithm for arranging the different words in a sentence into a tree structure, to help the model to parse the relations between the different words. GPT-3 has no equivalent: the process for parsing the relations between different words is learnt by the model. This is something the Transformer architecture is good at; but the

[172] Mohit Iyyer et al., 'QANTA: A Deep Question Answering Model', 2014, https://people.cs.umass.edu/~miyyer/qblearn/index.html.

[173] In 2015 the authors released a dataset that they described as 'much larger', which was 31MB. The largest of the datasets used for GPT-3 pretraining, Common Crawl (filtered), was 570GB. Available at: https://people.cs.umass.edu/~miyyer/qblearn/index.html

same architecture works for other domains, e.g. analysing images or DNA.[174] The methods underlying GPT-3 are more generic than QANTA, and rely heavily on scaling up the compute and data that are used in training.

Stepping back, there are three (closely related) ways in which projects like GPT-3 have challenged the pre-existing scientific paradigm.

First, GPT-3 relies heavily on scale: increasing (a) the size of the model (i.e. its number of parameters); (b) the amount of computation used to train the model; and (c) the size of the dataset. These three go hand-in-hand: the millions of dollars' worth of compute would be wasted on a small model and small dataset (and vice versa).[175] GPT-3 is a concession to the 'Bitter Lesson' as laid out in an influential blog post by Rich Sutton, a well-respected AI researcher:

> The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law . . . Most AI research has been conducted as if the computation available to the agent were constant . . . but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the

[174] Yanrong Ji et al., 'DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome', *BioRxiv*, 1 January 2020, 2020.09.17.301879, https://doi.org/10.1101/2020.09.17.301879; Mark Chen, Alec Radford, and Ilya Sutskever, 'Image GPT', OpenAI blog, 17 June 2020, https://openai.com/blog/image-gpt/.

[175] Jared Kaplan et al., 'Scaling Laws for Neural Language Models', *ArXiv:2001.08361 [Cs, Stat]*, 22 January 2020, http://arxiv.org/abs/2001.08361.

> shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation.[176]

The amount of compute required to train state of the art AI systems has increased dramatically in recent decades. It has increased about 10 billion fold since 2010 (at which point there had already been around a 10 billion fold increase since the 1950s).[177] This is partly because the hardware used in these projects, including GPUs, has become more powerful. The general increase in computational power is a rising tide that lifts all boats. However, there also seems to have been an increase in recent years in the amount of money spent on state of the art results. Since AI performance is amenable to greater computation, instead of waiting for Moore's Law, some AI labs have pushed ahead by outspending others. This kind of competition benefits the industry AI labs who have greater financial resources. They rely on chaining together hundreds of GPUs which operate in parallel.

Moreover, engineering skill is required to make efficient use of these GPUs. The 'Contributions' section in a recent DeepMind paper makes this apparent, listing many authors as responsible for various aspects of such engineering (four authors responsible for 'model parallelism', seven for 'pipelining', and eight for 'hardware efficiency').[178] OpenAI has a specific 'Acceleration' team, which collaborates with other teams to speed up the training of large models.[179] Training large models, even for research purposes, has gone through a kind

---

[176] Richard Sutton, 'The Bitter Lesson', 13 March 2019, http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

[177] Sevilla et al., 'Compute Trends Across Three Eras of Machine Learning'.

[178] Note that some of the same people are listed under each. See Jack W. Rae et al., 'Scaling Language Models: Methods, Analysis & Insights from Training Gopher', *ArXiv:2112.11446 [Cs]*, 21 January 2022, http://arxiv.org/abs/2112.11446.

[179] An OpenAI job advert, 2021: 'The Acceleration team works on everything that makes models faster and more efficient, with the goal of accelerating progress towards AGI. Our team is composed of both researchers and engineers, who work together to develop new libraries, training algorithms, and model architectures. We frequently collaborate with other teams to speed up the development of new state-of-the-art capabilities. For example, we recently collaborated with our Algorithms research team on the DALLE-E model, which generates images from natural language text captions.'

of industrialization, with different engineers and researchers responsible for various aspects of the process of training and evaluation.[180] The scale of today's largest cutting-edge models has changed how they are produced, favouring larger teams, greater specialisation, and more engineering.

Second, GPT-3 other projects have remapped the relationship between different elements in the AI research process: methods, data, compute, models, and tasks. Standard practice in AI research follows the 'common task framework', i.e. the 'quantitative comparison of alternative algorithms on a fixed task'.[181] For example, the ImageNet dataset contains many images and accompanying descriptor categories such as 'miniature poodle'. The task is to predict the descriptor given the image. Authors will introduce a particular method (e.g. a particular model architecture) and compare performance against other methods on this fixed task. The process revolves around comparing different methods. The dataset and task are held fixed, and the performance of the model is a reflection on the strength of the method. The amount of compute used is often not explicitly controlled for, which is a potential weakness of the common task framework; although traditionally the different authors would be competing with reasonably similar compute budgets, and thus compute would be implicitly and roughly controlled for.

Projects like GPT-3 subvert this framework. This goes back to the 'scaling' approach. Instead of controlling for data, the researchers try to increase the amount of data used in training. Part

[180] See Tamkin et al., 'Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models'. 'One participant remarked that the scale of models like GPT-3 was reminiscent of large particle accelerator experiments, which require many people with diverse backgrounds to execute. For example, when training such large models, different teams with diverse expertise must collaborate to run experiments, build and maintain the computing infrastructure, develop the algorithms, and continuously interrogate the model's capabilities for possible problems (e.g., bias, misuse, safety concerns, etc.).'

[181] Mark Liberman, 'Fred Jelinek', *Computational Linguistics* 36, no. 4 (1 December 2010): 595–99, https://doi.org/10.1162/coli_a_00032 p.598.

of the innovation of projects like GPT-2 and GPT-3 is finding a new, bigger dataset. Similarly, the amount of compute is not held fixed, but increased in an attempt to improve performance. The methods used, as per the Bitter Lesson, are generic and designed to leverage the high computation and large dataset..

Moreover, the model is not just a reflection of the methods, but takes on a life of its own. The model can perform a wide variety of tasks, to varying degrees of performance. Evaluating this performance involves a lot of work. Hence, most of the GPT-3 paper is devoted to evaluation. The model becomes this big, mysterious object, and the process of trying to understand the model requires a lot of work and skill. In addition, the authors compare GPT-3 against other models, such as BERT, but the comparison is not like-for-like (GPT-3 has one arm tied behind its back, because it is not fine-tuned on the given task). The argument of the paper is *not*, 'the small differences in GPT-3's design make it better than BERT', but rather, 'look at these interesting new capabilities that emerge at scale'. The authors appear to show no interest in demonstrating that the GPT architecture is better than (for example) BERT's.

Third, there has been the rise of 'foundation models'.[182] The 2021 paper that introduced this term gave it the follow definition:

> A foundation model is any model that is trained on broad data at scale and can be adapted (e.g., fine-tuned) to a wide range of downstream tasks; current examples include BERT [Devlin et al. 2019], GPT-3 [Brown et al. 2020], and CLIP [Radford et al. 2021].[183]

[182] Rishi Bommasani et al., 'On the Opportunities and Risks of Foundation Models', *ArXiv:2108.07258 [Cs]*, 18 August 2021, http://arxiv.org/abs/2108.07258.
[183] ibid. p.3

An important feature of such models is that they have a large space of latent knowledge and skills. Such knowledge and skills must be elicited, normally through fine-tuning the model on a particular task – although as we have seen, GPT-3 goes one step further by achieving decent performance without any fine-tuning. Foundation models represent a new scientific paradigm based around *transfer learning*, i.e. taking a model and adapting it to tasks for which it was not originally trained.[184] A model like GPT-3 or BERT can give rise to its own research ecosystem of people figuring out what tasks the model can and cannot be adopted to.

Given these scientific developments, the question becomes: how do they affect what KCRs are viable? I will now turn to this question.

3.2 The compute moat

In some ways, this new breed of models might be harder to govern. A well-known paper from 2021 asks, can language models be too big?[185] For example, with larger datasets it is harder for the researchers to understand the data on which the model is trained. Foundation models also come with a range of capabilities and so are especially dual use. However, at the same time, the shifting scientific landscape presents new opportunities for governance. I would argue that the emerging scientific paradigm and the KCR used to govern GPT-3 are natural bedfellows. The most important change, which I will focus on here, is that the scale of GPT-

[184] The leader of one AI company tweeted in 2021: 'Very few people realize that there's a massive tech paradigm switch happening between old-school ML (dare I say "big data"?) and new generation ML (transfer learning).'

[185] Emily M. Bender et al., 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.

3 makes it harder to replicate. In the next subsection I consider three additional ways in which the scientific changes have supported the new KCR.

A useful point of comparison with GPT-2 is that Connor Leahy has now been working on replicating GPT-3. He is one of the leading members of Eleuther AI, which he describes as a "loose research collective". Despite deciding against sharing his version of GPT-2, he now believes that GPT-3 should be open sourced so that researchers can study the model. His motivation is that we need to deeply understand how GPT-3 works because future versions could be very risky. He wants to open up GPT-3 for low-resource researchers to do interpretability and alignment research on the model.[186]

In the summer of 2021, Eleuther released a 6B parameter version of the model. In February 2022 they released a 20B version.[187] OpenAI's GPT-3, trained two years earlier, is still nine times larger. Connor explained that part of the challenge was the engineering difficulties that come with training the model at such scale:

> [W]ith GPT-2 it was really just kind of like: plug the parts together. . . . With GPT-3, it's very different, in the sense that there are like very non-trivial engineering challenges to be solved. So, I couldn't have done this alone. Like there was no way I could have done this as a one man job. With Eleuther, you know, 10 hackers in a cave or whatever, it's been surprisingly tractable.
>
> [...]

[186] Where 'alignment' is about ensuring AI systems follow human values and are amenable to human oversight and control.
[187] Connor Leahy, 'Announcing GPT-NeoX-20B', EleutherAI Blog, 2 February 2022, https://blog.eleuther.ai/announcing-20b/.

> The main difference is that with GPT-2, you could still fit the model on a single GPU generally. But that's, of course, not even close to the case with GPT-3. You have to split up the model across many different machines.

The engineering challenges have been getting easier with time, with new engineering frameworks for training large models. Eleuther has used Microsoft's DeepSpeed code library for training the model in parallel across different GPUs. Connor told me that, although this was not plug-and-play, the problems were tractable with 'five to 10 good engineers'.

The greatest bottleneck has been getting access to enough compute. Initially Eleuther was still using Google's TFRC scheme. This was not sufficient:

> . . . Google just could not give access to enough TPUs to train it, it just wasn't possible. GPT-3 is truly massive. And there's no way they could give us access to enough TPUs for long enough to actually train a GPT-3 model.

The insufficient access to compute meant that by December 2020 the Eleuther team was ready to give up on trying to train a GPT-3-sized model. That changed when they were approached by CoreWeave. CoreWeave was originally a Bitcoin mining company, and so they owned a large number of GPUs. They were pivoting into cloud computing. They planned to buy more NVIDIA GPUs and rent them out to people training large models. Connor told me:

> So, the deal was: we test the hardware, we figure out what do you need to train these kinds of models . . . because they don't have in-house capacity ML engineering talent.

> And then they buy [the hardware]. We get to train our model on it and release it for free. And everyone's happy.

Jack Clark, the previous head of OpenAI's Policy team, had been critical of Google TFRC's sponsorship of the GPT-3 replication. Clark wrote in his weekly newsletter in March 2021:

> Google's actions here are confusing. On the one hand, the company publishes AI principles and periodically goes on publicity drives about 'responsible AI'. On the other hand, Google is enabling the release of a class of models with some non-trivial ethical challenges via a process that lets it sidestep accountability. It's hard for us to know what Google believes as an institution, here.
>
> . . . It'd be interesting to understand the thinking here - does TFRC become the means by which Google allows open source models to come into existence without needing to state whether it has chosen to 'release' these models?[188]

One of Eleuther's members replied on Twitter:

> I can verify that the person who runs TFRC knows exactly what we are up to. We've discussed it with him several times, largely in an attempt to get enough non-preemptible TPU access to scale to 100B models. Unfortunately that wasn't in the cards.

[188] Jack Clark, Import AI Newsletter #241, March 2021. Available at: https://jack-clark.net/2021/03/22/import-ai-241-the-2-million-dataset-small-gpt-3-replications-imagenet-gets-a-face-blur-update/

Aside, as with chapter 1, we can interpret events in the context of the competition between NVIDIA's GPUs and Google's TPUs. Since the switch away from Google's TFRC scheme, Eleuther is now using the code from NVIDIA's Megatron project (see chapter 1), combined with code written with Microsoft's DeepSpeed. DeepSpeed was also likely released to help build the ecosystem around Microsoft's cloud computing service, which uses NVIDIA GPUs.[189] CoreWeave put in a large order of GPUs from NVIDIA and Eleuther's code might be useful for others training large models on NVIDIA GPUs. Arguably, then, the episode is a win for NVIDIA.

When I interviewed Connor in April 2021 he told me that what was slowing Eleuther's project down was still access to hardware. He said:

> So currently, funnily enough, the bottleneck is literally hardware. Is that due to the chip shortage, CoreWeave just can't order enough GPUs. They've already ordered like hundreds of GPUs, and they're just not arriving. So that's what we're currently bottlenecked on.

The global chip shortage has delayed the replication project. It is interesting to see how the global supply chain for high-end hardware has had a direct impact on the time period during which OpenAI (and other AI companies) can control access to GPT-3-like models.

Unequal access to compute, combined with the compute-intensiveness of the current state of the art, has paved the way for a certain kind of 'theory of change' for OpenAI. They build AI

[189] See Rangan Majumder and Andrey Proskurin, 'ZeRO-Infinity and DeepSpeed: Unlocking Unprecedented Model Scale for Deep Learning Training', *Microsoft Research* (blog), 19 April 2021, https://www.microsoft.com/en-us/research/blog/zero-infinity-and-deepspeed-unlocking-unprecedented-model-scale-for-deep-learning-training/.

systems that are on the frontier of what is currently computationally possible. They can then govern those AI systems by deploying them through the API, trying to steer how the models are applied before they become widely available.

This is simultaneously a business strategy and a way of contributing to AI governance. OpenAI was always intended as a vehicle for governing AI. At its conception, Elon Musk, one of the original founders, was quoted in the media:

> "We discussed what is the best thing we can do to ensure the future is good?" he said. "We could sit on the sidelines or we can encourage regulatory oversight, or we could participate with the right structure with people who care deeply about developing AI in a way that is safe and is beneficial to humanity."[190]

In this way, they viewed the company as an active way of governing AI, in comparison to state regulation. The theory of how OpenAI would contribute to AI governance has changed over time. Initially, the theory was that OpenAI would make AI more accessible, preventing any other lab from monopolising it. This was based on the idea – like with Grover in chapter 1 – that the best defence against AI harms is AI itself. In a media interview, founders Elon Musk and Sam Altman explained how they wanted AI to be 'freely available to everyone'. The interviewer asked, 'If I'm Dr. Evil and I use it, won't you be empowering me?'. The transcript continues:

> Musk: I think that's an excellent question and it's something that we debated quite a bit.

[190] John Markoff, 'Artificial-Intelligence Research Center Is Founded by Silicon Valley Investors', *The New York Times*, 11 December 2015, sec. Science, https://www.nytimes.com/2015/12/12/science/artificial-intelligence-research-center-is-founded-by-silicon-valley-investors.html.

> Altman: There are a few different thoughts about this. Just like humans protect against Dr. Evil by the fact that most humans are good, and the collective force of humanity can contain the bad elements, we think its far more likely that many, many AIs, will work to stop the occasional bad actors than the idea that there is a single AI a billion times more powerful than anything else. If that one thing goes off the rails or if Dr. Evil gets that one thing and there is nothing to counteract it, then we're really in a bad place.[191]

I do not know what sociological factors explain OpenAI's shift away from its original theory of impact. One possible explanation could be turnover in the people running the company (Musk left, and many people joined). Among the community of researchers concerned with long-term risks from AI, there is certainly no consensus in favour of wider proliferation being risk-reducing (anecdotally, the opposite position seems more popular, but I do not have good data on this). Another plausible background factor for the shift could be increased anxieties around China's involvement in AI and the Chinese state's use of AI for surveillance, often relying on technologies developed in the US and other Western countries. This has not followed the theory that (from the extract above) 'many AIs will work to stop the occasional bad actors'.

The emerging strategy, visible with GPT-3, involves building compute-intensive models and then selectively deploying them for non-harmful applications. I suggest the following, general description of the 'compute moat' strategy:

[191] Steven Levy, 'How Elon Musk and Y Combinator Plan to Stop Computers From Taking Over', Blog, *Backchannel* (blog), 11 December 2015, https://medium.com/backchannel/how-elon-musk-and-y-combinator-plan-to-stop-computers-from-taking-over-17e0e27dd02a#.fto7v7a1r.

Where some class of AI systems must be trained using an amount of computation that is prohibitively expensive or inaccessible for most actors, the actors training those AI systems are in a better position to govern their development and deployment. Governing AI development is easier because there are fewer AI developers to coordinate or regulate.[192] *Similarly, governing AI deployment is easier because the lab can choose which capabilities are made available for what applications, in the absence of widespread proliferation. The strategy is to undertake computationally expensive AI development projects in an attempt to realise these governance opportunities.*

OpenAI's approach to GPT-3 is an example of this 'compute moat' strategy, specifically relating to AI deployment. GPT-3 was a natural fit with the compute moat strategy. ROBIN said:

> [I]n the case of GPT-3 . . . the input to the production function that's the most scarce and that makes the biggest contribution . . . is the compute. . . . [M]odels are basically like compressed compute. . . . Whereas, with the data and the algorithm, it's less big of a deal. And, since GPT-3 was very similar to GPT-2 architecturally and training-wise and so forth, those parts aren't that sensitive. And the data, while requiring some expertise to build, was generally less difficult.

[192] CLEO, an academic AI researcher I interviewed, said: 'Okay I'll just say: right now these methods, these models aren't, I think, all that scary. But maybe in a few years, they will start to be. And at that point, like, I guess the fewer the actors that have access to these, the more chance we have to have some sort of coordinated response. Again, just like nukes.' Similarly, ROBIN told me: 'Just from a . . . governance perspective, I think there are definitely upsides to there being different conglomerations of compute that are larger than others, that sort of give you more control. And if everyone had exactly the same amount of compute . . . then that would potentially be dangerous, you know, from a coordination perspective.' For literature, see: Sverker C. Jagers et al., 'On the Preconditions for Large-Scale Collective Action', *Ambio* 49, no. 7 (1 July 2020): 1282–96, https://doi.org/10.1007/s13280-019-01284-w; Stuart Armstrong, Nick Bostrom, and Carl Shulman, 'Racing to the Precipice: A Model of Artificial Intelligence Development', *AI & Society* 31, no. 2 (n.d.): 201–6.

ROBIN contrasted this with other models like AlphaFold, DeepMind's protein prediction system, which relied more heavily on 'algorithmic secret sauce than it being a moat based on compute'.

Nevertheless, CAMERON described the compute moat around GPT-3 as 'tiny', and 'a year, at best'. ROBIN told me: 'one of the things that currently makes me anxious is: how do we prepare, as a company, as a society generally, for a world in which there are more available, more powerful open source language models'. ROBIN listed some strategies for keeping a competitive advantage: serving the models at low latency[193] and high reliability, updating the models with new data, and bringing out newer, fine-tuned versions of models. But ROBIN nevertheless conceded that an open source version:

> would run the risk, particularly on the low end of the ethical scale, of under-cutting some of the constraints that we're trying to impose around: requiring humans to be in the loop, and prohibiting certain malicious use cases . . . [F]rom a malicious use perspective, the issue looms large.

Looking forward, if labs spend even more money on training large models, the compute moat could grow. ROBIN said:

> Right now, yes, GPT-3 was a very expensive experiment. But, you know, it's still in the millions-ish range. If like five years from now, or 10 years from now, there were experiments in the range of billions, then that's going to be well beyond the amount that you could get just as a favour from a cloud provider, or with academic credits to

[193] "Low latency" means the user receives the output of their model very quickly after submitting their query.

> TPUs, or something like that. So it's like, I think it's likely that the long-term trend is towards a small number of corporate and government actors having a relatively big moat, at least in some areas.

Another relevant variable is whether other high-resource actors decide to open source their models. As I mentioned above, the strategy does not depend on having a monopoly over a model, but rather convergence in the strategy or norms adopted by the pool of labs that have built similar models. A couple of AI companies have adopted the same approach as OpenAI: Cohere, based in Canada, and AI21 Labs, based in Israel. Both companies trained GPT-3-like models and charge for access via an API. Both have usage guidelines that are very similar to OpenAI's.[194] OpenAI has demonstrated that there is a possible business model here, which gives these companies one reason not to open source their models. There have also been a number of GPT-3-like models developed for research purposes and kept private, including: Megatron-Turing NLG from Microsoft and NVIDIA;[195] Gopher from DeepMind;[196] PanGu from Huawei;[197] and HyperCLOVA from a South Korean company.[198]

---

[194] In AI21's terms of use, I even found a telling mistake: in many cases when they mention their own company's name (AI21), it hyperlinks to OpenAI's company charter. The drafters were clearly paying attention to OpenAI's thoughts on ethics and governance.

See also ROBIN on norm-setting: 'And while we maybe haven't been explicit about the norm-setting that's going on, I think, you know, part of the reason why we made the documentation publicly available . . . is to shape how people think about the governance of language models. . . . There's an object-level safety dimension, as well as more of a meta-level norm thing going on, when we do things like production reviews of things before they move into practice, which is not universally adopted by companies that have APIs. In fact, like, that's like.. as far as I know it's pretty rare to do stuff like that, to have that hands-on oversight of how an API is being used. So, I think there definitely is a norm setting dimension, we just maybe haven't been as explicit about it as we were in the GPT-2 context, and partially because it's just like a different context. In GPT-2, it was more of a delta in sort of like researcher/academic norms, whereas in this case, it's more about commercial deployment, and there's a less clear precedent that we're breaking.'

[195] Ali Alvi and Paresh Kharya, 'Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model', *Microsoft Research* (blog), 11 October 2021, https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/.

[196] Rae et al., 'Scaling Language Models'.

[197] Wei Zeng et al., 'PanGu-α: Large-Scale Autoregressive Pretrained Chinese Language Models with Auto-Parallel Computation', *ArXiv:2104.12369 [Cs]*, 26 April 2021, http://arxiv.org/abs/2104.12369.

[198] Boseop Kim et al., 'What Changes Can Large-Scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-Scale Korean Generative Pretrained Transformers', *ArXiv:2109.04650 [Cs]*, 28 November 2021, http://arxiv.org/abs/2109.04650.

3.3. Ease of use

If the scale of GPT-3 helps to create a situation where users have no choice but to use the API, there are two factors which mean that the API is a natural way of using GPT-3. Thanks to certain developments in the underlying science, the API KCR has become more attractive.

First, the size of GPT-3, in terms of its number of parameters, is large enough that people cannot use the model on their own hardware. In a practical sense, an API is the most accessible way of letting people use and study the model. As CAMERON put the point:

> [A]cademics aren't going to be able to run inference off of a 200 billion model – they'd need to have a cluster, they'd need to do multi GPU model sewing; it's a very rare set of skills. But if you can give them an API to it, it makes it easy for them to do a certain kind of research on this stuff.

The same issue has cropped up with Eleuther's attempt to open source GPT-3 models. Connor told me:

> I think it does have an impact. In that like, you know, we recently released like, a 2.7 billion parameter model. So, like 100th the size, and we constantly get people bitching on our Discord that this is unusable. Like, "this can't be used." It is the case that even if all the work for you is done with GPT-3, actually running it is a hardware investment on the order of hundreds of thousands of dollars.

A more recent example: when Eleuther released their 20B GPT-3 model in February 2022, one AI researcher tweeted that the 'full model weights can be downloaded for free'. One of

the replies was: ‘Download to what… these things are getting too big for my hard drive’.[199] My conversation with Connor continued:

> Toby Shevlane:
>
> Because in a way, I guess, talking about openness – the word “openness” is a little bit vague. And in a way, it's more open to have an API in some way, like I know-
>
> Connor:
>
> In some ways, yeah. I agree.
>
> Toby Shevlane:
>
> Because people can actually use it.
>
> Connor:
>
> . . . Yeah, I agree with that. Like “democratisation” – what does it even mean, when you're talking about models like this?

The API is easy to use, with the model running on OpenAI’s servers. The fact that this is by far the easiest way to interact with GPT-3, due to its size, likely reduces the pressure upon OpenAI to release an open source version.

Second, because GPT-3 often does not need to be fine-tuned on a given task, the user can access a large portion of the model’s capabilities even with just a simple input-output interface.

[199] I have corrected a spelling mistake in this quotation (‘too’ from ‘to’).

GPT-3 has sparked the idea of ‘Software 3.0’. This builds upon an influential blog post by the well-known AI researcher Andrej Karpathy.[200] Karpathy observed that traditional software (Software 1.0) is written directly by programmers using code. With deep learning (Software 2.0), the eventual program, i.e. the model, is not directly written by the programmer. The programmer instead writes more hands-off code that will search for a program that performs well on a given dataset. After GPT-3, another researcher, Chris Olah, playfully suggested on Twitter that we are now at Software 3.0. Here, the programmer designs a prompt, i.e. the text that is fed into the model, which specifies what they want the model to do, and the model can identify what is being asked of it. The documentation for the GPT-3 API reflects this idea, saying: ‘You can “program” it by crafting a description or writing just a few examples of what you’d like it to do.’ There is a section on ‘Prompt Design 101’, as if the text that the user supplies to the model is an engineering craft.

When it works, prompt engineering is easier than fine-tuning, which involves collecting new data and putting the model through a second stage of training. I put to ROBIN my claim that this favours the API approach. He said:

> . . . I do think that there is some element to what you're saying, yeah, that the larger number of things that you can do without access to the model internals, the more potentially useful it is to have only this kind of like black-box interface to it.
>
> . . . And, I mean, just speaking from my own experience. Just a few years ago, it would take fine-tuning of GPT-2 in order to get qualitatively interesting behaviour on

[200] Andrej Karpathy, ‘Software 2.0’, *Andrej Karpathy* (blog), 11 November 2017, https://karpathy.medium.com/software-2-0-a64152b37c35.

> some tasks. . . . And it's just easier when the model kind of already has a lot of that knowledge, like built into it. And it's just a matter of writing the prompt.[201]

I do not want to overstate the point, because fine-tuning GPT-3 still leads to performance improvements over prompt engineering.[202] However, at the same time, the OpenAI API attempts to capture that territory too: since December 2021, users can fine-tune GPT-3 on their own data and access the resulting model through the API. The API provides an easy-to-use interface for fine-tuning: the user uploads a labelled dataset, and the fine-tuning occurs remotely on OpenAI's servers. Moreover, recent scientific evidence suggests that larger models are actually more data-efficient, suggesting that a smaller dataset will be needed for fine-tuning.[203]

Whether through prompt engineering or data-efficient fine-tuning, GPT-3 has a property that we can call *pliability*. The model is pliable in that it can be adapted to a specific task with less input from the engineer or user than would traditionally be required. Pliability is closely connected to generality: generality is the model's latent capacity to generalise to a range of

---

[201] A more recent version of GPT-3, the 'Instruct' model, takes this a step further. OpenAI researchers fine-tuned GPT-3 to respond better to direct instructions than the original model. See Long Ouyang et al., 'Training Language Models to Follow Instructions with Human Feedback', *ArXiv:2203.02155 [Cs]*, 4 March 2022, http://arxiv.org/abs/2203.02155.

See also JAN, an NLP researcher, describing how GPT-3 changes the development pipeline: 'There is a more traditional pipeline, which is replacing one thing in typically data analysis, but with BERT embeddings, or things like that. Where basically, the model is just a part of a much . . . bigger pipeline, and you have a lot of steps where you have actually humans involved in the pipeline. So that's kind of the traditional way. And people have started to.. Well, I mean, mostly OpenAI, I would say, have started to push for like more "direct use", where the model is actually not just a part of a bigger pipeline, but becomes the pipeline now. . . . So mostly this is reflecting the type of models: you have models which can just be used to extract information. So, BERT, these models, they cannot generate stuff if you want, they just kind of give you embeddings, just as vectors basically. And you have some types of model which can generate things, and then they can be used directly as an output.'

[202] Rachel Lim, Michael Wu, and Luke Miller, 'Customizing GPT-3 for Your Application', OpenAI, 14 December 2021, https://openai.com/blog/customized-gpt-3/.

[203] Tom Henighan et al., 'Scaling Laws for Autoregressive Generative Modeling', *ArXiv:2010.14701 [Cs]*, 5 November 2020, http://arxiv.org/abs/2010.14701; Ting Chen et al., 'Big Self-Supervised Models Are Strong Semi-Supervised Learners', *ArXiv:2006.10029 [Cs, Stat]*, 25 October 2020, http://arxiv.org/abs/2006.10029.

different tasks, and pliability is about how much input (data, engineering effort) it takes to get the model to do so.

My argument is that pliable models naturally favour a kind of centralisation. Previously, engineers developing applications with the model would have a greater need for their own, local copy of the model, so that they can adapt it to their specific application. This kind of development pipeline benefits from an AI research community that open sources its models. In contrast, with GPT-3, more of that work – of adapting to a specific task – is done by the model itself. This does not remove the need for third party developers, who OpenAI still relies upon. But it allows for a development landscape where all these different, specific applications flow through a single point, which is the API. The pliability of GPT-3 makes the API more feasible.

This connects back to the Varshney et al paper discussed in chapter 1.[204] Recall their argument that, due to the prevalence of fine-tuning, a model's social impact is determined *after* it has left the hands of a lab like OpenAI. However, the GPT-3 API brings this back in-house, either through the prompt engineering or the API's fine-tuning functionality. This is empowering for the lab, which can – using the API – exert control over how the model is used even when it has been adapted to a specific task or dataset.

4. The API as a platform for governance

In this section, I discuss the API as a tool of governance, explaining why it is more powerful than the 'staged release' regime we encountered last chapter.

[204] Lav R. Varshney, Nitish Shirish Keskar, and Richard Socher, 'Pretrained AI Models: Performativity, Mobility, and Change' (arXiv:1909.03290 [cs.CY], 2019).

4.1 How the API works

The simplest way of interacting with GPT-3 is on the "Playground". This is a web interface that OpenAI provides. The user types the input into a text box, and clicks 'generate' to see the model's output. The interface gives the user several knobs and dials. They can select from a range of different model sizes. They can choose how long the model's output should be. They can also choose how the model's outputs are sampled: the model itself does not give a single prediction for the next word, but a probability distribution over many. Therefore, the user can choose the 'temperature', where a high temperature means that more unlikely words will be sprinkled in, and zero temperature means that each word will be one the model deemed most likely to appear in that spot. When GPT-3's outputs are displayed, each word can be colour-coded to show how likely the model deemed it. (Note: technically, the model's outputs are not *words*, but chunks of words, called 'tokens'. Short words, like 'dog', contain one token, but longer words contain multiple tokens.)

The API is like the Playground except the user interacts with the model through a programming language. This means that the interaction with the model can be integrated into other programmes. This is useful for third party developers who want to build GPT-3 into their products. It is also useful for researchers studying GPT-3, who can write code that, for example, tests the model on a particular benchmark. Over time, OpenAI has added additional features to the API, giving users deeper access to the model:

1. By default, users can see the model's top 5 next-token predictions and their corresponding probabilities. Users can apply to increase this.

2. Users can now get ‘embeddings’ from the model. Embeddings are not the model’s predictions about the next token, but one step earlier. Embeddings are a high-dimensional representation of the results of the model’s information-processing.
3. Users can create a new, modified version of the model, and then access that model through the API. They do this by fine-tuning the model on their own data.

For convenience I will refer to both the Playground and the actual API as simply ‘the API’. To get access to the API, users need a log-in. Currently, anybody in approved countries can make a log-in and start using the API, whereas previously there was a waiting list. There is a pay-as-you-go system. For the largest model, $1 will buy (very roughly) 10,000 words of text (as input or output).[205] Researchers can apply for subsidised rates via OpenAI’s academic access programme (see section 6).

The API is clearly intended, first and foremost, to be built into applications by third party developers. The ‘usage guidelines’ set out in detail what kinds of applications are permitted. At a high level, these guidelines prohibit applications that: post content on social media or news websites; involve targeted advertising or classification of individuals on demographic grounds; interact with people in a high-stakes or sensitive setting, such as psychiatry or credit checks. Deceptive and misleading applications are prohibited, as are applications that are political in nature, or involve adult content. OpenAI also will not approve applications that are too open ended. In the extreme case, for example, consider an application that just acts as a wrapper around GPT-3, allowing users full access to the model – this would circumvent the

[205] As of February 2022, the price for the largest model is 2¢ per 1,000 tokens. Short words contain one token, with larger words having more (a token roughly corresponds to one syllable). The price for a token is the same regardless of whether it is fed into the model or generated by the model.

other controls that the API imposes. The usage guidelines are evolving over time, as the company gains experience from people using GPT-3.

Developers building applications with the API must go through an approval process before launching their application, otherwise their access can be shut down. OpenAI staff will check that the proposed application conforms to the rules. They might suggest changes to the application – for example, asking that the application require user log-ins. In parallel, CAMERON said they also give developers easy-to-use tools that might make the application safer. For example, OpenAI can provide access to a tool that classifies outputs based on toxicity. CAMERON told me:

> [If we] assume all of our developers are well-intentioned but lazy, how we can probably extend our control is by making tools that we can devolve to them that they will end up using because they help them with their goal. . . . [M]ost entrepreneurs, in an ideal world, are going to be choosing the safest thing we offer.

The safety of the application is therefore achieved by a mixture of (a) enforcing rules, by leveraging the power that the API setup affords OpenAI, and (b) a softer kind of power, where OpenAI tries to nudge developers into certain best practices.

When the application is up-and-running, OpenAI can follow-up to make sure that it is still conforming to the usage guidelines. Sometimes this will involve trying out the application. ROBIN said:

> So it might be something you would want to look at their website. . . . [E]xamples of things that we might consider are: trying out the application and red teaming[206] it via getting a log-in from the actual developer . . . [F]or example, we have a standing meeting internally OpenAI from various people across policy, safety, etc, to talk about API kinds of issues, and one of the kinds of things that will come up there is like, *okay, I want to just try it out and see*; . . . we'll get a few test usernames from the developer and try it out ourselves.

As ROBIN told me, 'some of it is stuff that we can track programmatically, like if we wanted to enforce that a certain prompt is always used at the beginning of each query, that's something we can check'. OpenAI also builds automatic tools to screen for particular misuses. CAMERON described these tools:

> We're developing classifiers to classify stuff being generated. And we're trying to do this obviously in a privacy-centric way. And it's entirely for safety. I'll give you an example. We developed a political classifier a while ago. . . . [I]f you're interested in knowing if say GPT-3 is being used to generate propaganda for a presidential election in the US, you really want a political classifier.

In March 2021, GPT-3 was generating an average of 4.5 billion words each day.[207] This is clearly too much content to screen manually. The automatic tools are necessary because they can operate at scale. CAMERON told me:

[206] "Red teaming" in this context means to adopt the perspective of somebody trying to misuse the application.
[207] OpenAI and Ashley Pilipiszyn, 'GPT-3 Powers the Next Generation of Apps', OpenAI Blog, 25 March 2021, https://openai.com/blog/gpt-3-apps/.

> I guess our attitude is: tech is ultimately so much higher leverage and more scalable. And so we need to kind of, we want to get to a world where tech does 95% of it. And that's not tech solutionism at all, which is like a common failing. It's more like it's just the base of the pyramid.

I was not told the details, but it seems highly likely that OpenAI will be using fine-tuned GPT models as classifiers. Text classification is exactly the kind of capability that has seen improvement in the past several years thanks to the rise of large language models. This example is part of the wider trend that, as AI advances, it provides both offensive and defensive capabilities.[208] There is a parallel to Grover (see chapter 1), with a difference: in the API case, the classifier steps in earlier in the process, i.e. where the text is produced rather than where it is posted. This is made possible because the API provides a natural platform for monitoring and blocking misuse, unlike open-source.

An example of monitoring: one GPT-3 application is AI Dungeon, which is an online text-based adventure game. The player can guide the storyline, and the game uses GPT-3 to continue the story. In April 2021 it emerged that some players of AI Dungeon were using the game to generate sexually explicit stories about chilren. This was picked up by OpenAI's monitoring, and OpenAI asked the creators of AI Dungeon to add a filter that would prevent these kinds of stories from being generated. Some users complained online about the new filter, arguing that it flagged too many false positives (e.g. references to an "8-year-old laptop") and raised privacy concerns.[209]

---

[208] Ben Garfinkel and Allan Dafoe, 'How Does the Offense-Defense Balance Scale?', *Journal of Strategic Studies* 42, no. 6 (2019): 736–63, https://doi.org/10.1080/01402390.2019.1631810; Toby Shevlane and Allan Dafoe, 'The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?', in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20 (New York, NY, USA: Association for Computing Machinery, 2020), 173–79, https://doi.org/10.1145/3375627.3375815.

[209] Tom Simonite, 'It Began as an AI-Fueled Dungeon Game. It Got Much Darker', *Wired*, accessed 11 April 2022, https://www.wired.com/story/ai-fueled-dungeon-game-got-much-darker/.

Some misuses of the model are harder to detect. In the summer of 2020, Liam Porr, an undergraduate at UC Berkeley, used GPT-3 to write a series of self-help blogs. He did not have API access himself, but found a PhD student at Berkely who was willing to run Liam's prompts through the API and send back the results. Liam wanted to demonstrate GPT-3's capabilities by showing that it could write posts that successfully tricked people into believing they were human-written. Over the course of two weeks, he generated about one blog post per day, and posted them to various websites, especially the Hacker News forum. He would write a title, a short summary, and the first couple of sentences. He would then select from around five to ten GPT-3 generations, with very minimal editing. The posts were very popular: for example, one post – *Feeling unproductive? Maybe you should stop overthinking* – became the top post on Hacker News. The posts ended when Liam decided to stop after a couple of weeks, rather than OpenAI noticing and intervening. Liam told me:

> . . . I think that it is going to be very difficult to be able to prevent all forms of malicious use for this AI, just because from their end, like the way that language is used, it's not is not necessarily obvious, just by looking at text, how it's being used. And if it's being used in a good way or a bad way. For instance, with a blog post, like the blog posts that I was making, it's not clear at all from looking at them that they're being used in any way that is not within the intended use case. But given the context of it being on a blog and not being labelled GPT-3 text, then it suddenly becomes an issue. So from OpenAI's perspective, it's a very hard issue to be able to moderate from their end.[210]

[210] See also ALEXIS: 'I feel like it would not be that hard for me to pretend to be like, a researcher studying fake news and how to combat it and just secretly run like a fake news factory on the side. Right? And just use my same API tokens for both.'

In other words, whether or not some activity constitutes misuse does not just depend on the content of the text but also the wider context, which is not necessarily visible to OpenAI.

Finally, the API is designed to prevent users from stealing GPT-3. It is possible to train one model using the outputs of another model as data.[211] The model's outputs can be a better source of data than actual text data, because they are richer. For example, instead of seeing how an actual sentence finished, you can look at the various different ways in which GPT-3 predicts that the sentence could have finished, along with their respective probabilities. This is presumably why the API only outputs five predictions by default. In addition, the API comes with usage quotas, which users must apply to increase. ROBIN told me:

> And so, as far as I understand, [model stealing] would only really be an issue at the largest scale of our customers. So, it has to be pretty egregious, like one of our biggest customers pulling a pretty big wool over our eyes. So, it's not the sort of thing you could do with free access, or just with a few hundred dollars or something.

Overall, these different features of the API help to explain why it is a powerful governance tool. Next, I will go one level more abstract, and describe the properties that emerge from the different features we have seen so far.

4.2 The governance strengths of the API

The API has two key properties that help with preventing misuse.[212] I will call these *risk disaggregation* and *iteration*.

[211] Florian Tramèr et al., 'Stealing Machine Learning Models via Prediction APIs', *ArXiv:1609.02943 [Cs, Stat]*, 2 October 2016, http://arxiv.org/abs/1609.02943.
[212] See further: Toby Shevlane, 'Structured Access: An Emerging Paradigm for Safe AI Deployment', *ArXiv:2201.05159 [Cs]*, 13 January 2022, http://arxiv.org/abs/2201.05159.

Risk disaggregation

One of the difficulties of governing GPT-2 and GPT-3 is that a single model can be used in so many different ways. Miles Brundage, a member of OpenAI's policy team, wrote in his PhD thesis: 'The central dilemma of the GPT-2 case is that the very same language models can be used for creative, commercial, scientific, or malicious purposes.'[213]

One response to this dilemma could be to prioritise the beneficial applications and open source the model. This approach was favoured by some of my interviewees; for example, JESSIE said: 'I think it would be wrong to censor yourself just because there's a possible negative impact, if you can also argue that there are positive ways to use your technology.' Similarly, some researchers assume that the beneficial applications will outweigh the harmful ones.

Underlying these responses is the assumption that we are asking the binary question of: should the model be open sourced or not? The staged release of GPT-2 took one small step towards removing that assumption, demonstrating that models of different sizes could be shared at different times. This is an attempt at risk disaggregation: some of GPT-2's capabilities are present in the larger models but not the smaller models. However, it is very crude: the smaller GPT-2 models are, generally speaking, less performant across a wide range of tasks – regardless of their potential social impact. As I wrote elsewhere, 'manipulating model size is a very imprecise method for controlling how the model is used.'[214]

---

[213] Miles Brundage, 'Responsible Governance of Artificial Intelligence: An Assessment, Theoretical Framework, and Exploration', *ProQuest Dissertations and Theses* (Ph.D., Ann Arbor, Arizona State University, 2019), 30, ProQuest Dissertations & Theses Global (2344721343), https://search.proquest.com/docview/2344721343?accountid=13042.

[214] Shevlane, 'Structured Access'.

The GPT-3 API is a game-changer because the question is no longer: which models should be open sourced and which models should be held back? No longer must the issue of social impact be considered at the level of whole models. The API allows the lab to make within-model distinctions between different applications. The level of analysis shifts downward and the question is now: does this specific application seem beneficial or harmful? As we saw above, the process is now to look at applications proposed by third party developers and evaluate them on their merits. The third party developers now become part of the system through which the lab tries to filter out the harmful from beneficial uses of the model, in that they bring their proposed application to OpenAI and submit it for approval.

Sometimes a user of the API will not be building an actual application, but directly using GPT-3 in a more informal way. CAMERON explained that in such cases it can be difficult to draw distinctions between what should be permitted and would should not be permitted:

> One thing we've done about the Liam stuff [see above] and other things is try to communicate a lot more to our developers about what is and isn't appropriate. Because it's like, you try and write this in terms and conditions, right? It's really fucking hard to legally do the difference between, say, you, Toby, have access to GPT-3, you occasionally show it to some people at FHI, you occasionally show it to friends, you occasionally quote it a bit in a research paper, that's totally fine. But then there's a version of it, where you go and get a talk in London, have GPT-3 up behind you, and are like: *we're going to ask it about Mohammed now*. And we'd be like, *God, well we don't want you to do that.* And trying to write legally binding things that let you do the former and not the latter is really hard.

In other words, the limiting factor becomes the lab's capacity to articulate and communicate distinctions between permitted and non-permitted uses, alongside the lab's capacity to monitor for compliance. Notwithstanding the bureaucratic challenges involved, it is important to step back and notice how far the API takes us away from making model-by-model assessments of the potential harms and benefits. In fact, CAMERON's statement of the difficulties of making low-level distinctions can be re-read as a statement of how much potential control the lab now has. The API brings into view a whole new world of low-level distinctions between different kinds of uses.

Importantly, this new level of monitoring and control means that the risks of GPT-3 can be considered in a more elaborate way. Recall that OpenAI's discussions of the risks of GPT-2 centred around the proliferation of fake text across the internet (social media bots, catfishing, and spam). Since the GPT-3 API opens up a new level of fine-grained control over GPT-3's applications, it opens up a whole new space of risks to be considered. This is reflected in GPT-3's use case guidelines, which are very detailed. For example, here are some applications that were not problematised around GPT-2 but are now deemed risky in the GPT-3 guidelines:

1. Classifying tweets by political sentiment;
2. Classifying people based on protected characteristics;
3. Search engine optimisation for articles and blogs;
4. Summarising content that 'may be sensitive politically, economically, medically, or culturally';
5. Extracting personal information from data;
6. Medical diagnosis.

My point is that crude levers of control (such as the lever of when to share a bigger model) invite crude discussions of risks, whereas fine-grained levers of control invite fine-grained discussions of risk. After all, the point of the discussion about risk is to inform the KCR. If the KCR can only address risk in a low-resolution way, then there is little point in feeding it a high-resolution picture of the risks. The API therefore creates demand for a more elaborate picture of the model's risks.

Empirically, the differences in discourse between GPT-2 and GPT-3 might be confounded by other factors. Two confounders could be: (a) GPT-3 is more capable and so has a wider range of possible risks (although GPT-2 could likely have performed many of the applications listed above); and that (b) OpenAI has had more time to consider the risks of language models. But if my point is nevertheless correct, it lends weight to the general argument, introduced in the introductory chapter, about how actors decide which risks to pay attention to. They first consider what risks they can actually do something about, and that strongly shapes subsequent discussions of risk.

Iteration

The Collingridge dilemma says that knowledge about the social impacts of a technology normally arises only after the technology's deployment has become irreversible.[215] The dilemma is partly about the pace at which knowledge about risks is produced, and partly about the irreversibility of technological deployment. The GPT-3 API tackles both aspects of this problem, and does so more successfully than the staged release of GPT-2.

[215] David Collingridge, *The Social Control of Technology* (London: Frances Pinter, 1980).

First, the API allows the lab to see how the model is being used. This is difficult when the model is open source – a regime that has no mechanism for the lab to track who is using the model for what purposes. Hence with GPT-2, OpenAI had to rely on internet-based research to see traces of GPT-2's use. In contrast, with the API, we have seen that both (a) the lab can analyse what kinds of content the model is being used to produce; and (b) application developers must submit their plans to OpenAI before launching. The API puts the lab in a much stronger informational position. The API might not reveal the social impacts of the applications, but knowing about the applications' existence and prevalence is a good place to start.

Second, the deployment of the model is reversible. Users have no local copy of the model and so rely upon continued access to the API. If it comes to light that a specific application of GPT-3 is socially harmful, OpenAI can prohibit such applications going forward. That said, prohibiting a previously permitted application area would be disruptive to OpenAI's customers operating in that area, which provides a disincentive for them to do so. In practice, the likely resolution might be that the application developer agrees to make some alteration to their service, as we saw with AI Dungeon above.

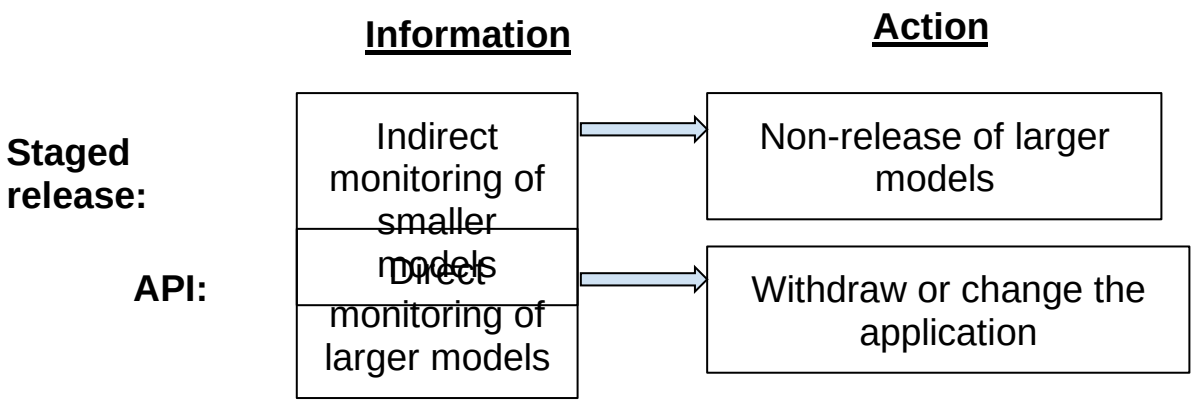


Figure 2: The API gives the lab better information, and a stronger mechanism for intervening.

Both the API and the staged release take an iterative approach. With staged release, the logic is that the lab can learn about possible misuses of the larger models by observing how the smaller models are used. This is not perfect because the larger models can be misused in ways that the smaller models cannot – otherwise there is no point in delaying their release. Because the release of each model is irreversible, the staged release approach involves going step by step: releasing a model, watching for misuse, and then taking another step into the unknown by releasing a larger model.

The API is very different. Each step is reversible, and so the lab can immediately grant access to the largest model and later refine what uses are permitted. Hence, the usage guidelines are a 'living tree' (to use a legal phase). When OpenAI released the guidelines, they tweeted: 'As we gain more experience operating the API in practice we expect to expand and refine these categories.' More generally, the CEO of OpenAI, Sam Altman, said in a panel discussion in 2021:

> [S]afe AI systems require iterative deployment. There is a belief in some efforts in the world that you should build a super powerful AI in a vacuum, and then once you're totally sure it's safe, you should push the button and launch it. I do not think that will work . . . And so, I think safe, iterative deployment, that is always under control, and the ability to pull it back, is how we're going to get to these long-term safe systems.[216]

[216] Global Emerging Technology Summit, 13 July 2021. Available at: https://www.youtube.com/watch?v=MkJs-eRPABg&t=29546s&ab_channel=NSCAI

Similarly, when discussing the AI Dungeon example, ROBIN told me that, while OpenAI had been "ahead of the curve" on monitoring for political content, this case was something they were "somewhat surprised by". He said: 'There's no clear way to predict all the sorts of things that will arise in advance. So, I think that's one reason why taking a sort of iterative approach is good'. The API makes this possible.

3.3 Information vs tools

We have seen that the API approach is better equipped for preventing misuse than staged release. Where did these new powers come from? I argue that from GPT-2 to GPT-3 there has been a paradigm shift, where the GPT-2 KCR treated the model as information, whereas the GPT-3 KCR treats the model as a tool.[217] For the purposes of this argument, 'information' is interchangeable with 'knowledge', and 'tool' is interchangeable with 'artefact', 'technology', or 'machine'.

An AI model, like any software, is simultaneously informational and a tool.[218] The code and parameter values are a set of instructions that can be stored, copied, and shared as information. And yet, on the right computer, the model is a tool that can be used for practical ends – e.g. to sort people into groups, or to write a poem. The 'open source models' KCR treats models as information, by allowing people to download a copy. The API is fundamentally different because people do not download the model at all. The model is stored in the memory of OpenAI's GPUs, and people interact with GPT-3 as a tool.

This is important because information has properties that make it difficult to govern. Models like GPT-2 and GPT-3 are dual use, and so the lab needs to construct some kind of filter that

[217] See also: Shevlane, 'Structured Access'.

[218] Timothy R. Colburn, 'Software, Abstraction, and Ontology.', *The Monist* 82, no. 1 (1999): 3, https://doi.org/10.5840/monist19998215.

blocks misuse while letting through other applications. If the model is treated as information, the obvious kind of filter is one that gives people some of the underlying information but not the risky part. This trick works well for factual information. For example, a sniffer dog allows police officers to tell whether a bag contains certain drugs without revealing the otherwise private contents of the bag. This is the classic example for the concept of 'structured transparency', which is about finding clever ways of communicating necessary information whilst keeping closely related information private.[219] Privacy-preserving analysis of health datasets follows the same logic, where the analyst is not given direct access to personal health data and yet can extract population-level insights.

However, AI systems are different from factual information. The problem is *not* that some identifiable subset of the AI system has harmful applications. The problem is that the AI system, as a whole, has a range of applications, and some of those applications will be harmful. Therefore, it is not usually possible to break the AI system down into smaller parts in a way that neatly tracks normative distinctions between different applications. This difficulty is illustrated by Microsoft's DialoGPT model, which was a GPT-2-inspired chatbot. Microsoft shared most of the model, but left out the code that allows users to actually retrieve the model's outputs. This was in response to their concern that the chatbot was prone to offensive outputs. However, as I wrote elsewhere:

> This is no solution to the mixed nature of the model's outputs, because either users cannot get the model to produce any text whatsoever, or they find substitute versions of the missing code online and thereby have access to the full functionality of the model. Perhaps users who are determined enough to do the latter are on average more

[219] Andrew Trask et al., *Beyond Privacy Trade-Offs with Structured Transparency* (arXiv:2012.08347 [cs.CR], 2020).

> likely to use the model responsibly, but even so, this is a very imprecise and insecure method of controlling how people use the model.[220]

As I put it, the problem is:

> there exists no sweet spot where the user knows enough to achieve only the beneficial ends. Either they know too little, and so cannot use the technology at all, or they know too much, and so they have too much leeway.[221]

This explains why the GPT-3 API is better than 'staged release' at *disaggregating* different applications (as I argued above). Staged release treats the family of GPT-2 models as information, and selectively discloses some of that information. In this way, it is an outgrowth of the 'open source models' KCR. The API is different because it treats GPT-3 as a tool that can be operated from afar. The lab can then place restrictions on how users operate the tool, as we saw above.

The information vs tools distinction can also explain why the API is better suited to *iterative* deployment than staged release. Information is difficult to revoke. Once it has been communicated, it is difficult for the donor to change their mind and wipe the memory of the recipient. One counterexample is that Amazon has the ability to remove books from users' Kindles, as they demonstrated when they recalled *1984* and *Animal Farm* from users' devices after realising they had no legal right to sell the books.[222] However, this is only possible because the software running on Kindle devices grants Amazon that power. When somebody

[220] Shevlane, 'Structured Access'.
[221] Ibid.
[222] Ian Kerr, 'Digital Locks and the Automation of Virtue', in *From 'Radical Extremism' to 'Balanced Copyright': Canadian Copyright and the Digital Agenda*, ed. Michael Geist (Irwin Law, 2010), 247.

downloads a model from GitHub onto their computer, the lab that trained the model does not have this kind of power to reach into users' computers. Even in the Kindle example, the information is still not entirely revocable: users can find ways around the DRM (digital rights management) on their ebooks and save the files elsewhere, out of Amazon's reach. With GPT-3, users have no local copy, and so cannot keep GPT-3 away from OpenAI's reach. Their access to the tool can be shut off, analogous to how somebody's keycard access to a building can be terminated.

The same information vs tools distinction also helps to explain why OpenAI has focussed on governing models rather than the insights contained in research papers. The GPT-3 paper, and related papers about 'scaling laws' (i.e. measuring how much a model's performance increases with more data, compute, and parameters),[223] have been controversial in some quarters. As we saw in section 2, some AI researchers concerned about the risks of AGI have been critical of OpenAI's decision to publish these papers. The concern was that the papers illuminate a clear direction for progress towards building AGI, and therefore speed up AGI timelines without equivalently advancing knowledge of safety techniques. As I described above, Connor Leahy described the publication of these papers as 'one of the most dangerous things anyone has ever done'. OpenAI staff themselves took these concerns seriously before publishing the papers, and I was told that both GPT-3 and the scaling laws papers were delayed.

However, scientific insights are difficult to govern. It is hard to predict what other actors will do with the information, and it is even harder to subsequently influence or control what they do with it. ROBIN explained:

[223] Kaplan et al., 'Scaling Laws for Neural Language Models'; Henighan et al., 'Scaling Laws for Autoregressive Generative Modeling'.

> We try to, as a matter of course, to ask those sorts of questions of: what are the likely impacts [of publishing a paper]?[224] But I'd say, roughly [the] shape of a lot of those conversations is that there are a lot of abstract arguments that can be made one way or the other about the potential upsides and downsides, but it's often hard to adjudicate them at that level of abstraction. And sometimes you can look to analogues of, "here's how this similar paper was received," or "here's what happened last time," or whatever. And that can somewhat reduce uncertainty. But at the end of the day, it's again still hard even with analogues; every paper is unique in some respects. So, sometimes it comes down to just weighing different variables that are very hard to compare.

The lab has much less control over how other actors react to a paper's insights, in comparison to the fine-grained control over model usage bestowed by the API. The criticism of the API would be that it is like the drunk man looking for his keys under the streetlight, because that is where he can see, despite having dropped them in another location. The API is good at what it does, i.e. preventing direct model misuse, but some AI researchers would argue that most of the impact of GPT-3 comes from the insights about scaling.[225] GPT-3 and the scaling laws papers play directly into the geopolitics of semiconductors, demonstrating the importance of hardware for progress in AI capabilities. And as we saw above, multiple

[224] ROBIN listing considerations: 'Some of the relevant considerations there, which we did think about include: making deep learning more scientific (and which potentially be beneficial from a safety perspective). Versus people just throwing whatever they can at an experiment, knowing roughly what loss to expect is a move in the direction of more rigour. There's also potentially incentivizing certain actors or making certain actors, including governments, more likely to invest in a lot of compute, and accelerating racing. There could be updating of people's timelines and stuff like that.' ROBIN also mentioned that the authors of such papers have a career interest in them being published.

[225] As Connor put the point: 'So there's a saying about the nuclear bomb, that the the only secret about the nuclear bomb was that it was possible. And I think that applies to GPT-3 too. The scary thing about GPT-3 is not GPT-3 itself . . . The problem is the scaling laws. . . . I think 99.9% of the damage of GPT-3 was done the moment the paper was published.'

Chinese research groups have followed OpenAI in building GPT-3-like models. The level of control that the lab has over GPT-3 is therefore a mixed picture: OpenAI has fine-grained control over the model itself, but no control over the fundamental ideas. This links back to my wider point that, in my case study, we generally observe an *opportunistic* kind of governance. Labs like OpenAI govern what they can, subject to structural constraints. The governance regimes I study have all arisen on islands where the lab (or, in chapter 3, the conference organisers) have some agency, surrounded by a sea of anarchy. Nevertheless, I would still maintain that the API is a more powerful governance tool than what has gone before.

5. The relationship between science and society: open source vs API

5.1 The agency of researchers

The API allows the lab to extend its tentacles out into society. As JAN, an NLP researcher, put it: 'they take responsibility for how everyone will use their model, which is a very strong responsibility.'

Such an assumption of responsibility is not possible under the open source regime. It is no coincidence that, of the different KCRs found in this thesis, the GPT-3 API is both (a) the strongest assumption of responsibility by a lab over the social impact of its models, and (b) the biggest departure from the open source regime. The API brings into view 'downstream' actors like application developers and end-users, giving the lab a platform for governing them. Under the open source regime, this part of the pipeline (model deployment and use) is hidden from view and difficult or impossible to influence. The open source regime pigeonholes AI researchers as simply producers of knowledge and artefacts, narrowing the role that they can play in AI governance.

When a researcher open sources a model, they abdicate control over how that model is used. For researchers who would prefer not to think about social impact, this is convenient: ignorance is bliss. As we saw in chapter 1, AI researchers sometimes dismiss questions of societal impact by pointing to their own lack of agency. Recall the quotation from FRANKIE, an NLP researcher:

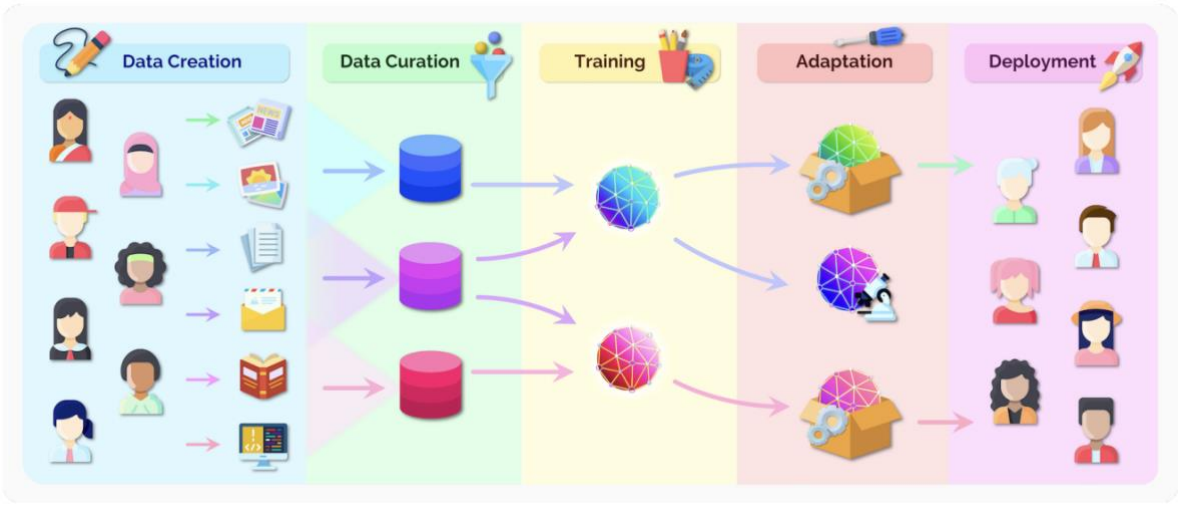
> What's the [solution] for models being used in ways that we don't like, whatever that way is? It's to stop the use. That's not something that the research community can possibly address. That's something that's external, like that's where the thing gets used. That's not something that researchers choose or do anything about. It's a policy problem.

The OpenAI API moves this boundary between what's internal and external to the research community, by capturing the territory of 'where the thing gets used'. The same ambition – of breaking down barriers between research and real-world use – is sometimes voiced within the AI research community, especially by researchers interested in AI ethics.[226] However, the assumption is normally that the wall around the research community exists *inside the minds of researchers*. The remedy is therefore to encourage mainstream AI researchers to think more about social impact – see, for example, the broader impact statements discussed in the next chapter. Such initiatives work *within* the existing publication and open source regime. Instead, I would argue that the barrier between research and 'where the thing gets used' is a real one. The open source regime directly creates a division of labour between AI researchers

[226] Karën Fort and Alain Couillault, 'Yes, We Care! Results of the Ethics and Natural Language Processing Surveys', in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (LREC 2016, Portorož, Slovenia: European Language Resources Association (ELRA), 2016), 1593–1600, https://aclanthology.org/L16-1252.

and the actors deploying AI systems. This regime provides little room for AI researchers to have any agency over the social impact of their technologies. Without grasping this, even very well-intentioned AI researchers and ethicists end up making proposals that are unlikely to seriously affect AI's social impact.

As an example, consider the paper *On the Opportunities and Risks of Foundation Models*, from academic researchers at Stanford University.[227] Their Figure 3 sketches the relationship between AI researchers training foundation models and wider society:[228]



Society comes into the picture as the people who originally created the training data, and as the people affected by deployment. The authors state: 'Appropriately, as we're interested in social impact, *people* occupy both ends of the pipeline.'[229] This is an honest attempt to build a more society-focussed vision of the AI research process. However, I would argue that the envisioned science-society relationship is impoverished. No doubt, students of Bruno Latour will be quick to point out that the researchers are 'people' too, but that is not my focus. My focus is: what agency do the AI researchers have?

[227] Bommasani et al., 'On the Opportunities and Risks of Foundation Models'.
[228] Copied with consent from the lead author.
[229] Ibid, p.8

The implication from the *Foundation Models* paper is that researchers should focus on design choices, such as what data to include, or how to train the model. The authors note that 'many researchers' and practitioners' purview is restricted to the training stage', and so offer the mantra, **'Think ecosystem, act model**.'[230] The authors acknowledge a difficulty here, that the model 'can be adapted to many downstream applications, sometimes by an entirely different entity for unforeseen purposes'. The authors suggest a solution: researchers should evaluate the model's performance on tasks related to different downstream applications, and write this up so that others can see it. The researcher's opportunity, then, is to focus on (a) how they build the model, and (when that fails), (b) how they evaluate and document the model.

This approach works best for a certain category of risk: risks attributable to either (a) a defect in the model, or (b) a defect in users' understanding of the model's capabilities. The archetypal risk here is bias, i.e. that models absorb racial, gender, or other stereotypes, and therefore, when used in decision-making processes, make biassed decisions. Another concern is that models have better knowledge of certain cultures and certain languages, such as Western culture and the English language. Another is that users will overestimate the capabilities of a model and therefore use it for invalid or inappropriate applications, such as screening candidates for a job. In such cases where the risk comes from a defect in the model or users' understanding thereof, the role of the researcher is then to remedy these defects, either through better model-building or better documentation. The user's behaviour can be positively influenced by giving them better tools: a model with fewer limitations, and better information. For example, one avenue of research, aiming to make language models more

[230] Ibid, p.9

ethical, involves training multilingual models using text data from many different languages.[231]

The problem is that many AI risks cannot be addressed like this. For instance, the category of misuse risk is left unaddressed. AI technologies like language and vision models can be used for a variety of purposes, and it is normally difficult or impossible to build the model in a way that discriminates between these. Even including minority groups in the training data can backfire. A 2018 paper published by Wiley from a group of Chinese researchers was very controversial: it was about building face recognition systems that identify Uyghur people.[232] There are media reports of the Chinese state using similar systems to surveil the Uyghur minority population.[233] So, should researchers train their models on images of Uyghur faces or not? It is difficult to see how any distinction between ethical and unethical uses of these models could be erected and enforced inside the code of the model itself. Similarly, improved documentation is also poorly suited to misuse risk. If researchers clarify, for example, that a language model is excellent at producing fake news, and identify its areas of weakness, this hardly puts bad actors in a worse position. More information is not always better.[234] The mantra of 'think ecosystem, act model' has serious limitations. And yet it finds favour among AI researchers, I would argue because they have no way of shifting to 'think ecosystem, act ecosystem' without fundamentally changing the KCRs of the field.

[231] BigScience, 'BigScience Large Language Model Training Launched', accessed 12 April 2022, https://bigscience.huggingface.co/blog/model-training-launched.

[232] Richard Van Noorden, 'The Ethical Questions That Haunt Facial-Recognition Research', *Nature* 587, no. 7834 (18 November 2020): 354–58, https://doi.org/10.1038/d41586-020-03187-3.

[233] 'AI Emotion-Detection Software Tested on Uyghurs', *BBC News*, 25 May 2021, sec. Technology, https://www.bbc.com/news/technology-57101248.

[234] Nick Bostrom, 'Information Hazards: A Typology of Potential Harms from Knowledge', *Review of Contemporary Philosophy*, no. 10 (2011): 44–79.

Similarly, 'structural risk' is another category of risk that is poorly addressed by 'think ecosystem, act model'.[235] Some scholars have raised concerns that AI will have adverse structural effects on society. For example, one concern is that authoritarian governments will become more stable, meaning that oppressive regimes persist for longer.[236] Another concern is that AI will lead to mass unemployment, which could be destabilising for democratic politics.[237] The CEO of OpenAI recently tweeted:

> it makes me very uncomfortable that so much new technology (especially ai, not trying to evade blame) naturally benefits authoritarian governments more than democratic governments.
>
> have to fight this hard.

As with misuse, it is difficult to see how researchers could avoid any such risks using the limited tools provided by the existing publication and open source regime. Again, the lack of agency that AI researchers have in this area is evident from their proposals for how the AI community should be acting. For example, their discussions about the impact of AI on power in society often gravitate back to a narrower topic: power *within* the AI research process. For instance, Pratyusha Kalluri, an AI researcher at Stanford University, wrote a commentary in Nature titled: *Don't ask if artificial intelligence is good or fair, ask how it shifts power*. The article opens with a description of how law enforcement agencies and other actors are using AI to 'monitor and predict our behaviour', and poses the question: 'how is AI shifting

[235] See Remco Zwetsloot and Allan Dafoe, 'Thinking About Risks From AI: Accidents, Misuse and Structure', Lawfare, 11 February 2019, https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.
[236] Allan Dafoe, 'AI Governance: A Research Agenda', July 2017, https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf.
[237] Ben Garfinkel, 'Is Democracy a Fad?', *The Best That Can Happen* (blog), 26 February 2021, https://benmgarfinkel.wordpress.com/2021/02/26/is-democracy-a-fad/.

power?' Nonetheless, the suggested proposals all relate to power dynamics *within* AI research rather than the wider world, such as: (a) increase the pay of workers who label data used in AI research; (b) where people give their data for use in AI research, allow them to opt out; and (c) discuss how AI shifts power in research papers. Though laudable, these proposals do not come with a clear explanation for how they might actually change the structural impact of AI on power in society. It is not clear that even a maximally egalitarian AI research community would be able to build AI systems in such a way that they are power-distributing, rather than power-concentrating, in society. The shaky assumption is that the technology is so flexible that it will absorb whatever values were present at its invention and then propagate those same values throughout society. This example is very similar to 'think ecosystem, act model'. In both cases, the ambition is to think on a societal level, but then the researchers restrict themselves to what they can actually affect – at least, given the existing separation between research and deployment.

Another similar example comes from a professor at Stanford, part of a research group that aims to train large language models like GPT-3. In a talk from 2021 (available online) he said:

> I was quite influenced by reading Phil Rogaway's article a few years ago on the moral character of cryptographic work, which I would highly recommend. He talks about cryptography as an inherently political tool, with massive social, structural consequences. But the exact same thing could be said about AI. So technology is not neutral. What kind of technology we develop can have a directional impact on society. And we need to work with social scientists and humanists to figure out what kind of

> society we really want to build. With this in mind, how should we change foundation models?[238]

He continues by offering a way forward: to train models in a more distributed way. For example, the compute from training the model could come from many different sources, linked together. This would have 'huge . . . implications on who could even train these models'. Again, the assumption is that – if AI is an inherently political technology – it will absorb the politics of the social context in which it was built.[239] It is interesting to see how the API allows for a completely different approach. The API allows the lab to, in effect, go out into society and manage how the model is being used. This makes these other strategies, such as training models in a distributed way, look very indirect and under-theorised in comparison.

One final example of the same phenomenon. AI researchers sometimes seek refuge in the idea that the negative impacts of AI are concentrated in specific kinds of AI, such as computer vision systems. The hope is that they can then avoid working on those areas, and their work will be more socially beneficial as a result. For example, JAN told me:

> I've not worked on computer vision. Because I think facial recognition is actually a very bad technology. And that's the main problem with computer vision. It's that, today, it's really one of the biggest applications – I think the most used application of computer vision – is facial recognition.

[238] The talk is available on YouTube (see around 25 mins): https://www.youtube.com/watch?v=dG628PEN1fY&ab_channel=StanfordHAI

[239] Aside, one interpretation could be: these researchers are conflating their own interests, as academics who do not have the resources to train GPT-3-sized models, together with the interests of individuals in society who might be negatively impacted by AI technologies.

However, this position often involves making distinctions between different types of research which are, in practice, very interconnected. A recent trend within AI research is that different subfields are converging. This is predicted by the *Bitter Lesson* essay (above), in that, as the field moves towards more generic methods, with less tailoring to specific tasks, the methods used for (say) computer vision and NLP will become similar. Indeed, the transformer architecture that has become very popular in NLP also became popular in computer vision and other areas. In 2021, one AI researcher tweeted the following meme with the caption, 'the ML community and Transformers':

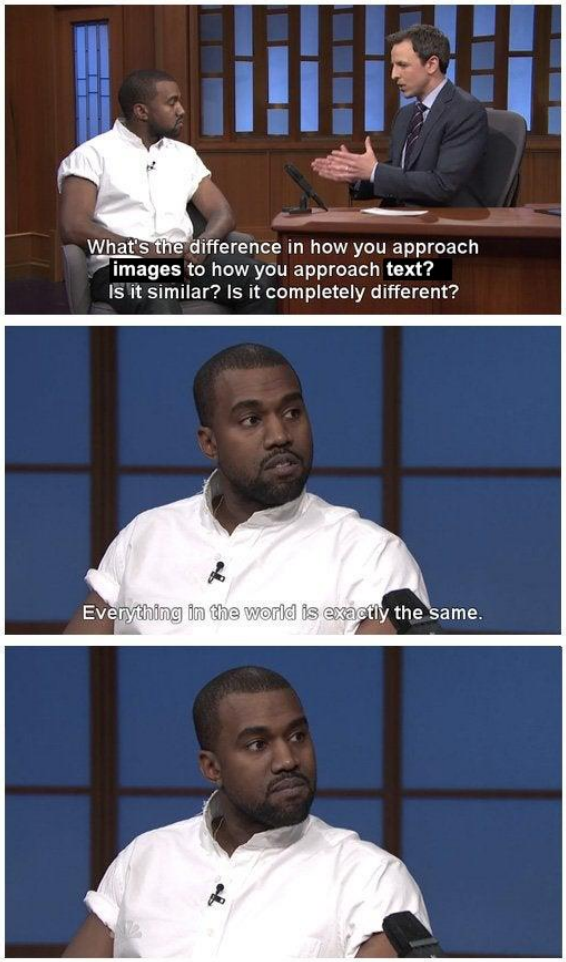


Figure 3: The joke is that AI researchers apply the same methods to different classes of problem.

In a world where ‘everything in the world is exactly the same’, it is becoming increasingly difficult for AI researchers to target their social impact by working on a particular subfield of AI capabilities. Above, we saw that the GPT-3 API excels at *risk disaggregation*: separating out different applications of the model, allowing some to go ahead and blocking others. This kind of intervention works because it occurs at the deployment stage, where the AI system is applied towards specific ends. Innovation in fundamental model-building methods normally has a much messier relationship with social impact, because those methods are compatible with a variety of AI technologies and applications.

Stepping back, the problem is that researchers who want to steer the societal impact of AI have few effective tools for doing so, because their role, under the existing publication and open source regime, is just to make research. As such, they can try to make changes to how they do research: what area they focus on, how they obtain the training data, and how they write up their results. But these actions have no strong mechanism for steering AI’s social impact. Again, we see the proverbial drunk man looking for his keys under the streetlight, despite having dropped them elsewhere. The fundamental problem is researchers’ lack of agency under the open source regime. This is a key reason why the API is so significant: it adds a powerful toolbox for the research lab to shape the social impact of its technologies.

5.2 The agency of governments

One criticism of OpenAI’s API has been that it centralises power in the hands of the lab. In a sense, this is spot on. As we have seen, the effectiveness of the API as a governance regime depends on its ability to influence and control what people do with the model. This is no doubt a kind of power. However, I argue that governments would be in a better position to regulate AI in a world where the most powerful AI systems are behind APIs rather than open source. This is a significant point, because if true, it suggests the API increases the overall

governability of AI, rather than just shifting power to AI labs. It is part of my wider argument that the open source regime makes it difficult to govern AI in the interests of society.

Institutional and technical structures like the API will give state regulation something to grasp onto. Rarely does state regulation invent, from thin air, the underlying basis for its implementation. As Foucault put it, the 'state can only operate on the basis of other, already existing power relations'.[240] Therefore, the API should not be viewed as handing power to the lab *instead of* giving it to democratic governments. Rather, the API sets up a system where it is possible to make enforceable distinctions between applications of AI that are allowed and not allowed. If states step in and regulate how AI developers grant access to their AI systems, structures like the API will be helpful. For example, state regulation could specify that AI developers like OpenAI must not allow their AI systems to be used for misinformation campaigns; or state regulation could demand that AI systems on the API service meet certain safety standards.

An early example comes from the proposed EU AI Act. This draft legislation would impose certain duties on the 'providers' of AI systems. A provider is defined as an actor that 'develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge'. Consider the scenario where a research group trains a large language model, open sources it, and then another company builds an application using a fine-tuned version of the model. The 'provider' imagined by the draft legislation is split between two actors. The risk management practices in Article 9 are supposed to 'run throughout the entire life-cycle' of a high-risk AI system, which would surely include the pretraining phase. Similarly,

[240] Michel Foucault and Gordon, *Power/Knowledge : Selected Interviews and Other Writings 1972-1977* (New York: Pantheon Books, 1980), 122; see also Michel Foucault, *Discipline and Punish : The Birth of the Prison*, 1st American (New York: Pantheon Books, 1977), 27.

the data governance rules in Article 10 refer to the collection of the training data, which would naturally include the pretraining data for a large language model. And yet the research group would not count as a 'provider', because it does not aim to actually put the model into service itself. These obligations therefore have no actor to latch onto. This is a product of the fact that the open source regime allows for so much important AI development work to be done under the institutional heading of 'research', even where the artefacts created during such research become central components within a plethora of different real-world applications.

Now consider the same scenario, except that the prevailing KCR in the AI research community is API-based. Either research groups have their own APIs, or they upload their model to third party repositories that host them via APIs. If the research group views the model as simply a research project, they can simply grant API access to other researchers and *not* companies looking to apply the model. Then, the draft legislation would not be relevant. If instead the research group does want the model to be built into applications, they can either (a) limit those applications to 'low risk' ones, reducing their legal obligations, or (b) follow the obligations of the legislation, even for 'high risk' systems, e.g. by being more careful about what data the model is trained on. In this scenario, it is much easier to see how the vision of the legislation's drafters maps onto reality. The API regime organises the different actors (research groups; application developers) in a way that is easier for the law to regulate – there is a cleaner separation between research and 'providers' of AI systems.

In addition, enforcement also looks easier in the API world. In a case like GPT-3, there is a provider (OpenAI, in this case) that has a bird's eye view of all the different ways in which the model is being used by thousands of other, smaller providers. Where the draft legislation prohibits certain applications – such as AI systems that exploit the vulnerabilities of elderly

or mentally disabled people – it is a much simpler task for the regulator to ensure that OpenAI does not allow any such applications than to conduct a broad-based market surveillance effort.

In other words, labs like OpenAI, in taking greater responsibility over how their AI systems are used, are not necessarily just grabbing power for themselves, but putting themselves into a position where democratic governments can better regulate AI. Indeed, my interviewees from OpenAI welcomed government regulation. For example, CAMERON told me:

> You don't want small companies [like OpenAI] making decisions about what should and shouldn't be done with tech, but it's something we kind of have to do in the absence of any decent regulatory regime or norms.[241]

My claim has been that a move towards APIs, and away from open source, would better enable democratic governments to regulate AI. This analysis turns a certain narrative about 'democratisation' on its head. It is common for AI researchers, and those in the AI industry more generally, to use the term 'democratisation' simply as another word for wider proliferation of the technology. It is about a larger group of actors being able to use AI technologies.[242] The background assumption is that the use of an AI system is empowering,

[241] Similarly, in response to the controversy over Clearview AI, a company that mass collects images of people's faces and identities to build a facial recognition system, a senior member of OpenAI's policy team tweeted: 'Absent quick action by the federal government (cautiously optimistic on this...), there will soon be "Clearview for X" for a wide variety of X's--powerful AI capabilities w/ few if any constraints. Voice identification, gait recognition, astroturfing, etc. . . . Sufficiently cheap hardware and powerful algorithms exist to do extremely Big Brother-y stuff at ~global scale already for not that much money. All the more reason to set guardrails for such things ASAP.'

[242] See, for example: Mark Riedl, 'AI Democratization in the Era of GPT-3', *Medium* (blog), 25 August 2020, https://mark-riedl.medium.com/ai-democratization-in-the-era-of-gpt-3-8b91891f91cb.
According to Reidl, democratisation involves:

- 'Having access to powerful AI models
- Having access to algorithms
- Having access to computing resources necessary to use algorithms and models
- Being able to use the algorithms and models, potentially without requiring advanced mathematical and computing science skills'

and wider proliferation will spread that power more broadly. For reasons already discussed, this assumption is shaky. AI systems may well empower their users, but those users will often be actors like law enforcement agencies who already have a lot of power. At the same time, notice what does *not* come hand-in-hand with wider proliferation: the collective ability of the population to set rules on what kinds of AI should, and should not, be deployed in society. The open source regime does not give each individual in society a vote over the future of AI. It actually makes this more difficult.

If my argument in this section is correct, then the API – at least in democratic countries – paves the way towards a more democratic alternative to the open source regime. By 'democratic', here I mean 'facilitating collective decision-making over important decisions'. We should no longer draw a naive connection between open source and democracy, and (insofar as anyone does this) between the API and authoritarian power. Instead, if we must use political imagery, I would argue the democratic vision of AI deployment, though not yet realised in practice, would come from democratic oversight and control of APIs (and other AI services).

Finally, I want to end by bringing the analysis back to James Scott's *The Art of Not Being Governed*. There is an irony here. For the hill people studied by Scott, a lack of writing is, he suggests, one of the strategies for avoiding the reach of states. Avoiding writing things down is a form of closedness, shielding the culture from outside inspection from nearby states. In AI research, openness achieves the same ends but through the opposite strategy. Instead of protecting knowledge by concealing it, scientific papers and open source models are a kind of denial of ownership. Everybody owns the knowledge, and so nobody does. This goes hand-in-hand with the claim that risks from AI are collective risks, to be dealt with by society,

rather than residing with the AI research community. The open source regime means that AI researchers – despite occupying an important part of the AI development and deployment pipeline – control very little, and so have little to offer to regulators. In contrast, the API regime has been interpreted by some people as a move towards greater 'closedness', because the underlying model is not visible. But for the purposes of regulation, it is not closed at all. It offers a tangible structure that could form the basis of state intervention.

6. The API and the relationship between industry and academia

In this section, I will argue that with the trend towards large models, an important distinction is emerging between *building* these models and research seeking to better *understand* them. Academics are well-placed to do the latter kind of research, which also has a claim to being useful for reducing risks from AI. The key question is then whether the API facilitates academics to do this kind of 'understanding' research on GPT-3.

6.1 The shifting relationship between academia and industry

GPT-3 has added to a feeling among academic AI researchers that there are certain areas of research where they cannot compete with the industry labs. Academic research groups generally do not have nearly enough compute to train GPT-3, nor the time to allocate to solving all the engineering problems that come with training such a large model. Even if they did, it would be difficult to justify an investment on the order of $10m for a single paper. CAMERON recounted a conversation with the 'head of engineering' at a North American university:

> But also they don't have the incentive. He was like: *look, we kind of ammortise costs over papers. It's very hard to put 10s of millions of dollars or millions of dollars into something which yields one paper.*[243]

The share of papers at AI conferences like NeurIPS and ICML authored by Fortune 500 companies has climbed from roughly 10% in 2000 to roughly 30% in 2019.[244] Within this, large technology companies like Google, Microsoft and Facebook contribute the most papers. The question is how the industry labs coexist alongside academic AI researchers. KAHL told me that, increasingly, there are areas of research where academics cannot compete:

> So, back in 2016, when I was considering: should I be an academic or should I go to Google or DeepMind or OpenAI? I was trying to figure out how much of an impact the compute resources would have on the sorts of research you could do. . . . At one extreme, maybe it has no impact. Versus: maybe you kind of build the same systems, but the industry ones work better because they have more resources. All the way up to: there are kinds of research that just don't make sense to work on unless you have very large amounts of compute – that they're just things that companies can do conceptually that like we can't do in academia. And five years ago, we were near the lower end of that scale, that academics and industry labs were basically building the same systems, except the ones in industry were larger, they were trained on more data, and they got better performance for that reason. And just in the last couple years, it started to look more like there are kinds of research you can do in the industry that are qualitatively different from what we can do in academia.[245]

[243] Recall my argument from chapter 1 that the existing publication system incentivises incremental work.
[244] Daniel Zhang et al., 'The AI Index 2021 Annual Report', *ArXiv:2103.06312 [Cs]*, 8 March 2021, 38, http://arxiv.org/abs/2103.06312.
[245] The interview was in 2021.

GPT-3 is an extreme example of this phenomenon. When the GPT-3 paper won a Best Paper Award at NeurIPS 2020, one researcher complained on Twitter that, although it was ‘an incredible piece of work’, ‘no academic lab could have won that award for that work’. FRANKIE, an NLP researcher, described the impact that GPT-3 has had on the community:

> I've been in the research community for about a decade, and the entire time there has been this idea that industry has more resources and can do bigger things than academia. I think GPT-3, especially, . . . makes this gap so much bigger. Because it's really clear that you get something… Self-supervised training at enormous scales actually produces something qualitatively different than what we've seen before. This is pretty broadly recognised. . . .[E]ven without NLP researchers, really, [OpenAI has] transformed a large percentage of the NLP research community.

In the process, projects like GPT-3 have arguably made certain lines of research redundant, as per the Bitter Lesson (above). FRANKIE said:

> GPT-3 isn't going to magically solve everything. But it does feel a little bit irrelevant to be asking some of the questions that we've been asking about smaller models, because I mean, once you get to a certain scale, a lot of questions just don't matter anymore.
>
> [. . .]
>
> One very large class of papers in academic publishing is: I'm going to create some task-specific model architecture that better captures some inductive bias of the

> domain that I'm trying to model, and is largely pretty specific to that task, and might actually help quite a bit. But if GPT-3, with some large self-supervised pre-training, just captures what's necessary, that whole class of paper goes away.

Some researchers mourn the shift, complaining that scaling is not scientifically interesting. One researcher tweeted sarcastically:

> GPT-3 is out with 175B parameters. I'm positively giddy with anticipation for GPT-4 with 400B parameters. That kind of progress is what motivated me to dedicate my research career to AI![246]

If training large models is scientifically uninteresting, and too expensive for most academic researchers anyway, it raises the question of where academics might fit into this new research direction. A theme that emerged from several of my interviews was the distinction between building models and contributing to *understanding* of those models. KHAL made this distinction:

> I think research we do academically will predominantly contribute to our understanding. When someone does build a strong AI system, it's going to be some organisation with massive resources. And they – in terms of . . . getting things to work better – they can do the hill climbing on their system that we can't. Whereas I think we can contribute to understanding.

---

[246] Compare this other tweet:

> The year is 2030
> Google, Microsoft & Open AI have merged
> Preprint on Arxiv (acquired by Github): "Fully Connected Neural Networks is all you need" - SOTA on all tasks
> In the crumbling remains of Stanford University, 1 voice whispers: "Ur not using symbolic reasoning"

Here, ‘understanding’ covers at least two areas of research: (1) evaluation of capabilities and limitations, and (2) interpretability. (1) involves getting a better understanding of the model’s behaviours, including what it can and cannot do, and (2) involves explaining those behaviours, normally by reference to the internal functioning of the model or the dataset on which it was trained. The importance of interpretability work was a very big factor behind why Connor Leahy hopes to replicate and share a GPT-3-like model. Like many others in the AI safety research community, he believes that for future, very powerful AI systems, researchers will be better equipped to prevent catastrophic failures if they can effectively monitor the system’s internal processing. During our conversation it became clear that Connor thought that this work was not only important for safety, but scientifically interesting:

> Toby Shevlane:
>
> [The] view of GPT-3 that I've been getting from you . . . is almost like it's this big, unexplored thing. Like the oceans, before we have any understanding of the ocean; or like the cosmos. That inside this model, there's all this cool stuff going on that we completely don't understand.
>
> Connor:
>
> That's exactly how I think about GPT. There is true empirical science to be done. This is science. We have discovered things to explore, and that we do not understand. No one knows what happens inside of GPT-3.

Connor wants GPT-3 to be open source so that ‘low-resource academics’ can contribute to this scientific challenge. Appendix 1 provides additional extracts, from my interviews and from tweets, about the academia-industry relationship in the GPT-3 era.

Similar to interpretability, another line of inquiry asks how a model's capabilities emerged throughout the training process. A 2021 paper from DeepMind, *Acquisition of Chess Knowledge in AlphaZero,* studies, for example, the progression throughout training that led to AlphaZero's particular approach to opening play.[247] To connect this line of work to background concerns about AI risk: the idea is that, if researchers better understand how the training process leads to certain behaviours within the system, they can better steer training so as to avoid any risky behaviours from emerging. In fact, this same logic explains why researchers at labs like OpenAI and Anthropic (a breakaway group from OpenAI) view research into scaling as safety-relevant. Scientific theories that reliably predict what capabilities and behaviours a model will learn, given certain inputs to the training process, make the process less 'mysterious' and more under the control of researchers.[248] (That said, as we saw above, this latter kind of 'understanding' is intertwined with methods for building stronger capabilities.)

In other words, the 'capabilities vs understanding' distinction does two things at once. First, describes a possible emerging division of labour between industry and academia: industry pushes the frontiers of capabilities, while both academia and industry work on better understanding AI systems. Second, it suggests one reason why this 'understanding' work, which could be done by academics, could contribute to reducing risks from AI. This is similar

[247]Thomas McGrath et al., 'Acquisition of Chess Knowledge in AlphaZero', *ArXiv:2111.09259 [Cs, Stat]*, 27 November 2021, http://arxiv.org/abs/2111.09259. For example, see this finding from the paper, pp.19-20: 'Figures 4, 7 and 6 suggest a sequence: that piece value is learned before basic opening knowledge; that once discovered, there is an explosion of basic opening knowledge in a short temporal window; that the network's opening theory is slowly refined over hundreds of thousands of training steps.'

[248] ROBIN told me, on the question of the pros and cons of publishing the scaling laws papers: 'Some of the relevant considerations there, which, you know, we did think about include: making deep learning more scientific (and which potentially be beneficial from a safety perspective). Versus people just throwing whatever they can at an experiment, knowing roughly what loss to expect is a move in the direction of more rigour.' Anthropic's website lists 'AI as a Systematic Science' as one of their research principles, and mentions scaling laws underneath.

but slightly different to the argument we saw in the previous chapter around Grover and GPT-2, where the idea was that access to the model could help academics find mitigations to the harms of text generation. With GPT-3, there is a shift: the model is not necessarily best categorised as a text-producing system, but perhaps as step closer to AGI. The model has very general capabilities, and it takes sustained research to properly understand its behaviours. This provides a nature role for academics in studying the model, contributing to the field's general understanding of large language models, and thereby making it more likely that labs will be able to spot dangerous properties of these models in future.

6.2 The API as a platform for academic research

We have seen that a few different factors all point in the same direction, suggesting that academics might want to contribute to scientific understanding of models like GPT-3 rather than building them: (a) resource constraints; (b) judgments about what is scientifically interesting; and (c) the potential to reduce risks from AI. The key question is whether the API facilitates them in this endeavour.

On one end of the spectrum of possibilities, the API could be for commercial use only, and shut out academic researchers. Academic researchers, if they hoped to study models like GPT-3, would first need to find some way of building and open sourcing these models. As we saw in section 3, this would then undermine the API regime, by allowing anyone to circumvent the restrictions on how GPT-3 can be used. Therefore, it is not in OpenAI's interests to make the API closed to researchers, because it would increase the pressures upon replication.

On the other end of the spectrum, we could imagine a case where the API is extremely easy for researchers to use, and it is flexible enough that it allows them to study GPT-3 in any way

that they can imagine. The API would then become another one of the convergent factors encouraging academic AI researchers to contribute to scientific understanding of models like GPT-3. The API would play into the existing division of labour between industry and academia when it comes to large models, where industry builds them, and both industry and academia seek to understand them.

The GPT-3 API is somewhere in between these two poles, and its exact location is open for debate. The researchers aiming to build open source, GPT-3-like models are naturally on the side that says the API is poorly suited for research purposes. The BigScience project, organised by the company Hugging Face, relies on researchers from many different institutions. On their webpage they motivate the project as follows:

> [W]hile recent models such as GPT3 . . . show interesting behavior from a research point of view, such models are private and not accessible to many academic organizations. Moreover, even when accessible, these tools have not been designed as research artifacts and for instance, lack access to the training dataset or checkpoints which makes it impossible to answer many important research questions around these models (capabilities, limitations, potential improvements, bias, ethics, environmental impact, general AI/cognitive research landscape).[249]

A similar statement can be found in the *Foundation Models* paper from researchers at Stanford University.[250]

[249] 'Introduction', BigScience Workshop, 2021, https://bigscience.notion.site/Introduction-5facbf41a16848d198bda853485e23a0.

[250] 'Foundation models start to roll back this positive trend. Some models (e.g., GPT-3) are not released at all (only API access to a limited pool of people). Even datasets (e.g., for GPT-2) are not released.' Bommasani et al., 'On the Opportunities and Risks of Foundation Models', 11.

Generally, I would argue that, while the API is a better platform for academic research than the above extract suggests, it could still go further. I will highlight the ways in which the API facilitates academic research, and also identify areas where it does not. This is an important issue: if the API is effective at preventing misuse, at what cost? Are risk-focussed KCRs such as the API worse for the research community, or is it possible to eliminate such tradeoffs? Moreover, aside from misuse, we have seen a different theory of how AI risk comes into the picture. The idea is that, if models are too closed off from the research community, scientific understanding will be hampered, ultimately leading to more risky AI systems being built.[251] Does the API have an answer to this concern? Finally, recall the related discussion in chapter 1 about how the 'staged release' regime aimed to foster a kind of distributed evaluation of GPT-2, producing risk-relevant knowledge about the model over time. This would include, for example, better knowledge of the model's capabilities. To what extent does the API improve upon staged release in fostering collective understanding of the model?

Scope of access

As of November 2021, there is no need to obtain approval from OpenAI before getting access to GPT-3. OpenAI requires approval for developers launching an application, but not for other uses, such as personal or research use. Researchers can therefore get easy access to the model. Before this point, when the API was in 'beta' mode, researchers would have needed

[251] Connor put a lot of weight on this consideration. I asked him: 'Is (a) or (b) more important? Where (a) is: we are *able* to align . . . these AGI models that we make. And (b) is: we're able to prevent people from taking AGI, even if it starts off as aligned, and making it unaligned and then having it let rip on the world?' (My logic was that (a), progress in alignment research, is made easier by open sourcing models, at least by Connor's lights; but it comes at the cost of (b), which is that it becomes more difficult to control how people modify and deploy the relevant AI system.) Connor said:

> So I genuinely think (a) is such a huge issue that (b) is basically trivial. Like I do not think about (b) at all . . . Eliezer [Yudkowski], in a tweet, once said it's kind of like....So someone was talking about the US vs China AI race, and Eliezer described it as, it's like two monkeys racing to who gets a bite of the poisoned banana first. It's that we don't know how to align these things at all to anything. . . . Until we can do any of that, it's completely irrelevant if the "bad guys" get access or the good guys, because it's gonna paperclip up anyways.

to apply for model access. I do not have any statistics on how many researchers had access during this period. Like with GPT-2, there was a specific academic access program, which in this case gave away free API access to academics looking to study a broad range of topics.[252] When I spoke to CAMERON in November 2021, four months after API beta launched, he described the academic access programme:

> Well, we have more than 15 different universities on the platform now. Some of the universities have top level PIs who themselves are responsible for 10 investigators on 10 different projects. So we devolve a huge amount of the research choice to the people who talk to us.

Already by this point, there were many more academics on the programme than the equivalent programme for GPT-2 (see chapter 1). The API makes this process much simpler from OpenAI's side, compared to, for example, sending academics a copy of GPT-2, which requires more trust. In addition, I got the impression that many researchers had API access even without having gone through the academic access route. On Twitter, for example, I saw one researcher complaining that they did not have access when it felt like 'everybody on twitter' already did, and the next day they tweeted thanking the CTO of OpenAI for giving them access. For myself, I gained access after reaching out to somebody at OpenAI asking to interview them about GPT-3 and the API.

[252] The online form for academic access read: 'We're initially scoping our research collaborations to the following areas, though we welcome suggestions for areas of focus in the future.' The areas listed were: fairness and representation; robustness; model exploration; interdisciplinary research; and misuses potential. The description for 'model exploration' begins: 'Models like those served by the API have a variety of capabilities which we are yet to explore.'

Even though the documentation for the API focuses on rules for *developers* building GPT-3 into actual applications, it does seem like researcher access is a core function of the API. For example, somebody at OpenAI who works on the API said on a podcast (February 2022):

> We have these two camps of users: the researchers and the developers. The developers keep telling us, 'hey, I just want one button; I just want the best model to come out.' And a lot of the researchers want to fiddle more with the parameters. And I think we can probably satisfy both for a long time.[253]

And there are plenty of academic papers that rely on API access (see below). The biggest exception might be that the API only works in certain countries. I counted 152 on the list of supported countries, but this does not include China or Russia (as of March 2022). The webpage states that OpenAI are 'working hard to increase the number of locations we can provide safe access to'.[254]

Depth of access

What can researchers do with GPT-3, and what can they not? How is their access different from having an open source version of the model? ROBIN told me in March 2021:

> …on a theoretical level, out of all the possible things that you can do to the model, they're much more constrained. They can't make any changes to the underlying weights. They can't fine-tune it arbitrarily. They can't remove layers, they can't inspect the activations; they can't do all sorts of things.

[253] 'Peter & Boris — Fine-Tuning OpenAI's GPT-3', Gradient Dissent, March 2022, https://wandb.ai/fully-connected/podcast.
[254] List available at: https://beta.openai.com/docs/supported-countries

ROBIN continued that in practice, the difference is not as large as it seems. One reason is that researchers doing low-level interpretability research do not need access to the largest language models, because there is still so much that we do not understand about smaller, simpler models. The other reason was that there is a lot of research that can be done using the API. To see this, we can look at the different kinds of research that people have done using the API:

Kinds of question	**Examples**
What capabilities does GPT-3 have?	Hendrycks et al evaluate GPT-3's performance on high school mathematics problems.[255]
What safety/ethical problems does GPT-3 have?	Lin et al study how 'truthful' GPT-3's responses to questions are.[256] Abid et al find that GPT-3 captures anti-Muslim bias.[257]
Why does prompt engineering work?	Min et al ask: if you give GPT-3 an example, what is it about the example that helps GPT-3's performance?[258]
Does GPT-3 actually do 'few-shot learning' as the paper claims?	Perez et al argue that GPT-3's performance depends on the fact that the user has experimented with different prompts, which is not 'true' few shot-learning.[259]
What is the best way of using GPT-3 for a given application?	Muennighoff offers a method for using GPT-3's outputs to search across text,

[255] Dan Hendrycks et al., 'Measuring Mathematical Problem Solving With the MATH Dataset', *ArXiv:2103.03874 [Cs]*, 8 November 2021, http://arxiv.org/abs/2103.03874.
[256] Stephanie Lin, Jacob Hilton, and Owain Evans, 'TruthfulQA: Measuring How Models Mimic Human Falsehoods', *ArXiv:2109.07958 [Cs]*, 8 September 2021, http://arxiv.org/abs/2109.07958.
[257] Abubakar Abid, Maheen Farooqi, and James Zou, 'Persistent Anti-Muslim Bias in Large Language Models', *ArXiv:2101.05783 [Cs]*, 18 January 2021, http://arxiv.org/abs/2101.05783.
[258] Sewon Min et al., 'Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?', *ArXiv:2202.12837 [Cs]*, 25 February 2022, http://arxiv.org/abs/2202.12837.
[259] Ethan Perez, Douwe Kiela, and Kyunghyun Cho, 'True Few-Shot Learning with Language Models', *ArXiv:2105.11447 [Cs, Stat]*, 24 May 2021, http://arxiv.org/abs/2105.11447.

	comparing this against other approaches.[260]
How can GPT-3 be used to create training data for other models?	Bonifacio et al use GPT-3 to create a new dataset that trains a different kind of language model (retriever models).[261]
How GPT-3 be improved so that it makes fewer mistakes?	Madaan et al combine GPT-3 with another algorithm to improve its performance.[262]

Table 1: Academic study of GPT-3 through the API.

Some of these papers evaluate GPT-3 (its capabilities and weaknesses); another sheds some explanatory light on how GPT-3 works (i.e. the paper on prompt engineering); and others show how GPT-3 can be combined with other methods to create something new. The papers show that all of this is possible with the API. With the API, GPT-3 has become a research artefact, even without anyone outside OpenAI and Microsoft having the model itself. The papers describe how the researchers interacted with GPT-3 and under what settings. For example, Hendrycks et al state:

> We use the 'Ada' GPT-3 model which has approximately 2.7 billion parameters, and the 'Davinci' model which has approximately 175 billion parameters. Since we are performing few-shot evaluation, we construct our prompt by prepending 8 problems with correct answers (but not step-by-step solutions due to space). Using temperature 0, models output up to 20 tokens for the final answer.

[260] Niklas Muennighoff, 'SGPT: GPT Sentence Embeddings for Semantic Search', *ArXiv:2202.08904 [Cs]*, 23 March 2022, http://arxiv.org/abs/2202.08904.
[261] Luiz Bonifacio et al., 'InPars: Data Augmentation for Information Retrieval Using Large Language Models', *ArXiv:2202.05144 [Cs]*, 10 February 2022, http://arxiv.org/abs/2202.05144.
[262] Aman Madaan et al., 'Memory-Assisted Prompt Editing to Improve GPT-3 after Deployment', *ArXiv:2201.06009 [Cs]*, 16 March 2022, http://arxiv.org/abs/2201.06009.

In other words, the API's knobs and dials have become part of the scientific process in some way. I would say that the API, despite being primarily intended as a commercial product, has 'gone native' within the research community. What's more, it is interesting to see that there is a small research ecosystem around the model. This makes sense given that the model's capabilities are broad and difficult to know; but it is noteworthy that having an API instead of an open source model still lets this research ecosystem grow.

It has been interesting to watch the research community's collective understanding of GPT-3 unfold in a distributed fashion.[263] It is very far from being a model that just researchers at OpenAI know about; and many different researchers have played a part in contributing to the collective understanding of GPT-3. Beyond the papers above, a few illustrative examples are:

1. Connor described two junior researchers, not working at OpenAI, as probably 'the best users of GPT-3 in the world; they're probably the most masterful in the entire world'.[264]
2. CAMERON told me that OpenAI researchers had internally tried to get GPT-3 to generate software code, but: 'then our users came along . . . and showed much better coding capabilities than we had ourselves got out of the GPT-3 model, because these people had just done better prompt design and other things.' BELLAMY, a university-based postdoctoral researcher, gave the same example and described it as

[263] CAMERON: 'Any organisation developing [generative models] should have the attitude that it doesn't know what it doesn't know, and these generative models have a load of capabilities which you won't ever realise are there. So you have to get people to experiment with them. And so the way that I think of the API is like, yeah, maybe it's interesting from a commercial standpoint, but I'm really bullish on this stuff for access to models in the future in general.'

[264] Two of their papers: Laria Reynolds and Kyle McDonell, 'Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm', *ArXiv:2102.07350 [Cs]*, 15 February 2021, http://arxiv.org/abs/2102.07350; Laria Reynolds and Kyle McDonell, 'Multiversal Views on Language Models', *ArXiv:2102.06391 [Cs]*, 15 February 2021, http://arxiv.org/abs/2102.06391.

a benefit of ‘making models public’. (Showing that the API challenges prevailing conceptions of what it means to make a research model ‘public’.)

3. Eliezer Yudkowski, a well-known advocate of AI risk, wrote a Twitter thread about how he thought GPT-3 could be, in one case, ‘pretending to be stupider than it is’. The example came from a software programmer who had, via AI Dungeon, tried to test whether GPT-3 could classify whether a set of parentheses were balanced or unbalanced (e.g. ‘(( ))’ or ‘ ( )))’). The simulated conversation was between the user and a character within AI Dungeon called John. Yudkowski’s hypothesis was that the model was getting some of the answers wrong because that was what the character of John would do. Yudkowski is fearful of powerful, future AI systems that misrepresent their own capabilities to their users.

4. A researcher on OpenAI’s policy team posted to Twitter their attempts to get GPT-3 to generate poems in the style of Robert Burns, an 18th Century poet. Another Twitter user replied: ‘18-19c lit scholar here to say that actually is pretty good. "The feudal chain, the despot's rod/Shall snap"—yep. Idk if it sounds like Burns, exactly, but it does sound like a vision of the future written in 1789.’ This is a good example of how both (a) GPT-3’s capabilities are being evaluated in a distributed way, and (b) evaluating GPT-3 sometimes involves qualitative judgments from domain experts. This is a far cry from QUANTA, above, the evaluation of which can largely be reduced to a single number (the percentage of correct answers on the evaluation dataset).

5. Similarly, Liam Porr (see above) told me that his aim was to demonstrate the writing capabilities of GPT-3, which are otherwise ‘hard to quantify’. His aim was to ‘translate [how good it was] into numbers’. The people reading the blog posts were part of an experiment, where Liam was trying to show that they would believe the

> posts to be human-written. This is a kind of distributed evaluation, albeit one that OpenAI did not want to happen.

Finally, the papers and examples I have given so far do not even take advantage of newer features that have been added to the API. An important addition is the ability to create a fine-tuned version of GPT-3, with new data, then access that modified model through the API. ROBIN highlighted the role that fine-tuning can play in evaluating the model's capabilities:

> [T]hings like fine-tuning are an additional level of access in some sense. If you can not only use the weights, but also modify the weights, that can allow a different lens on what the model is capable of, and what its knowledge is and does and doesn't consist of.

It will also be possible for researchers to work on new datasets that, when used for fine-tuning of models like GPT-3, improve the performance or safety of such models. Researchers at OpenAI have done such research: a 2022 paper shows how fine-tuning GPT-3 can make it give responses that are less toxic, more truthful, and more helpful to the user.[265] In this way, the GPT-3 model becomes something that can be improved upon: not necessarily by training a whole new model, but by using GPT-3 as a starting point. This is the idea behind the term 'foundation models': that they are a foundation to be built upon.[266] The API's success so far indicates that it is possible for a model to be a foundation model in this sense even without the weights being open source.

[265] Ouyang et al., 'Training Language Models to Follow Instructions with Human Feedback'.
[266] Bommasani et al., 'On the Opportunities and Risks of Foundation Models', 6.

However, there are certain areas where the API, as a platform for research, is less flexible than open source models. These limitations are not necessarily inherent limitations of APIs, but at least they show the drawbacks of how OpenAI currently deploys GPT-3. I would highlight five areas:

1. Quotas. FRANKIE told me about a collaborating with a university student:

> There's a limited quota that we have access to. And so she has to think, *okay, which experiments do I run? Can I afford to do them on the GPT-3 biggest model? . . .* It's interesting, because at the end of the month, it's like, *okay, we have this much left, I need to be sure to run everything before my quota resets.*[267]

This might depend upon whether the researcher is on OpenAI's academic access programme. CAMERON told me that one research group on OpenAI's academic access programme became the 'third biggest customer for a couple of weeks'. CAMERON continued: 'except we give academic access away. So we had to have this conversation with them where we were like, your experiments... we don't know how to say this, but they're costing us a lot of money.' FRANKIE suggested a different approach, where the academics simply use their own funding to pay for as much usage as they need. However, as with all of the limitations I list here, the quotas are partly in place to make it difficult to either (a) misuse the model at scale, or (b) extract the underlying intellectual property from the model.[268] In future it might be that these problems can be overcome, either through gaining trust that specific academic researchers will not do these things, or by better monitoring to check that they are not.

[267] Similarly, ALEXIS told me that they had not used the API but: 'I'd probably get rate limited pretty quickly. You know, I'll just run like way too many experiments, expecting it to be essentially free to call this API, but really, it's not.'

[268] In this case, by training a new model on the outputs of GPT-3. See Tramèr et al., 'Stealing Machine Learning Models via Prediction APIs'.

2. Inspecting the weights and activations. Researchers cannot inspect the weights of the model, nor the 'activations', i.e. how much each neuron is active in response to a given input. Both of these are useful for interpretability research. For example, Connor told me about his research with Eleuther on the models that they were training. The research involves taking a GPT model and training a classifier (e.g. a funniness classifier, or a rudeness classifier) that takes, as an input, the activations of a specific layer of the GPT model. Connor said they found:

> the deeper you go in the network, the better the performance climbs, until the last layer. The very last layer, for some reason, the performance collapses; it just throws out all this information. . . . So the second-to-last layer is actually the most semantically rich.

This is not research that could have been done on the API, because the API does not give the activations of different layers of the model. Connor claimed that '99.9% of the interesting stuff" with GPT-3 is at this 'intermediate' stage, before the model gives a final prediction.

3. Modifications. We have seen that researchers can fine-tune GPT-3 on new data. However, researchers cannot modify the model in other ways. For example, researchers cannot identify neurons as performing a particular function in the model and then test this by going into the model and changing the relevant weights.

4. Dataset. The dataset for GPT-3 is not publicly available. I do not know whether this is for legal reasons (the dataset will contain much copyright material) or an attempt to delay

replication. JAN argued that the lack of dataset is limiting for researchers trying to study the model:

> [T]hese models, they are just reflections of what they are trained on. . . . And so if you can't investigate the data distribution, you're very limited in how you can understand this model. . . . A lot of the excitement from these models was like, *Oh, they can do zero shot stuff, they can do stuff that they were not trained on, they can do new stuff.* But if you have no access to the dataset, how are you exactly sure that these are new stuff, and this is just not . . . somewhere on the internet?

One possibility could be for the lab to give researchers an interface for interacting with the dataset even without downloading it – for example, searching across the dataset to see whether, for certain outputs, GPT-3 is just copying from the dataset.

5. Models throughout training. Above I described a DeepMind paper that compared different versions of a chess-playing system, corresponding to different stages in its training process. This would not be possible with GPT-3 because only the final versions of the models are available on the API. Similarly, a recent paper from Anthropic looks at a specific point in the training process where a large language model went through a sudden jump in its capabilities, and compares the versions either side of this jump. OpenAI could serve on the API tens or hundreds of different iterations of GPT-3, alongside related statistics such as loss curves (i.e. how well the model was performing across the training run).

Missed opportunity: how is GPT-3 actually being used?

Beyond these five areas, which all compare less favourably against an open source approach, I want to highlight one potential missed opportunity. We saw above that the API allows the

lab to know what applications are being developed using the model. OpenAI occasionally advertises particular applications that have been built by GPT-3.[269] However, OpenAI has not provided analysis of all the different applications: what they are, and what role GPT-3 plays. This could help to shed light on the social impact of models like GPT-3.

To elaborate this point: discussions about the risks of models like GPT-3 often take, as a starting point, the fact that these models *produce text*. The question then becomes: what could be the harms from producing text? The main risks that fall out of this analysis are that the text could be biassed, misleading, or part of some fraudulent or disinformation effort.[270] This is a kind of *deductive* analysis, in that it starts with an assumption about what the model does and then reasons from there about the ways it could go wrong. This would contrast with an *inductive* approach, which would look at GPT-3's applications and study what effects they are having on society.

When social scientists come to engage with GPT-3, their research is normally slotted into the existing theory of GPT-3's risks that has been arrived at through the deductive approach. For example, social scientists, as part of OpenAI's academic access programme, study whether humans find GPT-3-made propaganda believable. The assumption is that GPT-3-like models could be used for propaganda, and then the empirical research comes in to evaluate, within an experimental setting, whether such propaganda would be effective. My claim is not that this kind of research is ill-conceived, but rather that it is a subset of the possible social science research to be done. It is the kind of research that can be easily commissioned via the API: the lab gives API access to the social scientists, and they can run experiments using the API that speak to GPT-3's possible societal impacts. The research on how GPT-3 repeats anti-

[269] OpenAI and Pilipiszyn, 'GPT-3 Powers the Next Generation of Apps'.
[270] See, for example: Laura Weidinger et al., 'Ethical and Social Risks of Harm from Language Models', *ArXiv:2112.04359 [Cs]*, 8 December 2021, http://arxiv.org/abs/2112.04359.

Muslim steroytypes follows the same approach. The researchers identify the bias in the model's outputs, and then assume (fairly) that these outputs could infect a whole range of different applications. One final illustration of the point: Weidinger et al go through many different possible risks of large language models, including biassed outputs, misleading outputs, and propaganda.[271] For each risk, the authors tag whether it is 'observed' or 'theoretical'. However, 'observed' does not mean 'observed in the real world'. In practice, it means: observed by people experimenting with the GPT-3 API (or some other language model). What has been observed is the model's behaviour, rather than the societal problem itself.

The understanding of GPT-3's social impact could benefit from an inductive approach, which would complement the above research efforts. Social scientists would start by looking at the applications that have been built using the API. Even before doing any empirical research on these applications, I would argue that we can already see how this might add a new dimension to the understanding of GPT-3's risks. From looking at the applications that have been announced, many of them are not about *producing text* per se. For example, certain applications use GPT-3 as part of a semantic search pipeline. Sorting through search results has a different flavour to generating text: the model is being used to sort through information, rather than as some kind of discursive agent. Also, Connor Leahy's biggest fear about future GPT models is that they will supply the 'world model' (i.e. the understanding of the world) for an agent that takes open-ended actions in the world. He told me that, for him, the most interesting thing about GPT-3 is 'not the text it generates', which is 'downstream' of the model's understanding of the world. As an early indication of this possible future: GPT-3 has already been adapted into an agent that browses the internet.[272]

---

[271] Ibid.

[272] Reiichiro Nakano et al., 'WebGPT: Browser-Assisted Question-Answering with Human Feedback', *ArXiv:2112.09332 [Cs]*, 12 March 2022, http://arxiv.org/abs/2112.09332.

In this light, the existing focus on text generation looks like the inverse of the same mistake that OpenAI made when thinking about their CLIP model. As we saw in chapter 1, ROBIN was surprised that the model, which was supposed to *classify* images, was being widely used within a pipeline for *generating* images. These examples caution against overreliance on a kind of analysis which starts with the technical properties of the model (e.g. GPT is a 'generative' model, hence the 'G') and, from there, makes over-simplistic projections about how the model will be used.

A step towards a more inductive approach would involve a comprehensive analysis of the different ways in which GPT-3 is actually being used. This would include typologies of the different kinds of applications. There are different ways of slicing up GPT-3's applications: for example, by business function (customer relations, human resources, and so forth), or by the mechanism through which GPT-3 is useful (search, classification, conversational agents, etc.). Alongside such typologies, OpenAI could produce statistics on the scale at which GPT-3 is being used for different kinds of applications. Hypothetically, OpenAI could find that 60% of GPT-3's usage (operationalised, for example, in terms of number of queries) is devoted to customer relations, for example, or for lawyers searching across documents. The question of GPT-3's social impact could then be attacked from that angle. At present, any such facts are not widely known. Revealing them could add an additional level of transparency over risk-relevant knowledge beyond the transparency that comes from researchers having API access.

At the same time, this inductive approach would also allow for a more structural approach to risk analysis, in that social scientists could try to study the structural impact that these models

are having on society.[273] Of course, this is a more difficult kind of research, and it requires time for these structural changes to become evident. Therefore, although I have argued that the API helps towards tackling the Collingridge dilemma, it is not a full solution. The feedback loop – between the lab and the wider society, which I discussed above – will not always be a short one, where researchers quickly identify a dangerous capability and it is withdrawn from the API service. This will need to be done where possible. However, (absent a process for making very accurate forecasts of social impact, which nobody currently has) this approach cannot fully substitute for the slower, more inductive process, where researchers study the actual impacts of AI as they unfold.

Conclusion

At the top of the chapter I quoted CAMERON saying that GPT-3 was a moment where 'the future lurches into your present'. I would make a similar point, with different emphasis. Trends in the computational power of hardware used within AI research, combined with the spending budgets of typical AI research labs, broadly sets a certain timeline for when researchers will arrive at different AI capabilities. With GPT-3, OpenAI raced ahead of this timeline, perhaps by a year or two, mainly by spending unusually large sums of money on compute. This presented the lab with an opportunity to try and govern the AI capabilities that they had arrived at.

The lab seized this opportunity, almost by accident. The research team adopted a particular strategy for addressing AI risks: delay publishing the paper, fill it with analysis that sought to characterise the weaknesses of the model, and decline to open source the model.[274] The

[273] Zwetsloot and Dafoe, 'Thinking About Risks From AI'.

[274] It is unclear whether GPT-3 would have been open sourced if the company had never pursued the API. My hunch is that it would not have been open sourced, at least not for a long time (perhaps until a very similar model had been open sourced).

strategy was to keep GPT-3 as something that existed in the future, at least beyond the walls of the lab ("the future is already here; it's just not evenly distributed").[275] More generally they wanted to avoid speeding up AI timelines, fearing that the world does not have enough time to prepare for advances in AI capabilities. The instinct of the research team, therefore, was not to actively govern the model itself, but the perceptions of the model's significance in the wider world.

The API was a very different strategy for addressing AI risks. It was a product of the lab's commercial incentives. The API meant that everyone could see how capable GPT-3 was, undermining the strategy of downplaying its capabilities. But the API turned the model into something that could be governed, and more powerfully than any AI research model that had gone before it. The API, as a KCR, was imported from the commercial world, and brought with it a level of control over use that was foreign to the research community. Instead of assessing the risks of a whole model, the API allowed the lab to consider risks on an application-by-application basis. The lab could also iterate over time, changing who could do what with the model as they learned from experience. As such, in comparison to the open source regime, the API greatly expanded the responsibility that the lab could adopt over how the model was deployed in society. It redrew the boundaries between issues 'that the research community can possibly address' and issues that are 'external policy problems', expanding the territory of the former.

At the same time, the API made the model into a researchable artefact, pushing forward the research community's understanding of large language models. There are still areas where the open source regime does this better, and so room for the API to improve. But generally,

[275] The quotation is from William Gibson.

the API facilitated an understanding of GPT-3 that unfolded in a distributed way, with many more eyes studying the model than would have been possible internal to OpenAI. The API eludes a simplistic distinction between ‘open’ and ‘closed’, showing that a model can be closed with respect to misuse, open with respect to legitimate uses, and (at least somewhat) open with respect to further research. The ability to have all these things at once could ultimately prove important for reducing risks from AI.

CHAPTER 3: AN ETHICS REGIME FOR CONFERENCE PUBLICATION

Abstract: This chapter analyses a new ethics regime introduced at various AI conferences in 2020 and 2021, focussing on the NeurIPS conference. The new regime involved: (1) an alteration to the peer review process, whereby ethics reviewers would review papers flagged as raising ethical issues; and (2) a requirement that authors append to their papers a "broader impact statement" discussing the possible negative societal consequences of their work. I argue that the regime is functionally different from those discussed in previous chapters, in that it is not primarily aimed at controlling the proliferation of risky technologies. Instead, the regime became a primarily educational tool, encouraging researchers to reflect about the ethical implications of their papers, and facilitating contact with ethical experts to give feedback on such reflections. Instead of governing an external population of users, the regime serves to govern the research community: how the researchers think and talk about ethical impact. The problem remedied by the regime is not per se the societal impact of AI, but rather the perceived lack of reflection by researchers on such social impact. Hence the remedy for papers caught by ethical review would typically be a correction to the language the authors used in their broader impact statement. I argue that this development has its roots in the existing conference publication regime, which serves as a very weak foundation for governing the impact of AI research on society. The original pioneers behind the initiative had lofty aims, concerned with the impact of AI on democracy (among other things). However, the opportunity provided by the peer review process was not enough to directly address such risks. I cite various difficulties stemming from such reliance on the existing conference publication regime. These include: the idea that individual papers are the wrong level of analysis for governing social impact, and the institutional separation between research and AI applications (discussed in the previous chapter). These limitations

necessarily shape the kinds of risks the initiative can attempt to tackle, pushing attention away from large-scale, incremental, structural impact on society. I conclude that risk-based KCRs do not simply incorporate existing notions of risk, but often structure the ongoing conversation about such risks, and in a way that is biassed by the existing institutional framework (in our case, the conference publication regime).

1. Introduction

Broader impact statements (BIS) and ethical review were introduced by the NeurIPS conference organisers for the 2020 conference.[276] Papers were expected to include a statement about how the work might affect society, and ethics reviewers would review papers that were flagged as raising ethical issues. Since then, a number of other AI conferences have introduced similar measures.

In this chapter, I analyse this new, experimental KCR: the new ethics regime for conference publication. I focus especially on the NeurIPS conference, where the new regime was pioneered. Partly, the question is how the regime functions. How does the regime serve to reduce risks from AI? How is knowledge about risk produced and channelled into decision-making? I argue that the regime is not primarily aimed at reducing the proliferation of dangerous technology. In this regard, it is different from the regimes for GPT-2 and GPT-3 that we encountered in previous chapters.

To further illustrate this point, we can also contrast the new conference ethics regime with an example from nuclear physics. In 1939, before the outbreak of World War II, a number of nuclear physicists coordinated with an academic journal in an attempt to prevent the

[276] Neural Information Processing Systems: https://nips.cc/

publication of experiments about nuclear fission.[277] The attempt was driven forward by Leo Szilard, a Hungarian-American physicist who foresaw the potential for nuclear weapons and was fearful of their potential use by Nazi Germany. Wellerstein's history of nuclear secrecy describes how the scheme came together after an experiment from Szilard and others suggesting that a nuclear chain reaction was possible:

> It was a high-quality discovery in physics, but one that increased Szilard's fears of a Nazi bomb. As the scientists wrote up the results, Hitler was invading Czechoslovakia. Szilard's argument for self-censorship was taking on more weight. The Collumbia physicists met again and a compromise was reached: they would adopt a form of secrecy. Any new papers of fission would be sent to the *Physical Review*, who would register having received it. These registrations could, perhaps, be used to arbitrate later priority disputes. But the papers themselves would remain unpublished until a later date. It was a scheme that, ideally, would satisfy the need for priority without making the work immediately public.

(The scheme was ultimately unsuccessful because a team of French physicists published the key result in Nature.)

I highlight this example because, from one angle, it resembles the changes to the AI conferences. On their face, the new changes mean that the conference can decline to publish a potentially dangerous result. However, the new KCR has not moved in this direction. Indeed, it has moved in a different direction altogether. Much of the emphasis on ethics

[277] Alex Wellerstein, *Restricted Data* (University of Chicago Press, 2021).

reviews is, in practice, about issues such as research ethics (i.e. harms caused *within* the research process) and the way ethical issues are framed in the paper's writing.

As well as describing this evolution of the new regime, I also offer an explanation for why it has moved in this direction. I argue that the priorities of the new KCR are a result of its basis in the existing regime of conference publication. BIS and ethical review were not created on a blank slate. They do not fundamentally change the conference publication system, which still revolves around papers being submitted and reviewed. Reflecting discussions in the previous chapters, I argue that this system of papers provides a poor foundation for building a regime designed to tackle AI risks created by AI research. This institutional context is why BIS and ethical review have not been able to seriously confront the issue of the proliferation of dangerous technology.

As with previous chapters, I rely on interviews with AI researchers, between 2021 and 2022, and following conversations on Twitter. I interviewed one of the organisers for NeurIPS 2020, which gave me an insight into how the new regime was intended to work. I also read many ethics reviews, which are available for NeurIPS 2021 on the OpenReview platform. I made notes on 14 different papers and their reviews, and include reference to some of these in the chapter. I have also relied on public statements and blogs from conference organisers, especially for NeurIPS and NAACL, a natural language processing conference.

2. Background

Ethical review and the broader impact statements were both introduced for NeurIPS 2020 and went hand-in-hand. Ethical review came in response to a few papers at the previous year's conference which were perceived to be unethical. TAYLOR, who was one of the researchers organising NeurIPS 2020 and involved in the changes, told me that the previous

year's organising committee passed on the problem as 'something that really needs to be dealt with'. TAYLOR said:

> In 2019, there were a few papers that were accepted, and then after they were published, then people, (Twitter), said, *wait a minute, there are ethical concerns with these papers, why was there no consideration of potential harm and risk from publishing these?* And so that was one of the sort of assignments that was given to us for this year's programme chairs – was to think about how we would put such a process into place. . . . So that's really where it came from now, I'd say, is confronting the possibility for actual harm to come from papers, from work published, and wanting a way to deal with that.

I asked which papers from 2019 TAYLOR was referring to. They said:

> One of them was, it was a paper that.. it was a generative network that generated a face image from a voice input.[278] . . . And that was widely considered to be, as you know, something that was problematic.

This paper had been highlighted on Twitter by a well-known AI ethics researcher, who commented: 'Computer scientists and machine learning people, please stop this awful transphobic shit.' Another NeurIPS 2019 paper that was criticised on Twitter was about predicting whether an image used in a news article was from a left-leaning or right-leaning

[278] Yandong Wen, Bhiksha Raj, and Rita Singh, 'Face Reconstruction from Voice Using Generative Adversarial Networks', in *Advances in Neural Information Processing Systems*, ed. H. Wallach et al., vol. 32 (Curran Associates, Inc., 2019), https://proceedings.neurips.cc/paper/2019/file/eb9fc349601c69352c859c1faa287874-Paper.pdf.

publication.[279] The paper included a method for taking images of politicians (including Trump, Clinton, and Obama) and making their faces look more left or right leaning according to the model. The shared criticism across both of these papers seems to be one of validity. The papers make questionable assumptions about the domain they are modelling.

The conference organisers intended the BIS to dovetail with ethical review. The idea of BIS for AI papers originated with Hecht et al, who argued for this change in a 2018 paper.[280] Hecht et al argued that AI research is having a significant impact on society, and it is irresponsible that papers only mention the societal benefits and not the risks. TAYLOR told me that a colleague shared this paper with the 2020 programme chairs and it 'made a lot of sense to us'. TAYLOR said that one of the reasons for introducing BIS was the contradiction that authors discussed positive impacts of their work but not negative ones. The other main justification was that researchers should take ownership of the real-world impacts of AI:

> [T]here was the recognition that machine learning is used broadly in the world; this is no longer a speculative field. The algorithms that are developed really are changing society and changing the world. So I think it's important that we consider that and be.. think about that impact and write it in our papers. Take ownership of it.

The BIS would be an additional section at the end of the paper, enjoying an extension to the typical 8 page limit. All authors were asked to make such a statement, although that could include a simple statement that it was not applicable due to the nature of the work. As far as

[279] Christopher Thomas and Adriana Kovashka, 'Predicting the Politics of an Image Using Webly Supervised Data', *ArXiv:1911.00147 [Cs]*, 31 October 2019, http://arxiv.org/abs/1911.00147.
[280] Brent Hecht et al., 'It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process', *ACM Future of Computing Academy* (blog), 2018, http://arxiv.org/abs/2112.09544.

I am aware, the impact statements did not feed into the conference itself, e.g. there were not workshops organised around debating specific statements.

Ethics review and BIS went hand-in-hand. Papers would not be rejected because they had a bad BIS, but a good BIS could help to save a paper that otherwise raised ethical concerns. The BIS was supposed to give the authors the opportunity to show that they had considered the relevant ethical issues.

The reviewers or the area chair (i.e. the person overseeing the review process) could flag a paper for ethical review. The paper would then have dedicated ethics reviewers assigned to it, who were supposed to have relevant expertise. Their judgement would inform the decision on whether the paper was accepted or rejected. The ethics reviewer pool was selected by asking for recommendations from senior members of the 'FAT community' (Fairness, Accountability, and Transparency – a conference) and the ethics teams at industry labs that the conference organisers were connected to.[281] For the 2020 conference, ethical review was only applied to papers that were likely to be accepted.

Judging from my interviews and the social media discussion, the introduction of broader impact statements was not very popular with the research community. A survey of researchers in 2020 identified the same set of complaints that I had encountered:

> our respondents indicated a combination of nonchalance in their approach to the requirement ("This is just another exercise in 'doing something'. I expect these statements to become pro forma with time, since it will be possible to look at previous

[281] Information from TAYLOR, one of the organisers.

> years' papers for 'inspiration'."; "I wasn't particularly rigorous about it."), outright farce ("If I liked writing fiction I would be writing novels."), or perceived it as a burden ("one more burden that falls on the shoulders of already overworked researchers"). [...] Some respondents described the requirement as "too broad" or said they did not feel "qualified to address the broader impact of [their] work." Among those who supported the requirement, some found the thought process most valuable, and that it "forces researchers to reflect on the impact of their research."[282]

A couple of changes were made to the process for the 2021 conference. From 2021 there was no requirement for a specific BIS at the end of the paper, but the authors still had the extra page added to the maximum page limit (9 instead of 8). According to the ethics guidelines, authors were still 'expected to include a discussion about potential negative societal impacts'.[283] Authors submitting a paper had to fill out a checklist that asked whether they had done so. Second, the ethical review process was scaled up. For 2021 there were 105 ethics reviewers, and ethics review could be applied regardless of how likely a paper was to be accepted.

The BIS and ethical review were supposed to apply to a wide range of different ethical concerns. The checklist for NeurIPS 2021 illustrates the scope, stating:

> Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), environmental impact (e.g., training huge models), fairness considerations (e.g., deployment of technologies that could further disadvantage historically disadvantaged groups), privacy

[282] Grace Abuhamad and Claudel Rheault, 'Like a Researcher Stating Broader Impact For the Very First Time', *ArXiv::2011.13032 [cs.CY]*, 2020, https://arxiv.org/abs/2011.13032

[283] NeurIPS 2021, 'Ethics Guidelines', accessed 20 April 2022, https://neurips.cc/public/EthicsGuidelines.

considerations (e.g., a paper on model/data stealing), and security considerations (e.g., adversarial attacks).

At the 2020 conference, the average BIS was about seven sentences long.[284] Around 90% of papers gave a statement rather than stating it did not apply.[285] To give a flavour of the content, Appendix II displays five statements, randomly selected from a dataset of the NeurIPS 2020 broader impact statements.[286]

Only a small handful of papers went through ethical review. Thirteen received ethical review, and four were rejected on ethical grounds.[287] For the 2021 conference, eight papers were accepted conditional upon the authors making changes to the paper, and one paper was rejected on ethical grounds.

Variants of the ethical review and BIS process have spread to other AI conferences. Big NLP conferences (such as NAACL, ACL, EMNLP) ask for some kind of BIS and conduct ethical review. CVPR, a large computer vision conference, strongly encourages authors to discuss 'ethical and societal implications of their work in their papers', and reviewers are 'asked to positively weigh the depth of such ethical reflections'.[288] Authors at ICLR are encouraged to include an ethics statement in their papers, and reviewers can flag papers as raising ethical problems. ICML authors are also expected to highlight and address any risks of their work.

[284] Carolyn Ashurst et al., 'AI Ethics Statements -- Analysis and Lessons Learnt from NeurIPS Broader Impact Statements', *ArXiv:2111.01705 [Cs]*, 2 November 2021, http://arxiv.org/abs/2111.01705.
[285] Ibid.
[286] Ibid. Dataset available at: https://github.com/paulsedille/NeurIPS-Broader-Impact-Statements
[287] Hsuan-Tien Lin et al., 'What We Learned from NeurIPS 2020 Reviewing Process', *Medium* (blog), 16 October 2020, https://neuripsconf.medium.com/what-we-learned-from-neurips-2020-reviewing-process-e24549eea38f.
[288] CVPR 2022, 'Ethics Guideline', accessed 20 April 2022, https://cvpr2022.thecvf.com/ethics-guidelines.

3. How does the regime work?

If the aim of the GPT-2 and GPT-3 KCRs was to govern a population of users and application developers, the aim of BIS and ethical review is to govern the population of researchers. The mechanism is 'responsibilisation':[289] to prompt researchers to reflect upon the social impact of their work and (armed with that knowledge) to become active members of the drive to make AI research more ethical.

At the inception of the new regime for the 2020 conference, the organisers hoped that a body of knowledge about ethics and risk would emerge over time, in a bottom-up process. TAYLOR told me that the organisers intentionally left vague the question of what kinds of risks or ethical problems were relevant. In terms of the content of the BIS, TAYLOR said:

> I think that what is expected will change and it will become clearer once we have the precedent of this year. We have purposefully been fairly lenient in terms of putting this into effect and have not given people super clear guidelines, because we will get a range of replies. And after this year, it will be easier for people to write this because there will be precedent there.

Similarly, for ethical review, as a Nature article put it, 'reviewers were not given specific guidance on what constitutes harm to society'.[290] The hope was that AI researchers and the ethical reviewers would figure out, over time, what kinds of papers were risky or ethically

[289]Nikolas Rose, 'Governing the Enterprising Self', in *The Values of the Enterprise Culture: The Moral Debate*, ed. Paul Heelas and Paul Morris (Routledge London, 1992). An alternative term is 'subjectification' – see Leonard Lawlor and John Nale, eds., *The Cambridge Foucault Lexicon* (Cambridge: Cambridge University Press, 2014), chap. 85, https://doi.org/10.1017/CBO9781139022309.

[290] Davide Castelvecchi, 'Prestigious AI Meeting Takes Steps to Improve Ethics of Research', *Nature* 589, no. 7840 (23 December 2020): 12–13, https://doi.org/10.1038/d41586-020-03611-8.

problematic, as well as the underlying framework for what kinds of risk or ethical issues should be prioritised.

With different people in charge, the organisers' intentions for the 2021 conference was slightly different. There was still an emphasis on encouraging researchers to reflect.[291] However, instead of waiting for knowledge to arise bottom-up, there was greater emphasis on the ethical reviewers as a source of expertise. Rather than researchers' reflections becoming the source of a new ethical framework, they were part of a pedagogical process. An analogy would be to a student writing an essay that is then marked by a teacher. For example, see this statement from two of the organisers:

> Because this process was for educational purposes first, Ethics Reviewers were assigned to each flagged paper, regardless of how likely it was that the paper would be accepted on technical grounds. This was a change from last year and required us to recruit a larger pool of Ethics Reviewers. However, we believe that all authors can benefit from feedback on the ethical aspects of their research and the opportunity to integrate this feedback into future iterations of the work. While it would reduce the burden on Ethics Reviewers, only soliciting ethics reviews for papers likely to be accepted would counteract our goal of making this a constructive process for everyone in the community, rather than just operating as a filter to catch unethical content prior to publication.[292]

[291] One of the people involved tweeted: 'The objective of social impact statements is not to police work but to encourage people to be more actively thoughtful about their work. A good statement is informed & reflective - the goal here is to make ppl think more critically, not judge people on the ethics of their work. . . . [The organisers] have made it clear that the main criteria for success here is "Have you thought enough about this?" above anything else.'

[292] Samy Bengio et al., 'A Retrospective on the NeurIPS 2021 Ethics Review Process', *NeurIPS Blog* (blog), 2022, https://blog.neurips.cc/2021/12/03/a-retrospective-on-the-neurips-2021-ethics-review-process/. A similar quotation from the same post: 'At this early stage of adoption, ethics reviews are meant to be educational, not prohibitive. Our goal is not to police submissions, but instead to prompt reflection. The process that we implemented was intended to support this goal.'

Two kinds of power are contrasted here. The organisers are keen to stress that they are not relying on a gatekeeping kind of power, where the reviewers allow some papers to be published and not others. Instead, they rely on a kind of power that will be quickly recognisable to Foucauldian scholars. The aim is to mould the researchers into good 'subjects'. The researchers, by reflecting on how their work is connected to certain societal impacts, should become enlightened and empowered to play their own role in promoting ethics – for example, by choosing what projects to undertake in future. The same motivation is apparent in the original proposal for BIS, which states:

> There clearly is a massive gap between the real-world impacts of computing research and the positivity with which we in the computing community tend to view our work. We believe that this gap represents **a serious and embarrassing intellectual lapse.**[293]

Researchers' existing non-engagement with how their work affects society is seen as a failure of self-knowledge: they lack sufficient knowledge about their own actions and their proper role in society. This lack of self-knowledge is seen as holding researchers back from acting as responsible custodians of AI technologies.

The role of knowledge with the regime of BIS and ethical review is, therefore, similar but subtly different from the KCRs we encountered in chapters 1 and 2. There, the knowledge was about the misuse of GPT-2 and GPT-3 (actual and potential). That knowledge was a tool for helping OpenAI to decide what level of model access actors should be given. In other

[293]Hecht et al., 'It's Time to Do Something'. (Bold in original.) See also the section on expected outcomes of BIS: (1) 'Increased intellectual rigor'; (2) 'Researchers will be incentivized to change the technologies they create'; (3) 'More support for key governmental policies'; (4) 'Research that mitigates the downsides of existing and new technologies will be more strongly incentivized and motivated'; (4) 'We will be encouraged to engage more with the social science literature, which has a great deal of expertise in understanding the social impacts'.

words, it was knowledge pertaining to an external population (of users), created to help the lab better govern that external population. With BIS and ethical review, the question is still often about external social impacts. But the emphasis is more tilted towards researchers learning to be cognisant of these impacts and their own relation to them, so that they can better govern their own conduct. Knowledge of social impacts becomes a tool for governing researchers as much as it is a tool for governing society.

The question for ethical reviewers then becomes (to quote one of the organisers) 'have you [the authors] thought enough about this?'[294] This is different from a regime where papers are rejected if reviewers forecast that they would cause too much harm to society (which would be more analogous to the KCRs used for GPT-2 and GPT-3). Therefore, the NeurIPS regime treats papers' ethical problems as curable via more in-depth discussion from the authors. This is evident from the fact that eight of the nine papers judged most ethically problematic by the 2021 ethics reviewers were accepted after the authors redrafted the papers.[295]

For example, one of the 'conditionally accepted' papers at NeurIPS 2021 proposed a method for identifying individuals using 'stylometric' data: for example, how they play chess, or how they write.[296] The main example and experiments throughout the paper were about chess, but the paper implied that the methods were generic to the task of stylometric identification. The Program Chairs, taking into account the reviews, including two ethics reviews, concluded that the authors had not sufficiently addressed the ethical issues surrounding methods for

[294] Twitter post.

[295] Bengio et al., 'A Retrospective on the NeurIPS 2021 Ethics Review Process'. They also state: 'In many cases in which issues were identified, Ethics Reviewers simply recommended that authors reflect on the issues and include a discussion of them in the paper, either by expanding the discussion of potential negative societal impacts or being more explicit about limitations of the work.'

[296] Reid McIlroy-Young et al., 'Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in Chess', in *Advances in Neural Information Processing Systems*, ed. M. Ranzato et al., vol. 34 (Curran Associates, Inc., 2021), 24482–97, https://proceedings.neurips.cc/paper/2021/file/ccf8111910291ba472b385e9c5f59099-Paper.pdf.

identifying individuals.[297] They requested a revision to the paper that addressed these issues, and stated that they 'hope that the authors' added discussion could be used to send a strong signal commensurate with the gravity of the situation'. The authors added a section on 'Ethical Considerations' to the paper. This struck a different tone from the paper's introduction. The introduction implies that the authors were using chess as a test-bed, with the ultimate aim being generic tools for stylometric identification. The new argument in the ethics section was now that researchers must study identification in benign domains, like chess, so that we can better understand 'the power of stylometry', including 'investigating and quantifying these risks' before something similar is developed for more risky domains. The paper was then accepted. A simple change to the language of the paper – inventing a new, post hoc justification for the work – was sufficient.

In this incarnation of ethical review, the discourse found in the papers is a key object of ethical review. One of the 'most common' ethical issues identified at the 2021 conference, according to the organisers, was: 'Uncritically emphasising explicitly harmful applications, such as police profiling.'[298] Similarly, at NAACL 2021, one of the questions the ethical reviewers were asked to consider was: 'Do you find any ethical concerns with the way the research is described in this paper (e.g. language essentializing identity categories)?'[299] Most of the questions for ethical reviewers took the form, 'Does the paper describe…?' The ethical problems either come from (a) the assumptions the authors make in their writing; or (b) potential bad applications of the technology, but not in themselves, but rather as something insufficiently discussed in the paper. In this form, ethical review is not a tool for filtering out

[297] The Program Chair wrote: 'I do believe advances of stylometry in identifying chess players is a pretty benign application, however there's little discussion about how general this work might be in its ability to do stylometry in other domains'.
[298] Bengio et al., 'A Retrospective on the NeurIPS 2021 Ethics Review Process'.
[299] Emily M. Bender and Karën Fort, 'NAACL Ethics Review Process Report-Back', *NAACL 2021* (blog), 20 May 2021, https://2021.naacl.org/blog/ethics-review-process-report-back/.

harmful papers, but rather is a forum for incentivising researchers to change what they write in their papers.

Another similar example comes from a paper about text generation, accepted to NeurIPS 2021.[300] The paper proposes an algorithm for steering the text generated by language models like GPT-2 or GPT-3. For the ethics reviewers, the problem with the paper was not that the method could be misused (as per OpenAI's concerns around GPT-2 and GPT-3). Rather, the problem was that the paper discussed this topic in a flippant way. The paper had emphasised the benefits of text generation, to which one of the ethics reviewers responded: 'I urge authors to avoid general optimistic speculations'. The paper had also mentioned (in the vein of Grover: see Chapter 1) that 'AI systems could be leveraged to fight against misleading content and harassing material'. One of the ethics reviewers responded:

> Mitigating the risks of these systems is an extremely complex socio-technical problem that many are working to understand and solve. The fact that this is not acknowledged is a key ethical weakness.

Note how the 'ethical weakness' is specifically that the relevant point had *not been acknowledged*. The reviewer recommended that the authors adopt the language from their own review, and the authors obliged, directly borrowing the language about the 'extremely complex socio-technical problem'. The paper was accepted. This example again shows what is being governed: not risk per se, but how researchers *write about* risk. It is possible that over time this will have an impact on how the research community collectively thinks about

[300] Yufei Wang et al., 'Neural Rule-Execution Tracking Machine For Transformer-Based Text Generation', in *Advances in Neural Information Processing Systems*, ed. M. Ranzato et al., vol. 34 (Curran Associates, Inc., 2021), 16938–50, https://proceedings.neurips.cc/paper/2021/file/8ce241e1ed84937ee48322b170b9b18c-Paper.pdf.

the risks from AI, and measuring any such shift is beyond the scope of my research. My argument is not that the policy is necessarily ineffective, but rather that it does not directly address concerns around the proliferation of dangerous technology.

The BIS and ethical review process, in this more pedagogical format, relies on the expertise of the ethics reviewers. The expertise of the ethics reviewers is what grounds the evaluation of the authors' discussions of ethics. The organisers for NeurIPS 2021 recruited 105 ethics reviewers and catalogued them by their area of expertise, such as 'Privacy and Security'; 'Discrimination / Bias / Fairness Concerns'; and 'Inappropriate Potential Applications & Impact (e.g., human rights concerns)'.[301] The organisers listed, against each category, how many different papers were deemed to fall under each. The central challenge for the organisers was to set up a system where the right papers would be matched with the right ethics reviewers. One problem was that there were 'false positives' and 'false negatives', where the technical reviewers either incorrectly flagged a paper as raising a specific ethical issue, or failed to catch a paper that did. One of the people involved said at a workshop I attended, 'the ML community is not yet at a point where they can identify all the ethical challenges they face'. Whereas the previous year's organisers had imagined a bottom-up process for identifying ethical challenges, the 2021 process revolved more around the notion that ethics reviewers had specific expertises that they could use to educate authors. The organisers of NAACL 2021 took a related approach, and they stated that they 'wanted to ensure that potential societal impacts of work published at NAACL were considered from multiple different cultural perspectives.'[302] Again, the idea is that particular reviewers have certain kinds of expertise – in this case, certain cultural perspectives – that could enrich the discussion around each paper.

[301] Bengio et al., 'A Retrospective on the NeurIPS 2021 Ethics Review Process'.
[302] Bender and Fort, 'NAACL Ethics Review Process Report-Back'.

Table 2 sums up this section's discussion with a comparison against the KCRs for GPT-2 and GPT-3.

	GPT-2 and GPT-3	**Ethical review / BIS**
Aim	Prevent misuse	Make researchers more reflective and informed
Key questions	How will the model be misused?	Have the authors adequately discussed the issue?
Knowledge comes from	Monitoring real-world misuse; experimenting	Researchers' reflections; expertise of ethics reviewers
Object of governance	Actors in broader world	Researchers' sense of responsibility
Remedies	Delay proliferation; control use	Encourage more in-depth discussion from authors

Table 2: Comparison between how the different KCRs operate. The answers are crude and subject to exceptions. The aim is to give a stylised description.

4. Teaching an old dog new tricks

The original idea for BIS had lofty aims. The Hecht et al paper opens:

> The computing research community needs to work much harder to address the downsides of our innovations. Between the erosion of privacy, threats to democracy, and automation's effect on employment (among many other issues), we can no longer simply assume that our research will have a net positive impact on the world.[303]

The risk is pitched on a grand scale, including the future of democracy, and Hecht et al want researchers to make a contribution to reducing such risk. They suggested BIS because it was a tractable intervention in the context of the existing peer review process. Hecht et al state:

> At our inaugural ACM Future of Computing Academy meeting last June, many of us agreed that the computing research community must do more to address the downsides of our innovations. . . . After several months of discussion, an idea for acting on this imperative began to emerge: we can **leverage the gatekeeping functionality of the peer review process**.[304]

In other words, there is a grand ambition (reducing AI risks on a societal level) and a specific opportunity provided by the existing conference publication KCR. As with previous chapters, we see actors practising an *opportunistic* kind of governance, where they reach for the intervention that is available to them (even if it comes with serious limitations). At the same time, note where the opportunity comes from: the existing conference publication KCR. BIS and ethical review are possible because they build upon this existing institutional framework. Therefore, we must ask the historical institutionalist's question: how does the existing framework of conference publication constrain the new initiative? In this section I argue that

[303] Hecht et al., 'It's Time to Do Something'.
[304] Ibid. Bold in original.

certain weaknesses of BIS and ethical review can be traced back to the fact that they are built upon – and therefore tainted by – the existing regime.

I will argue the existing regime of conference publication, as a foundation, causes four problems for BIS and ethical review: (1) looking at individual papers is often the wrong level of analysis for understanding AI risks; (2) peer review incentivises bland ethics statements; (3) the binary decision to accept or reject a paper is a blunt tool; and (4) the research/applications separation (discussed in the previous chapter) still remains. I will consider each of these in turn. These four problems limit the extent to which BIS and ethical review can actually serve to reduce risks from AI, and generally lead to discussions of risk that are insufficiently tethered to reality.

1. Level of analysis

The first issue is the *level of analysis* problem. The conference publication regime works at the level of individual papers. Therefore, so do BIS and ethical review. Authors are expected to discuss the potential impact of their own, specific papers. There is a mismatch here, because technological change in society – including threats to democracy and employment – is normally not closely related to individual papers. When social scientists study the impact of the television on democratic politics,[305] or the impact of the dishwasher on female labour force participation,[306] the independent variable is naturally a technology or application thereof. Perhaps the television was the result of a thousand small innovations, but nonetheless, the impact of those innovations can only be properly understood in the context of the overarching technology of television.

[305] E.g. Sidney Kraus, 'Televised Presidential Debates and Public Policy' (Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers, 2000), /z-wcorg/, http://site.ebrary.com/id/10346756.
[306] E.g. Jeremy Greenwood, *Evolving Households: The Imprint of Technology on Life* (The MIT Press, 2019), https://doi.org/10.7551/mitpress/11268.001.0001.

Predicting the impact of a specific model like GPT-2 is hard enough, as we have seen. Papers at NeurIPS normally do not simply introduce a model, but are generally more methodological in nature. Methodological improvements could be relevant to a wide variety of different kinds of AI systems, from computer vision to NLP to game-playing agents. Most papers also do not add to AI capabilities in a cleanly identifiable way. They make an incremental addition to the toolbox of techniques relevant to building AI.

As a brief example, consider a 2021 paper by Abel at al, *On the Expressivity of Markov Reward*.[307] The paper is about algorithms for rewarding AI agents for good performance. The paper examines the limits of a popular algorithm for rewarding agents, demonstrating that there are some tasks for which the algorithm cannot produce a reward that accurately communicates how the agent is supposed to perform the task. This paper could have broad applicability across the field of reinforcement learning (where agents learn through collecting rewards while interacting with an environment), which is a large topic in AI. Most papers published at conferences like NeurIPS are methods-orientated, like this one.[308] They often study and contribute to the toolbox of algorithms used for building AI systems. A minority of papers are more applied in nature, attempting to solve some specific, real-world problem: at NeurIPS 2019, less than 20% of papers were filed under 'Applications'.[309] Most papers will instead contribute to AI applications in a way that is indirect, incremental, and could unfold over long time horizons.

[307] David Abel et al., 'On the Expressivity of Markov Reward', in *Advances in Neural Information Processing Systems*, ed. M. Ranzato et al., vol. 34 (Curran Associates, Inc., 2021), 7799–7812, https://proceedings.neurips.cc/paper/2021/file/4079016d940210b4ae9ae7d41c4a2065-Paper.pdf. I found this paper by looking at the list of Outstanding Papers on the NeurIPS 2021 website.
[308] See the breakdown by subject area for the 2019 conference here: https://neuripsconf.medium.com/what-we-learned-from-neurips-2019-data-111ab996462c
[309] Ibid.

An even more foundational example is the backpropagation algorithm, which is used to update the weights of a neural network during training. The algorithm was introduced in a 1986 paper.[310] However, the impact of the backpropagation algorithm is still being uncovered today. Before the significance of the algorithm could be appreciated, it needed to be (a) combined with other techniques, and (b) combined with larger amounts of compute than were available in the 1980s and 1990s. The algorithm does not have a specific connection to certain applications, but rather is used for every application that relies upon deep learning. Instead of linking specific papers to specific applications, therefore, an alternative picture is often more accurate: AI as a body of techniques (and knowledge about those techniques) accumulating over time, combining with one another and with other inputs (including compute). This picture can only be seen by zooming out from the paper-by-paper level.

This causes difficulties for BIS and ethical review. KHAL argued that it is difficult to connect specific papers to macro-level changes, even though they are ultimately connected:

> I think it's complicated to think about the relationship between individual papers and the downstream applications. And I don't think it's the case that the research we're doing is completely neutral to the downstream applications. Because like, where does the funding for AI research come from? In the US, the academic research is primarily funded by the military.[311] . . . But it's not something where you can look at the individual papers and say: *this paper is likely to have these effects on the military*, right?

Similarly, on Twitter, one AI researcher criticised BIS on 'level of analysis' grounds:

[310] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, 'Learning Representations by Back-Propagating Errors', *Nature* 323, no. 6088 (1 October 1986): 533–36, https://doi.org/10.1038/323533a0.
[311] I have not been able to find statistics on this question.

> I certainly agree we have an obligation to consider long term impacts, but this generally needs to take place at a different scale than the individual research paper.

The reply from a different researcher, who was defending BIS, was to concede the point but argue that consideration of broader impacts must be 'embedded within the scientific process somewhere' because otherwise researchers would not do it.[312] In other words, the leverage provided by the peer review system provides a rare opportunity for governance, even if it operates at the wrong level of analysis.

Moreover, consider a particular reinforcement learning paper submitted to ICLR 2022.[313] The original ethics statement began by stating that the paper addresses a problem that will be relevant to building AGI.[314] From there, the authors raised the concern that 'AGI might go out of human control in the future'. One of the reviewers of the paper complained that the statement 'lacks nuance', and recommended 'that it be revised to more carefully address *the specific ethical implications of this paper'* (my emphasis).[315] The authors revised the statement to remove the issue altogether:

> [C]urrently we cannot foresee any negative social impact of our work. The algorithm is evaluated on simulated robot environments, thus our experiments would not suffer from discrimination/bias/fairness concerns.

[312] The tweet reads: 'It's definitely possible that individual papers are the wrong time scale for thinking about societal impact. I do worry, though, that if this thinking is not embedded within the scientific process somewhere, most people (even well meaning ones) won't do it'.
[313] Rui Yang et al., 'Rethinking Goal-Conditioned Supervised Learning and Its Connection to Offline RL', *ArXiv:2202.04478 [Cs]*, 13 February 2022, http://arxiv.org/abs/2202.04478.
[314] See: https://openreview.net/forum?id=KJztlfGPdwW&noteId=DlWKGI7UkV6
[315] Ibid.

There are plenty of AI researchers who see AGI as an existential risk to humanity, specifically due to a loss of human control – the reason the authors had given.[316] However, the grand challenge of AGI is unlikely to be solved by a single paper. If authors are confined to narrowly addressing the potential effects of their specific papers in isolation, then that precludes discussion on the possible impacts of AGI. This is unfortunate, because if invented, AGI would likely be an extremely impactful technology.

The 'level of analysis' problem is most evident when we turn to the thorny question of whether some papers should be exempt from the BIS. The NeurIPS 2020 organisers stated on the conference website: 'if your work is very theoretical or is general enough that there is no particular application foreseen, then you are free to write that a Broader Impact discussion is not applicable.' Similarly, one of the organisers tweeted that authors can write that the BIS is not applicable if 'your paper is more foundational - e.g. core ML methods or theory'.

The difficulty is that so many NeurIPS papers could fall under this description. One AI researcher commented on Twitter: 'Isn't NeurIPS primarily for core ML methods or theory?' The justification for exempting these papers therefore proves too much, undermining the BIS enterprise more generally. I asked TAYLOR why the BIS was not applicable to the more foundational papers, and their answer was about the difficulties of pinpointing specific impacts:

> Because somebody who is writing a paper that gives you a proof for a bound on some algorithm or something . . . For that author to spend a lot of time speculating multiple levels away as to how this might get used in an application really puts us immediately

[316] See, for example: Stuart Russell, *Human Compatible: AI and the Problem of Control*, 1st edition (London: Allen Lane, 2019).

> into I think that long range, existential risk sort of speculation, that doesn't feel like it's going to be particularly helpful. I mean . . . it could be useful exercise, but I think that people become very, very uncomfortable with it because it's very, very far away from what they're doing. . . .
>
> . . . [T]he string of causation is so distant that I don't see that it's an actionable exercise. . . . [W[hen you get to that level of the fundamental – math and computational bounds and theory and core algorithms, and things like that – . . . then there isn't any reason to think about positive or negative or any specific outcomes, because it's sort of machine learning very broadly. It's the entire field.

This same argument could be made for many of the papers at NeurIPS – and indeed, perhaps many of the papers that will ultimately prove to be seminal. Because this work is a poor fit for BIS and ethical review, a sizable portion of the most important work is left unaccounted for. My claim is not that the scope of BIS should be extended, but that the regime is not built on solid foundations: the paper-by-paper nature of the existing publication regime causes problems for BIS.

From the extract, I would also highlight the final line about how fundamental papers throw their lot in with the entirety of AI. The response might be that authors should discuss the possible societal impact of AI generally. However, this comes up against a number of difficulties: (a) the authors do not have the time or expertise to make a valuable contribution at that level; (b) the statements would likely be repetitive across many different papers; and (c) it invites the criticism: why does this content need to be in the paper, rather than anywhere else? What justifies tying these reflections to this specific paper?

These three difficulties all stem from the attempt to tie reflections of broader impact to specific papers. In each case, it would be more natural to have large groups of researchers come together in workshops to collectively discuss AI risks and the future of AI. However, realistically, attendance at such a workshop would likely be patchy. I attended a NeurIPS 2020 workshop on 'Navigating the Broader Impacts of AI Research', and got the impression that most of the attendees were already keenly interested in AI ethics. Many of them were not authors at the main conference but came to the online workshop because they were interested in BIS and ethical review or certain risks of AI.

2. Incentive to avoid controversial topics in BIS

A barrier to high-quality, reflective BIS is that authors are anxious to avoid upsetting reviewers. The authors' main goal is to have their paper accepted, and so they do not want to write anything that the reviewers could disagree with.

This difficulty came through during my conversation with KHAL, an academic AI researcher. KHAL first noted their main objection to the BIS, that it 'incentivised low-quality discourse'. KHAL thought this had been borne out in the low quality of statements:

> So, I work on the algorithms for neural nets, which are like, fairly generic – the sorts of things that would support deep learning more broadly. And the sorts of things that people would write in the impact statements: I guess a lot of people would write about climate change – you know, maybe better optimizers will reduce carbon emissions or something like that. And it's not because they thought carefully about it and they did the cost benefit analysis and thought "this is a high-impact thing to do". It's just like, they maybe saw this mentioned on social media or something, and thought that was the sort of thing they're expected to say in this situation.

When I asked why this was the case, KHAL gave an institutional explanation:

> I mean, I think it's because it's part of the review process. Right? Because, what are your incentives when you're an author on the paper? Well, your incentives are to avoid anything that could possibly annoy the reviewers. And, I guess we all have experience with the conference reviewing process, when, you know, Reviewer 2 is convinced that this step of the proof is wrong. And being able to convince them that this step of the proof is actually correct is hard enough in the discussion process. And so if the reviewer just like disagrees with your opinions about some ethical issue . . . are you really willing to have your paper tank over that? And so, what winds up happening is that people just sort of limit themselves to the really uncontroversial things like climate change and unemployment and things like that.

Again, the difficulty is that the researchers' reflections about broader impact are hosted within this specific, existing institutional process of peer review. That brings in certain dynamics – namely, the relationship between the author and the reviewer – that are not necessarily conducive to good discourse.

Relatedly, a few people I spoke to raised the concern that lawyers and PR teams would check over the broader impact statements written by industry researchers. It would be too easy for a journalist to go through the conference proceedings, find a BIS from industry researchers, and write a headline such as, 'Facebook researchers admit that AI is a surveillance technology'. This likely meant that the statements from industry researchers were more bland than they might otherwise be. This undermines the aim of BIS, which was supposed to encourage researchers to discuss serious risks of AI. The problem might be less severe in a

different forum, such as a workshop with Chatham House rules. Again, then, the institutional template provided by the conference publication regime hinders the success of the new KCR.

3. Accept/reject as a lever

In the previous chapter, I argued that the API provides the lab with a fine-grained way of influencing how a model like GPT-3 is built into applications. We already saw that OpenAI's dilemma of whether to publish the scaling laws papers was more difficult because there were no equivalently fine-grained levers for affecting how the paper was received. In the same way, within ethical review, the decision to reject a paper is a very blunt tool.

Partly, the problem is that the decision to reject or accept a paper is binary. As the organisers of ethical review have been keen to stress, papers cannot be classified into 'ethical' and 'unethical' in a binary way.[317] One response could be that a paper should be rejected if it crosses some threshold: for example, reviewers predict that the paper will have a net negative impact on the world, or a significantly negative impact. Such proposals would run into two difficulties: (1) the field currently does not have a widely agreed-upon framework for evaluating risks; and (2) where such frameworks do exist, applying them would reject too many papers.

I want to focus on this latter difficulty, because it especially highlights the bluntness of using the accept/reject decision as a governance tool. Rejection is harsh. The community of researchers attending NeurIPS and other similar conferences would be extremely upset if a large number of papers – say, 50% – were rejected on ethical grounds. Furthermore,

[317] See also Bender and Fort, 'NAACL Ethics Review Process Report-Back'. They state: 'it is important to not treat ethics review as an all-or-nothing proposition. It doesn't make sense to think of papers as wholly "ethical" or wholly "unethical". Likewise, it would be impossible (or at least extremely impractical) to set up a process that would catch all papers that might be deemed "unethical".'

systematically rejecting such a large number of papers would dramatically change the nature of the conference. At NeurIPS 2020, only four papers (out of around 1900 accepted) were rejected on ethical grounds (0.2%). At NeurIPS 2021, it was only one paper (0.04%). (In comparison, nearly 7,000 papers were rejected for academic reasons.)

To further illustrate the point: consider the distinction between papers contributing to AI 'capabilities' and those contributing to AI 'safety' or 'alignment'. CHARLIE, an AI safety researcher, finds this distinction useful as a way of assessing different research projects (even if sometimes it is difficult to draw). They told me: 'I think the correct move for humanity right now, if we were fully coordinated . . . would be to significantly slow down AI research progress and refocus it on alignment.' Therefore, I asked them if the most dangerous papers will often be, crudely, the best papers – the ones that get the most citations and with the awards – because those will often make the biggest contribution to AI capabilities. They replied: 'Yeah, roughly something like that.' CHARLIE's position is comparable to Leo Szilard wanting certain breakthrough nuclear physics results to go unpublished: the risk comes from the capabilities that the papers might ultimately help to unlock.

This is not an extremely niche perspective. The main factor that concerned OpenAI about GPT-2 and GPT-3, as we saw in previous chapters, was that those models took big steps forward in capabilities. One researcher told me that in their opinion the best approach would be to publish work in AI safety, security, and policy, but not work that speeds up progress in AI capabilities generally. Similarly, in the Partnership on AI's white paper on publication norms, one of the factors they list for making research potentially higher risk is if it is paradigm-shifting. This presumably because paradigm-shifting research has high potential to contribute to capabilities progress.

However, rejecting most groundbreaking papers (except those on AI safety) is hardly a realistic option for the NeurIPS organisers.

To add to this analysis, consider how the GPT-3 paper was processed by the ethical review regime. The paper is considered by many people as the biggest step forward in AI capabilities in recent years. As we saw in the previous chapter, OpenAI delayed publishing the preprint, anxious that it would contribute too much to AI capabilities progress without doing the same for AI safety. OpenAI, and those criticising GPT-3, have also been very concerned about both the problem of bias in GPT-3 and the problem of misuse. Nonetheless, the paper was, of course, not rejected on ethical grounds at NeurIPS 2020. Indeed, it was given a Best Paper award. The function of the regime is not to block publication of such work, even if it is viewed by many as posing risks to society.

As discussed in previous chapters, risk-based KCRs must limit their ambitions in accordance with the underlying institutional tools at their disposal. Ethical review demands a framework that can brand 99% of papers as publishable – anything else would be too unpalatable for the NeurIPS community. Ethics-related rejection must be a tool deployed very sparingly. In other words, ideas about risk, and which risks should be targeted by the regime, must be reverse engineered to suit the practical limitations of the KCR.

This is important context for understanding the nature of ethical review currently practised. Above, we saw that ethics reviews do not primarily function as a way of filtering out high-risk papers. An explanation is because that setup would lead to too many rejections. Ethics review is therefore not a proliferation-centric regime (the term I used to describe the staged release of GPT-2). It is not similar to the journal that agreed not to publish certain nuclear physics results in the build-up to World War II (above). Hand-in-hand with this, the regime

is not directly concerned with the level of risk that might result from a piece of research. The task of the ethics reviewer is not to forecast or evaluate the chances of some harm occurring conditional on publication.

The emerging framework of risk is one that is much more compatible with an ethics review system that must reject virtually zero papers. In practice, this means diluting the focus on 'broader impacts' or 'negative societal consequences', and instead focussing on risks that are curable. As we have seen, the discourse contained within papers can be very easily rewritten, preserving the paper's acceptance. Therefore, writing-related risks are given more prominence, such as the risk that the paper contributes to some unhealthy discourse. Often, it is doubtful whether the question being asked is one of risk at all, but rather (as we have seen) authors' attitudes and level of 'education' on a particular topic.

A very similar shift, also away from 'broader impact' / 'negative societal consequences', is towards issues of research ethics. 'Research ethics', to borrow PAI's definition, refers to 'principles surrounding how research is conducted', which may include 'the protection of the welfare of human participants, the responsible handling of data, and other considerations generally covered by IRBs'.[318] The issue of research ethics was within-scope for NeurIPS ethical review.[319] It is another attractor state for the regime to move into, because it is easy for ethics reviewers to comment on. Moreover, as with writing-related concerns, the paper can often be saved by adding an extra paragraph.

[318] Partnership on AI, 'Managing the Risks of AI Research: Six Recommendations for Responsible Publication', 6 May 2021, https://partnershiponai.org/wp-content/uploads/2021/08/PAI-Managing-the-Risks-of-AI-Resesarch-Responsible-Publication.pdf. p.29
[319] NeurIPS 2021, 'Ethics Guidelines'.

For example, one paper submitted to NeurIPS 2021 introduced a new dataset: human participants had their brain activity recorded using an fNIRS headband while they did different tasks.[320] The tasks varied by the amount of mental workload required. AI systems trained on this dataset are supposed to predict an individual's mental workload based on the brain recordings. The paper was flagged for ethics review, but the ethics review did not focus on the potential impact of AI-powered brain-scanning headbands. Instead, the ethics review focussed on the issue of data protection. (This is despite the fact that the research had already gone through IRB approval, that the data was anonymous, and that the participants had consented.)

Another paper demonstrated that AI vision systems could be improved if, as well as being trained on image data, they were also trained on brain scans from monkeys looking at those images.[321] Again, the ethics review functioned as a backup to the IRB process, with the ethics reviewers asking for the paper to include additional details on how the monkeys were treated and details on the IRB approval. Such descriptions were added to the paper. I highlight this example because it shows how easy ethics review can be when the issue of societal impacts of AI technologies is not considered. The ethics reviewers can simply notice that certain details have been left out of the paper's text and ask for them to be added. The paper is not under the threat of rejection, unless the authors have breached research ethics, such as by mistreating the monkeys; and this would be an uncontroversial rejection. This version of ethics review is much easier than, say, evaluating what kind of future the research could make more likely, and conditioning the paper's acceptance on that exercise.

[320] Zhe Huang et al., 'The Tufts FNIRS Mental Workload Dataset & Benchmark for Brain-Computer Interfaces That Generalize', 20 August 2021, https://openreview.net/forum?id=QzNHE7QHhut.
[321] Shahd Safarani et al., 'Towards Robust Vision by Multi-Task Learning on Monkey Visual Cortex', 2021, https://openreview.net/forum?id=3KhhJxaufVF&noteId=2j4NwFHGKop.

To sum up my argument in this subsection: reliance on the existing institution of peer review means that the most powerful lever for ethics review is the 'reject' button; but the lever is too powerful to actually be used. This means that ethics review must become a less ambitious regime, leaving unaddressed the issue of the proliferation of potentially dangerous capabilities.

4. Gap between research and applications

In the previous chapter, I argued that the open source regime creates a separation between research and applications. Exactly the same point can be made about papers. The publication of the paper marks the point at which the research project ends. The paper contains techniques and insights that can be used by others, including for building AI applications in the real world. The paper is a kind of one-way communication between the researcher and the application developer: the researcher hands over techniques and insights, and receives nothing back. Publishing the paper allows the researcher to cash in on their work, because they get career benefits from the publication, without actually applying the techniques themselves. To use an economics term, the ecosystem is 'vertically disintegrated': generally, the research and the applications are made by different organisations.

As I argued in the previous chapter, an important aspect of this research/application separation is epistemic: the researcher does not know who is reading their paper and what they are building. The other aspect is one of control: the researcher cannot influence who builds applications using their insights.

BIS and ethical review are a bolt-on to the conference publication KCR, rather than a fundamental change. This means that the separation between research and applications is still there. There is nothing in BIS or ethical review that gives the researcher new powers to know

and influence applications. Without addressing this fundamental barrier, there are limits to what BIS and ethical review can achieve. Neither the authors nor the ethics reviewers have very good information about what applications will be developed using the paper's insights, or effective tools for shaping those applications.

The epistemic problem is evident from the following passage, where one of the NeurIPS 2021 organisers discusses how researchers should approach ethical reflection. The task is one of *imagination*:[322]

> I think that when we talk about the consequences of our research, I think it is very important to try and imagine having a conversation with people who are adversely affected. . . . Of course, that isn't as good as actually having a participatory dialogue, which is one step further. But I think that when you don't know what to do, there are steps that you can take to move your thinking forwards. Either in the privacy of your own home or ideally in collaboration with other people as well.[323]

Even if the researcher wanted to have a conversation with someone adversely affected by their work (which would be unusual), they would not know where to find such people. There is no list of the different applications that incorporate the researcher's work. Moreover, if the researcher did successfully track down an adversely impacted individual, and learned something, they have no way of stopping the application. Their status as the author on the paper that contributed to the application gives them no entitlement to demand that the application be altered or terminated.[324] The researcher's only option is to shift their research

[322] See also Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford, 'Overcoming Failures of Imagination in AI Infused System Development and Deployment', *ArXiv:2011.13416 [Cs]*, 10 December 2020, http://arxiv.org/abs/2011.13416.

[323] Talk available at: https://ai-broader-impacts-workshop.github.io/#Recordings

[324] Assuming they are not the holder of a patent.

agenda, and hope that their next paper does not also contribute to any harmful applications (again, using their imagination). As a point of contrast, the GPT-3 API relies less on imagination, because OpenAI can see what applications are being built.

This is a fundamental problem for BIS and ethical review. Authors and reviewers are expected to discuss potential negative societal impacts of the work, but in a setting of complete epistemic isolation from any such impacts.

This analysis could provide another explanation for why the regime has evolved to have less emphasis on 'negative societal consequences' and more emphasis on the easier-to-address issues of (a) the paper's writing, and (b) research ethics. Both these issues are within the scope of visibility and influence for researchers. It is possible to know that, *in the course of your research*, you have mistreated a monkey. It is possible to work out a rough, imperfect estimate of the carbon footprint of training the model and include that in the paper. On the writing side, it is possible to revise a paragraph about ethics (especially if the reviewer is willing to make concrete suggestions). It is possible to make your writing 'transparent about decisions such as compensation of crowdworkers' (as the NAACL organisers propose).[325] The relevant risks up for discussion naturally shift towards the ones that can be seen and addressed by the researcher.

In the following extract, the NAACL organisers draw the connection between, on the one hand, the epistemic difficulties around predicting societal impacts, and on the other, the greater emphasis on issues like research ethics:

[325] Bender and Fort, 'NAACL Ethics Review Process Report-Back'.

> Of course, no researcher is in a position to perfectly predict future impacts; thus the emerging best practices involve clear description of known information (energy usage, compensation of crowd workers, privacy protections for data subjects, etc) and thoughtful discussion of potential risks of application of the results.[326]

They are not suggesting that discussion of risky applications should be substituted out – they acknowledge this explicitly. But the passage reflects a certain balance of emphasis, where the risks of applications are de-emphasised (due to epistemic difficulties) and other issues (due to epistemic convenience) come to the fore. Similarly, the same post lists five areas of 'best practice': (1) crowdworker compensation; (2) data scraping; (3) essentialising identity characteristics; (4) the environmental impact of model training; and (5) 'other'. Potential societal impact is discussed under 'other'.

Finally, without tools for directly influencing applications, researchers must rely on framing their papers as being useful or not useful for certain applications. ALEXIS told me that researchers have a responsibility to selectively frame their work in this way:

> [W]here the responsibility comes in is, I guess, how you communicate your research. I think the small things like examples you use, when you put together some slides for your talk, they matter. If you have an example of a violent video game, where people are shooting each other, it does give the wrong impression to people about what these things are useful for – what they're going to be used for ultimately.

[326] Ibid.

> . . . [P]eople who are thinking about [your work],. . . like, *can I use this for my own application?* If you pitch it as a defence-related application, probably you're going to attract more people working on defence.

Accordingly, sometimes ethics reviewers will ask that authors reframe their papers specifically to discourage certain applications. I found a couple of examples of this when reading NeurIPS 2020 ethics reviews. In one case, the paper was about taking a fast MRI scan and using AI to reconstruct what a full MRI scan would have found.[327] One of the ethics reviewers was concerned that the method was not yet ready to be used in real-world medical applications, stating: 'I would not want these techniques to be used in real world medical applications until further validation studies are complete.' Therefore, the Program Chair, accepting the paper, stated that the authors 'should add additional comments that the paper should not be used for medical purposes without subsequent study by medical professionals'.

The other example is the paper on algorithms for guided text generation (mentioned above).[328] The ethics reviewers did not appreciate the paper's 'optimistic speculations' about certain applications, such as (to quote the paper) 'commercial and humanitarian translation services to help overcome language barriers'. One of the ethics reviewers argued that such statements are harmful:

> Statements like these run the risk of encouraging off-the-shelf usages of the method in high-stakes situations that can impact people's livelihood. Instead, the authors must acknowledge that while their method shows promise on several limited benchmarks, deployment in the real world requires a careful analysis of potential societal benefits

[327] Ajil Jalal et al., 'Robust Compressed Sensing MRI with Deep Generative Priors', 2021, https://openreview.net/forum?id=wHoIjrT6MMb&noteId=oqbk6Brs03.
[328] Wang et al., 'Neural Rule-Execution Tracking Machine For Transformer-Based Text Generation'.

and harms (e.g., the harms associated with furthering negative stereotypes against certain vulnerable groups).[329]

Therefore, BIS and ethics review can sometimes function to add warning labels to papers, encouraging application developers to proceed with caution. This is a caveat to my claim, above, that BIS and ethical review primarily regulate authors and their papers. In this case, the aim is to influence the behaviour of application developers (even if via regulating the authors and their papers).

However, application developers have no obligation to heed the warnings found in the paper's BIS – if they even read it. The institutional separation between research and applications means that researchers have very limited tools for influencing applications. If BIS and ethical review regime functions to add warning labels to papers, it is not because there is evidence that this will have a significant impact on applications. Rather, it is because this is one of the few levers that authors have within reach.

Upshot

The upshot of these different difficulties, I would suggest, is that BIS and ethical review have a distorting effect on how AI risks are talked about and understood. Partly, the distortion can be viewed in terms of the kinds of risks that are afforded more or less attention. Ethics reviews are most interested in risks that relate directly to the paper in question, and risks that can be cured through changes to the paper's writing. Ethics reviews are not primarily concerned with risks stemming from the gradual build-up and proliferation of potentially dangerous capabilities. The distortion can also be viewed in terms of the fraction of papers that are

[329] The authors adopted this note of caution into the paper. The wording was almost identical.

deemed problematic by the regime. Ethics reviews, especially, reinforce the idea that only a small handful of papers are problematic. The social construction of risk is not just about branding certain things as risky – it also involves branding other things as safe.[330] Ethics reviews help to position 99+ percent of AI research projects as risk-free.

Some of my interviewees have acknowledged that, in general, the links between certain papers and certain risks is somewhat arbitrary. ALEXIS said:

> I think if I knew anyone in my friend group, or like my peer group at Berkeley was working on autonomous weapons research, they would probably be ostracised, like, pretty likely ostracised. And yet, so many people work on computer vision and stuff that is so, so close to surveillance and defence and so on. And [laughter] no one thinks twice about that.

The regime of BIS and ethics review does not fight against these artificial distinctions, but helps to reinforce them. For example, the NeurIPS 2021 organisers wanted to surface papers for ethics reviews which might have gone erroneously unflagged by the technical reviewers. They therefore ran a 'keyword search late in the review process', looking for papers on 'surveillance' (among other things).[331] However, what counts as a paper 'on' surveillance? In practice, this means finding papers that mention surveillance. But, as we saw in chapter 1, my interviewees would often admit that *most* AI research is ultimately, even if indirectly, relevant to surveillance. The quotation from ALEXIS, above, also reveals this duality between what AI research projects are 'about' and the capabilities to which they ultimately

[330] Steve Maguire and Cynthia Hardy, 'Organizing Processes and the Construction of Risk: A Discursive Approach', *The Academy of Management Journal* 56, no. 1 (2013): 231–55.
[331] Bengio et al., 'A Retrospective on the NeurIPS 2021 Ethics Review Process'.

contribute. Ethical review currently appears more concerned with the former: the surface-level questions around how researchers package their work.

My intention is not to criticise ethics reviews for its own sake, but to highlight something interesting about the relationship between KCRs and risk. You might imagine that KCRs incorporate risks that have undergone some prior epistemic process – for example, a process of contestation between experts. Stephen Hilgartner's 1992 paper on the construction of risk paints a picture of a discursive struggle among specialist communities, often relying on scientific evidence, sometimes with commercial interests at stake.[332] Without disputing this possibility, I have evidence for an alternative picture. Risk-focussed KCR do not just incorporate existing visions of risk, but help to shape how those risks are constructed. We must pay attention to the limits in what risk-focussed KCRs can achieve, especially within their pre-existing institutional contexts, because those limits will affect what kinds of risks can be productively incorporated into the KCR.

Conclusion

I have presented a particular description of how BIS and ethics reviews function. They are a more inward-looking regime than one might have expected. There is less emphasis on controlling the proliferation of dangerous technologies, and more emphasis on researcher education, research ethics, and the sentences written in papers. I have sought to explain this by reference to the constraining effect of the existing conference publication regime.

[332] Stephen Hilgartner, 'The Social Construction of Risk Objects: Or, How to Pry Open Networks of Risk', in *Organizations, Uncertainties, and Risk*, ed. James F. Short and Lee Clarke (Boulder, CO: Westview Press, 1992), 39–53.

Much of these difficulties boil down to the fact that conference publication is a regime of *papers*. Papers have an awkward connection to AI risk, because they contribute to AI capabilities in an incremental and incomplete way. The regime of papers also does not provide a platform for controlling how the papers' insights are used – i.e. the regime provides very limited tools for researchers to steer AI applications. Building a new KCR as a bolt-on to this regime is like building a house in the sand. There is nothing in BIS or ethics review that can cure the deficiencies of the existing conference publication regime as a platform for governing AI risk.

This helps us to understand something fundamental about risk-based KCRs. They are highly dependent on the underlying institutional foundation on which they are built. In this case, we can see how the conference publication regime gives with one hand and takes with another. It provides a fortuitous institutional framework that papers already pass through and are vetted. The main conferences have good coverage across AI research, with researchers viewing conference publication as a bigger career achievement than only uploading a paper to a preprint server. However, at the same time, the existing conference publication regime is the source of the biggest weaknesses of the new ethics regime. In turn, these institutional weaknesses shape what kinds of risks are discussed in impact statements and ethics reviews. While 36% of AI researchers think it plausible that AI could cause a catastrophe on the level of an all-out nuclear war'[333], it is very unnatural to discuss this risk in a broader impact statement, and very difficult for ethics reviewers to take issue with a particular paper as increasing the likelihood of such a catastrophe. In this way, the institutional starting point (the existing publication regime) strongly affects not only the governance regimes that people can build in AI research but also the kinds of risks that can be productively discussed.

[333] Julian Michael et al., 'What Do NLP Researchers Believe? Results of the NLP Community Metasurvey' (arXiv, 26 August 2022), https://doi.org/10.48550/arXiv.2208.12852.

CONCLUSION

When I started drafting the introduction to this thesis, in October 2021, the median prediction on Metaculus (a forecasting community) for the arrival of AGI was 34 years away.[334] Six months on, as I write the conclusion in April 2022, the same prediction has become 28 years – apparently, six years have passed in six months. The same forecast includes a predicted one-in-five chance that AGI comes in the next 10 years. Such is the pace of progress in AI research, and humanity's struggle to make sense of the latest breakthroughs.

A series of new models have been released, too late for me to integrate into the thesis, displaying stronger capabilities than GPT-3.[335] The CEO of OpenAI has publicly mentioned the existence of GPT-4, but this has not yet been released (April 2022). A group of Chinese researchers, coordinated under the Beijing Academy of Artificial Intelligence, has published a 'roadmap' for training large models.[336] It is highly unclear what AI capabilities lie around the corner, and even less clear that these capabilities will be adequately governed.

This thesis has charted the early years of the current era. GPT-2 was a good starting point. On the one hand, it was an early success in the current trend of training large, generic AI systems using unsupervised learning. On the other hand, it represents an early attempt of an AI research lab to exercise control over how the model proliferates, fearing that its capabilities were dangerous.

---

[334] Forecasts available at: https://www.metaculus.com/questions/5121/date-of-first-agi-strong/

[335] Jack W. Rae et al., 'Scaling Language Models: Methods, Analysis & Insights from Training Gopher', *ArXiv:2112.11446 [Cs]*, 21 January 2022, http://arxiv.org/abs/2112.11446; Jordan Hoffmann et al., 'Training Compute-Optimal Large Language Models', *ArXiv:2203.15556 [Cs]*, 29 March 2022, http://arxiv.org/abs/2203.15556; Aakanksha Chowdhery et al., 'PaLM: Scaling Language Modeling with Pathways', *ArXiv:2204.02311 [Cs]*, 19 April 2022, http://arxiv.org/abs/2204.02311; Aditya Ramesh et al., 'Hierarchical Text-Conditional Image Generation with CLIP Latents', *ArXiv:2204.06125 [Cs]*, 12 April 2022, http://arxiv.org/abs/2204.06125.

[336] Sha Yuan et al., 'A Roadmap for Big Model', *ArXiv:2203.14101 [Cs]*, 20 April 2022, http://arxiv.org/abs/2203.14101.

Since GPT-2, we have seen newer forms of governance emerge. The GPT-3 API hinted at a departure from the prevailing approach of open source models, and imported a new, stronger form of control. AI conferences have also started exercising some limited control over risky research papers – even if so far they have only scratched the surface of the relationship between AI papers and risk. Governments have not yet stepped into the governance of AI proliferation, even if regulation of AI applications is starting to emerge.

In this concluding chapter, I will draw together what we have learnt throughout the case study. The chapter is divided into two main sections, addressing two clusters of questions underlying the research. First, I ask: how do risk-based KCRs work? In particular: how are they different from other kinds of KCR? How do they marshall knowledge about risks? And how do they structure relationships between different actors? Second, I ask: what are the limits of governing AI through KCRs? In particular: what are the factors that make governance easier or harder in this area? How does the existing institutional landscape of the field constrain attempts at AI governance? And more speculatively: what would be the right institutional home for the development of AGI?

1. KCRs and risk

1.1 The structure of risk-based KCRs

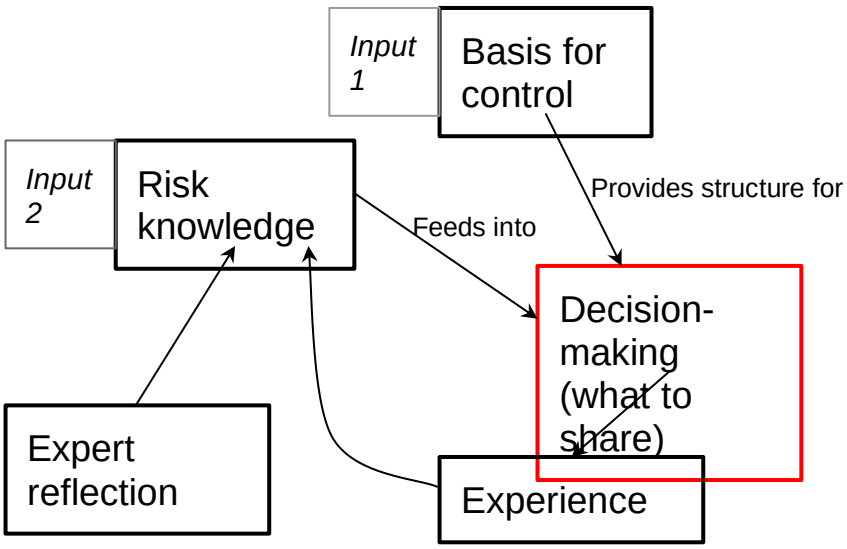


Figure 4 The structure of the risk-based KCRs from the case study.

All of the risk-based KCRs I have studied share a common structure, as set out in Figure 4. The central motif is *informed decision-making*. At their heart, each of the KCRs involve a decision-making process: when to share a given model; what applications to approve; or the acceptance/rejection of conference papers.

Two kinds of input flow into this decision-making process. First, there is the *basis for control*. The basis for control largely comes from the wider institutional context. In the case of conference publication, the basis for control is the conference publication system, including the peer review process. In the case of GPT-2 and GPT-3, the equivalent is that AI research labs are free to decide when and how to share the software they build.[337] The scientific community expects them to publish the work, which means that delaying can be reputationally costly (as was the case for GPT-2) but otherwise the decision is within the lab's prerogative. The basis for control is also subject to technical and practical factors. For

[337] For example, the lab has no institutional obligation to immediately share the work – unlike some of Hilgartner's examples from the Human Genome Project. For example, under the Bermuda Principles, researchers had 24 hours to upload sequence data after generating it.

example, we saw that GPT-3 was harder to replicate than GPT-2, making it amenable to a longer period of control.

The basis for control structures decision-making. It determines what the objects of the decision-making process are: models of different sizes (GPT-2), applications (GPT-3), and papers (conference ethics regimes). It also structures the decision space, i.e. what is being decided for those objects. For example, the decision space for ethical review includes the option of 'conditional accept'. For the GPT-3 API, the decision space includes, for example, requesting that a third party developer make an alteration to their application. For GPT-2, the decision space is more restricted, but could include waiting another couple of months before releasing another model. The basis for control, by structuring decision-making, plays a fundamental role within these different KCRs.

The other input to decision-making is *risk knowledge*. This is knowledge about the risks of AI, and especially the risks of the different objects (e.g. papers) within the decision-making process. We have encountered different kinds of risk knowledge in the case study. In some cases, there are concrete frameworks for categorising the different objects. OpenAI's use case guidelines for GPT-3 are the clearest example, specifying what kinds of applications are likely to be approved. The NeurIPS 2021 organisers also gave a list of different categories of risk, such as 'discrimination and bias', 'privacy and security', 'research integrity', and so on. At other times, risk knowledge can be more abstracted away from low-level decision-making; for example, the arguments for why AI poses an existential risk.[338] Risk knowledge can also be empirical in nature. For example, we saw how with GPT-2, OpenAI was on the lookout for examples of real-world misuse.

[338] E.g. see Nick Bostrom, *Superintelligence : Paths, Dangers, Strategies* (Oxford, UK: Oxford University Press, 2014).

The prominence of risk knowledge in my case study helps to show what is special about risk-based KCRs. All three risk-based KCRs involved mechanisms for producing risk knowledge. The KCRs do not just demand pre-existing risk knowledge, but organise the supply of new risk knowledge. Sometimes this is by eliciting expert reflection. We saw this with the original aim of the NeurIPS broader impact statement, which was supposed to create an evolving body of risk knowledge. In the case of GPT-2 and GPT-3, the KCRs took an experimental approach, partially exposing society to the models and looking for evidence of misuse. Similarly, I also briefly discussed OpenAI's CLIP model (an image classification model) which they stage released. OpenAI staff were surprised by how people were using CLIP (as part of an image generation pipeline) and took that into account when deciding on the release of larger models. Finally, the other approach is to encourage researchers to study the model and identify its risks. We saw this process for both GPT-2 and GPT-3, where OpenAI encouraged outside academics to study risk-relevant topics using the models.

It should not be surprising that risk knowledge is a key feature of risk-based KCRs. The purpose of the regime is to mitigate risk, which is a hazy and moving target. Risks are uncertain and difficult to characterise, especially in the case of emerging technologies. There is much epistemic work to be done in understanding the risks before they can be appropriately mitigated. The AI community does not have settled mappings between specific objects (such as 'large language models' or 'papers about reinforcement learning') and specific risks. The KCRs help to establish these links. The centrality of this process is in contrast with KCRs in other areas, such as patent protection, which involve less decision-making under uncertainty. Patent law sets fixed standards for what kinds of innovation can be patented and what cannot – such as works of art. The patent office might encounter some edge cases where the application of the rules is ambiguous, but there is no equivalent to the AI community's head-

scratching over the question: what even are the risks of this kind of work? The risk-based KCRs from my case study, therefore, all involve a form of decision-making that is hungry for risk knowledge.

However, there is nothing about risk-based KCRs that says they *must* have 'well-informed decision-making' as their organising principle. The secrecy rules of the Manhattan project arguably qualify as a risk-based KCR (the risk being that other countries would develop nuclear weapons). Yet, the policy was one of blanket classification, rather than looking at each document and considering the risks of releasing it to the public (the latter approach only came after the war, with the Smyth Report).[339] The risk-based KCRs from my case study are different because they do not take a blanket approach, but rather try to be selective about what knowledge and artefacts are shared. The closest thing to a blanket policy was probably the Machine Intelligence Research Institute deciding to make their research 'non-disclosed by default'.[340] There is also the fact that many large industry labs have recently trained large language models and not open sourced them without publicly explaining why, as we saw in chapter 3. If a rule of 'do not share large language models' is crystallising among many labs, that would be different from the three KCRs I have studied, which all involve more case-by-case decision-making. Perhaps we should expect the field's norms to become more settled over time, meaning that my case study can be characterised as the early stages of KCR development. AI risk, and how it should relate to AI proliferation, is currently pre-paradigmatic.

[339] Alex Wellerstein, *Restricted Data* (University of Chicago Press, 2021).

[340] Nate Soares, '2018 Update: Our New Research Directions', Machine Intelligence Research Institute, 22 November 2018, https://intelligence.org/2018/11/22/2018-update-our-new-research-directions/.

1.2 Comparing the three KCRs

	Staged release	**API**	**BIS / ethics review**
Aim	Control proliferation to prevent misuse		Responsibilisation of researchers
Basis for control: source	Leading-edge model-building	Leading-edge model-building + the API infrastructure	Peer review system
Basis for control: objects	Models of different sizes	How the model is built into applications	Language in papers
Basis for control: decision space	When to release	Vetting and shaping applications	Requesting revisions to paper
Risk knowledge	Monitoring model use (indirectly); academic studies	Monitoring model use (directly); academic studies	Authors' reflections; expert ethics reviews

Table 3: Comparing the three risk-based KCRs.

Table 3 shows the main points of comparison between the different KCRs. Here, I will further elaborate the comparison. As usual, the purpose is to better understand the KCRs, individually and collectively.

Aims

All three regimes ultimately aim to reduce the negative impacts of AI on society. Both staged release and the API aim to do this very directly. They aim to prevent misuse by preventing people from having access to AI capabilities that they might misuse. It is about controlling what technology to which people have access.

The regime of BIS and ethical review has more numerous and often indirect routes to affecting the impact of AI on society.

First, there is the concern that researchers are too ignorant of the societal consequences of their work. The aim is then to make them more responsible citizens, cognisant of how their work affects society. The hope is that this will lead to changes in behaviour that ultimately cash out in societal impact: for example, the researcher decides their research agenda with societal impact in mind.

Second, there is an aim of discouraging poor research ethics, i.e. harm caused within the research process – such as infringements to individual privacy during data collection. Third, the aim is to control what messages are sent by the papers. For example, one organiser of NAACL identified one class of ethical concerns as: 'the research question being pursued is ill-formed, and posing it (and especially appearing to have answered it) can do harm'.[341] This category includes, for example, papers about predicting an individual's sexual orientation from an image of their face. The aim here is to prevent readers from getting an incorrect impression of the nature of sexual orientation (or the limits to AI capabilities). There is a blurred line with the scientific review of the paper here, because the ethical concern is also a concern of validity: that the assumptions underlying the modelling exercise are incorrect (e.g. they assume too much about the connections between sexual orientation and an individual's face). But at the same time, the regime envisages the paper as a contribution to societal discourse. The idea is that there is a special kind of invalidity: invalidity that is prejudicial

[341] Emily M. Bender, 'Academic Freedom, Academic Integrity, and Ethical Review in NLP', *Medium* (blog), 22 June 2021, https://medium.com/@emilymenonbender/academic-freedom-academic-integrity-and-ethical-review-in-nlp-1db38153cd98. Emphasis removed from original.

against certain individuals. The harm is lending 'legitimacy to ideas that are used to oppress people'.[342]

Finally, there is the aim that comes closest to OpenAI's aim of preventing misuse of GPT-2 and GPT-3. The concern is that the technology advanced in the paper could be used in a way that causes harm: either maliciously, or because the technology contains defects (such as bias). However, as I argued in chapter 3, the regime does not ultimately attempt to address this by influencing what technology gets let out into society. Often, the hope is that, by changing how researchers frame their papers, or the information they put in them, those using the technology will be better informed and therefore will not use it for harm. I argued in chapter 2 that this approach (giving people better information) is not well-suited for all kinds of AI risk, especially misuse.

The aims of the regime are shaped by the regime's basis for control. If the basis for control is not strong enough to support a certain ambition, then that ambition will be sidelined. This might explain why the staged release of GPT-2 and the conference ethics regime attracted interest from different crowds of people. Staged release attracted attention from people who view the dangers of AI as primarily rooted in increasingly powerful AI capabilities (sometimes described as the 'long-term AI governance' community). The conference ethics regime initially sparked interest among such people, but ultimately has shown little evidence of being able to address the proliferation of dangerous capabilities. It has therefore attracted comparatively more interest from people who would describe themselves as working on 'AI ethics'. This community puts much less emphasis on increasingly powerful AI capabilities as the source of AI risk, and is more interested in the aims of the regime as I have stated them.

[342] Ibid.

Basis for control

The KCRs for GPT-2 and GPT-3 overlap in their basis for control. Both involve the lab being at the leading edge of AI capabilities – building models with capabilities that are not already covered by available models. This is an important fact, because the lead of labs like OpenAI is dependent on certain structural and technical factors, which therefore moderate the viability of these KCRs. One factor is the propensity of other labs to open source their work. Another factor is the required inputs to the research project, such as large amounts of compute and large datasets. On both counts, OpenAI's control over GPT-3 was stronger than for GPT-2, as we saw in Chapter 2. The KCR for GPT-3 was less of a hostage to replication than for GPT-2.

Adding to this, the GPT-3 API had a more sophisticated infrastructure for low-level control over how the model was being used. With GPT-3, OpenAI made a stronger departure from the open source paradigm, and instead opted for a regime where users remotely interacted with the model as it ran on OpenAI's servers. This gave the lab the power to vet and monitor applications being built with GPT-3. OpenAI was able to govern GPT-3 at a lower level of analysis than GPT-2. The questions were not about what models to release or keep back, but instead about precisely what kinds of application should be allowed on the platform. Moreover, the deployment is reversible: OpenAI could withdraw GPT-3, limit its scope, or swap it out for a model that they deem safer. That said, terminating applications might go against the lab's newfound business interests.

Further, the API also provided a platform for users to remotely modify the model (via fine-tuning). We saw in chapter 2 that one criticism of staged release was that the lab has no control over how the model is modified and combined with other software, which can be an important determinant of how the model is ultimately used. The API became a one-stop-shop

for using *and* modifying GPT-3, granting a wider basis for control to the lab. Overall, then, the basis for control was stronger for GPT-3 than GPT-2.

The peer review of conference papers, as a basis for control, has some advantage over the GPT-3 API. The API applies to the models trained by just one lab, whereas the conference publication system captures the majority of AI research papers. In this sense, it is a wider basis for control. However, as I argued in chapter 3, rejection is viewed as a very harsh intervention. The conference organisers might, on paper, have a lot of power over the work published at the conference, but in reality they do not have the authority to drastically change the conference by rejecting a wide range of papers on grounds of risk. This political reality weakens the basis of control. The decision space does not, except in extreme circumstances, include rejection, and instead involves requesting revisions to papers. In substance, a lab like OpenAI probably has stronger control over what its employed researchers publish than a conference. The lab still wants to keep its researchers happy, which does factor into decision-making. But nevertheless, to quote CAMPBELL, the lab 'has this corporate structure' and can say: '"I'm paying you $200,000 a year, this is the least you can do"'. The control held by industry labs is currently the strongest power over publication available in the institutional landscape of AI research.

Risk knowledge

KCRs both consume externally-made risk knowledge, and produce it in-house. In terms of the former, the staged release of GPT-2 was informed by prior thinking within AI governance about misuse risk, most notably through the Malicious Use Report and the workshop that went into it. Similarly, OpenAI's guidelines for GPT-3 rely on pre-existing ideas about the risks of AI. For example: the guidelines refer to 'high-stakes domains' like criminal justice; they use the concept of 'human-in-the-loop'; and they reflect pre-existing concerns that AI

will be used to undermine the epistemic health of democracy. Finally, ethics reviewers bring to the table their existing knowledge from working in relevant fields such as AI ethics and AI governance. There are not clear, systematic differences in the kinds of external risk knowledge consumed by the three different KCRs.[343]

Where the three KCRs differ is how they produce risk knowledge. One problem confronting all three is that they aim to confront risks to society, and must therefore produce knowledge about societal impact. This is challenging given the institutional context, especially the existing publication regime, which isolates researchers from society. With GPT-2 and GPT-3, OpenAI confronted this problem directly, taking a very empirical approach to producing risk knowledge. Both regimes relied on a feedback loop of: expose society to the model in some way; monitor for evidence of risk; then build that new risk knowledge back into the decision-making process. With GPT-2, OpenAI got some evidence against their original concern that GPT-2 would be widely used to create fake text online, since this concern was not materialising. With GPT-3, OpenAI over time found more ways in which the model was being misused – for example, finding that users of one of the story-telling applications were creating sexually explicit stories involving children. Over time, the policy team added to its knowledge of how the model might be misused. This kind of knowledge production – added to the new levers for controlling model use – was central to the new relationship between science and society that was instituted by the two KCRs.

Both KCRs, in parallel, also encouraged academic studies of the risks of GPT-2 and GPT-3. In the case of GPT-2, this was more in tension with OpenAI's concerns around proliferation: they struggled to delegate study of the model to more than a handful of research groups. With

[343] OpenAI's background motivations had more to do with concerns about AGI and existential risks, but these concerns were not institutionalised within the KCRs.

GPT-3, the API provided a natural platform for researchers to study the model without needing to be in close conversation with OpenAI staff members. This meant that GPT-3 has become a very well-studied model, at least in terms of its behaviours. Going forward, this kind of academic study could help to quickly identify the safety problems with a particular model, and more generally to foster research that adds to the community's collective understanding of how large models work.

The conference ethics regime took a different approach to producing risk knowledge. Part of the justification for BIS was that authors of papers – as technical experts on the topic – have something to contribute to discussions about its risks. We saw that the original hope of BIS at NeurIPS was that the statements would build up as a body of knowledge. In this framing of the regime, the authors were the knowledge producers. We saw how the regime grew to place greater emphasis on the ethics reviewers, who play an educational role.

This approach to producing risk knowledge is markedly different from staged release or the API. The approach is less empirical in nature, without directly collecting data about the social impact of the paper's techniques. My explanation is structural: the conference publication regime (the basis for the new KCR) provides no mechanism for any such data collection, maintaining the structural separation between science and society. This forces the regime to rely more heavily on authors' 'imaginations', as we saw, although ethics reviewers may also bring to bear more generic empirical knowledge from their disciplines.

At the same time, the regime does not primarily conceive of the paper as the moment when novel AI capabilities are introduced (as OpenAI did with GPT-2 and GPT-3). Therefore, there is less emphasis on trying to forecast the social impact of any such novel capabilities, and hence less need to produce knowledge that would be relevant to such a forecast. Instead,

the paper is conceived of as either (a) a platform where AI researchers frame their work to a wider audience, or (b) an opportune moment for exposing researchers to ethical discussion or obligations. With GPT-2 and GPT-3, OpenAI was more concerned about the proliferation of novel AI capabilities, and that created a strong demand for the production of risk knowledge.

2. The challenges of governing AI

2.1 Governability: opportunities and limits

The level of risk posed by AI will affect how strong the governance regime will need to be. My thesis has not taken a strong position on what these risks are and their relative importance. However, a useful exercise would be to start with the assumption that the most severe concerns – including existential risk to the future of humanity – are well-founded. We can then ask: on present trajectories, is the governance of AI going to be sufficient to address these major risks?

I would suggest the answer is 'no'. None of the regimes from the case study seem equipped to deal with existentially dangerous AI systems. If such AI systems are built in the same institutional context as GPT-3, for example, then it will not be long before a large number of actors can build them across the globe. The GPT-3 API showed how interaction with a single model can be closely governed, but there is no equivalent for the methods that went into building the model.

A range of institutional and technical factors determine the limits of governance in AI research. On the one hand, we have seen how the institutional and technical environment can provide opportunities for actors looking to set up stronger governance regimes. There is the

increasing prominence of industry labs, which have more top-down control than academic labs. There has also been the growth of large, computationally intensive models, which has helped to put these industry labs on the leading edge. So far, these factors have meant that there is a temporary window of time where the lab (and other like-minded labs) can decide how the model should be deployed. As soon as the work is announced, and especially after a paper is published, the clock for independent replication starts ticking.

Other factors make governance harder. Some of these also relate to the nature of AI research outputs. AI is dual-use: the same methods, or the same model, can be useful for a variety of different purposes, some beneficial and others harmful. We saw how the API provides a partial answer to this problem, excelling at 'disaggregation' of the model into different applications. Nevertheless, the problem remains, especially at the level of more fundamental, methodological insights. These are especially dual use (being more generic than models) and especially difficult to control: the field has not yet worked out an equivalent, for fundamental insights, of the arm's length interaction that we saw with GPT-3.

The institutional context also constrains AI governance, especially the development of risk-based KCRs. This has been a major theme throughout the thesis. AI research is first-and-foremost a scientific field, organised around the publication of conference papers. First, this institutional structure makes progress in AI very incremental. It is difficult to pinpoint any particular paper as a source of risk or worthy of control. Second, researchers have strong incentives to publish papers, creating an 'endless drive to produce more knowledge and artefacts' (as I put it in chapter 1, on GPT-2 replications). Third, the field is anarchic: it is full of different research groups who hold very few obligations to one another. This makes it difficult to organise a coordinated governance regime. Fourth, there is a strong separation between the development of AI techniques and their application in the real world. This makes

it very difficult for researchers to know and steer the applications of their work. Finally, a factor that is less about science and more about commercial incentives: cloud computing companies have an incentive to make widely available code that trains cutting edge models on their own hardware. These different factors combine to make it very difficult to control the proliferation of potentially dangerous AI research.

2.2 The distorting effect

The institutional context of AI research plays a strong role in limiting the ambitions of risk-based KCRs. This has two important consequences. First, most directly, risk-based KCRs leave many risks unaddressed. Second, the indirect effect is that risk-based KCRs shape the conversation about risks, amplifying some risks and sidelining others.

A recurring theme through the thesis has been: risk-based KCRs do what they can, subject to structural limits. For example, the GPT-3 API addresses direct model misuse, but did little to prevent proliferation of the fundamental insights. The conference ethics regime focuses on educating researchers about ethics, because the institutional context does not provide effective levers for guiding the proliferation of the paper's insights. To use the language above: the KCR's basis for control has a very strong influence in structuring the KCR and its decision-making. It defines the objects of the decision, as well as the available responses. I would argue that this has been the strongest factor shaping the KCRs from the case study. The process does not start with an appreciation of a certain risk and then, from there, reasoning about what objects must be controlled and how. The process starts with the limits on what objects can be controlled, how they can be controlled, and what risks can be made visible. The process then works backwards, identifying risks that can be addressed using the available tools.

Therefore, the risk-based KCRs in AI research have a tendency to create a little universe of objects, risks, and responses, within which the operation of the KCR appears principled. The GPT-2 regime focussed, for example, on the direct risks of text generation from GPT-2, and on the lab's control over the timing of model release. Part of the background context was that some OpenAI members were concerned about potential existential risks from AGI, but this was not a feature of the KCR. The KCR does not directly produce or consume knowledge about existential risks, for example. The hope was rather that the new KCR was a step in the right direction, shifting norms around sharing research in a way that might ultimately help for building the right KCRs for AGI-relevant research.

Another example is how the conference ethics regime handled the risks of AI surveillance. The issue of surveillance was clearly within the remit of the regime, but was packaged in a way that the regime could handle. The regime was never equipped to process the fact that a very large proportion of AI research is indirectly relevant to building surveillance technologies. Instead, recall the late-stage keyword search for 'surveillance' across the accepted NeurIPS papers. The regime was only equipped to govern how authors *write about* surveillance, and so that was what it did. More generally, risks that could plausibly hinge upon how authors describe their papers floated to the top of the agenda, and risks that depended on the gradual, field-wide accumulation of capabilities sank to the bottom. None of the three risk-based KCRs seriously address the latter, which has a claim to being the most dangerous thing about AI research.

The principled nature of risk-based KCRs is therefore somewhat illusory. They achieve a state of being *internally* principled, and do so by excluding risks that are too difficult to address. Despite having 'well-informed decision-making' as their organising principle, these KCRs are arguably no more rational or principled than a blanket policy. A lab delaying the

publication of all its papers for one year, for example, might seem arbitrary in comparison, but may well turn out to be a closer fit with how AI risk actually emerges. (Of course, such a policy would be insufficient, on its own, to tackle AI risk.)

My argument provides an institutionalist take on the academic question of how risks from technology are socially constructed. As a point of contrast, consider the 2013 study by Maguire and Hardy on the classification of chemicals as safe or toxic.[344] They study the discussions surrounding certain chemicals, and identify various risk assessment practices. For example, 'referencing' involves comparing the chemical to existing bodies of knowledge; whereas 'particularising' involves highlighting the unique properties of a certain chemical. Maguire and Hardy argue that it is through these kinds of practices that the risks of the chemical are socially constructed. I do not deny that these kinds of practices take place, including in AI research – for example, OpenAI referring to established norms in biosecurity and computer security when justifying the staged release of GPT-2. However, I claim that focussing only on these discursive practices misses too much. Instead, we must look further back in the causal chain, and understand the institutional structure within which the conversation about risk is taking place. The conversation about risk within AI research, especially when connected to KCRs, is a reflection of the institutional structure of AI research, and what it does and does not make possible.

For example, advocates of AGI as an existential risk have sometimes referred to an 'Overton window' of AI research, with existential risk possibly falling outside the window. My argument provides one possible explanation. Existential risk-focussed KCRs are unlikely to arise, because they would often involve drastic changes to the field; therefore, existential risk

[344] Steve Maguire and Cynthia Hardy, 'Organizing Processes and the Construction of Risk: A Discursive Approach', *The Academy of Management Journal* 56, no. 1 (2013): 231–55.

does not receive an invite to one of the key forums where AI risks are socially constructed. Recall CHARLIE, an AI safety researcher: 'I think the correct move for humanity right now, if we were fully coordinated . . . would be to significantly slow down AI research progress and refocus it on alignment.' However, there is no lever – no basis for control – within the existing institutional landscape that could accomplish this. Even in the broader impact statements, TAYLOR (a NeurIPS organiser) told me that 'long range, existential risk sort of speculation' would not be helpful because it is not 'an actionable exercise'. The governance of the AI research field (including its KCRs) is not very oriented towards existential risk because that would be very difficult and involve much structural change. Existential risk therefore gets little airtime within mainstream AI governance and AI ethics projects, including risk-based KCRs, which naturally form around more addressable risks. In other words, the institutional context plays a strong role in the process of risk construction.

One take-home from all this analysis might be a pessimistic view of AI governance. This view says: due to structural factors, AI governance cannot succeed in preventing the biggest risks from AI. This is the view taken by Connor Leahy, for example, who argued in our conversation that AI governance is 'hopelessly intractable'.[345] His approach instead was to focus on building technical solutions instead: low-cost methods for building AGI safely. He made an analogy to climate change: instead of getting nation states to agree to reduce emissions (which he saw as intractable), the best strategy is to make technologies like solar panels as cheap and effective as possible.

However, there is room for an alternative conclusion. The conclusion is that AI governance should focus on the structural forces that make certain governance regimes possible or

[345] Instead of 'governance', the word Connor actually used was 'coordination', but I assume he meant this as synonymous.

impossible. I have argued that the institutional context of the scientific field is the most important structural force in this regard. Therefore, in the final section I turn to the question of whether science is the right institutional home for the development of AGI.

2.3 The right institutional home for AGI development

If AGI is an existential risk, there are strong reasons to doubt that AGI development should be organised primarily as a scientific field. The institutional context of science would provide a hostile environment for actors seeking to control a technology like AGI. As we have seen, the AI scientific field has a number of properties that make governance difficult: incremental progress; strong incentives to publish papers above all else; fragmentation across many different research groups who would struggle to coordinate with each other; and isolation from the real-world applications of the research.

Under the scientific regime, the methods for building AGI will emerge one paper at a time. When a technology is built that is recognisably AGI, all other countries and companies with enough access to compute will be just one paper's worth of innovation away from building the same thing. The lab that built the first AGI (if such a line can be drawn) will therefore have very little ability to shape how AGI is built and deployed around the world. The lab may have invested heavily in research on AI alignment, but they will have to hope that all the other actors apply that research. Similarly, the lab may have a plan for how to deploy AGI in a safe way, preventing misuse, but they will have to hope that all the other actors have the same intentions. AGI will be something that happens to the world, rather than something that human actors have agency to shape.

This raises the question: how *should* AGI be developed? I do not have a confident answer to this question, but there is an alternative to the current status quo that should be considered.

The alternative is that AI research becomes more closely integrated with the AI industry more generally. There are plenty of industries where research is comparatively more intertwined with product development, such as chip design and pharmaceuticals. Moving AI research in this direction might remedy the current weaknesses of AI research as a platform for governance, even if it would come with new downsides to be addressed. I will briefly sketch out a vision for what AI research would look like in this more industry-centric future.

The field of AI research and development would be organised around products rather than papers. One class of products would be API access to models. APIs would also be available for training models from scratch, using methods that sit behind a black-box interface (unlike existing frameworks like TensorFlow which are open source). The other main business model would be vertical integration, where technology companies build AI into their products (e.g. autonomous vehicles; recommender systems within social media). Both these business models already exist, but rely heavily on open source AI research. Instead, when an AI company's research team makes an innovation, its design would be kept as a trade secret and built into the company's products. In some cases, these innovations would come from joint research efforts between different AI companies, and then all these companies would be able to apply the innovation to their products.

The AI industry would be regulated by governments. This would include standards on how to safely build AI systems. Such standards could also exist through coordination between the major AI companies. Insights from companies' AI safety teams, e.g. about particular failure modes, would be built into these standards. There would also be government regulation of what kinds of capabilities can be made available – the equivalent of OpenAI's rules around the use of GPT-3.

The academic field of AI research would still exist, but would not be focussed on building new AI capabilities. Instead, there would be even more emphasis on trying to understand AI systems; for example: how neural networks work, and how they acquire capabilities (and safety problems) during training. This is the vision of academic AI research that is oriented towards improving AI safety. AI companies would cooperate with academics (and could be mandated to do so by regulation). They could provide deep API access to models, including different checkpoints throughout the training process (see chapter 2). A fundamental challenge would be detaching knowledge of the 'secret sauce' for building an AI system from safety-relevant understanding of its behaviours, its internal functioning, and its evolution throughout the training run. If these can be severed, then AI academia could make a strong contribution to AI safety, helping AI companies to better understand the AI systems they build and informing safety standards.

The hope would be that this setup retains the upsides of the vision of AI being 'accessible to all', without the downsides. AI could be widely accessible in the sense that business and consumers could have low-cost access to AI services. AI services could also be adaptable, with businesses and consumers adjusting the AI system to fit their needs. However, access to AI would not be unconstrained. There would be rules in place to ensure that AI is used safely. Any existential risks posed by AI would (hopefully) be better understood than today, and this would guide what capabilities can be made available and under what circumstances. If it transpires that advanced AI is less risky than previously feared, the rules could be relaxed.

The structural change outlined here could come with its own potential downsides. One risk is that companies cannot be trusted to build AI safely, and governments cannot be trusted to achieve this via good regulation. On the other hand, the new structure would provide greater opportunities to limit the proliferation of dangerous AI technologies, and control how they

are applied. Ensuring those opportunities are realised would then become a domain of public policy.

Acknowledgements and conflicts of interests

I am very grateful to my supervisor, Bettina Lange, for excellent and patient supervision throughout the DPhil. I am also grateful to my colleagues at the Centre for the Governance of AI where I worked as a researcher alongside my DPhil research. This provided an exciting working environment for my research, and taught me a great deal about AI and AI governance. I am grateful for the ESRC for DPhil funding, and the Long-term Future Fund for some additional funding nearer the end of my DPhil.

Conflicts of interest: On 25 April 2022, I joined DeepMind as full-time Research Scientist. DeepMind and its parent company Alphabet have commercial interests in AI, including the kinds of AI technologies discussed in this dissertation. Two of my interviewees were from DeepMind, and two more from Alphabet more generally. I finished writing the draft thesis before joining DeepMind, but have since made additions in response to comments from my examiners and supervisor.

Appendices

Appendix I: Additional extracts describing the industry-academia relationship

BELLAMY (postdoctoral researcher at a university lab):

> [T]he papers [industry labs] publish tend to be different from academic labs. . . . They can run experiments at a larger scale, so they can train much larger models on a lot more data. So they have these internal data sets, for example, in computer vision or in NLP, that they can use to train their models. . . . And so the academic works, at least the applied ones generally, are usually about, say for example, taking one of the pre-trained models made available by a large industry lab and maybe analysing how exactly it works, like are what are some of its strengths, weaknesses, things like that. So more like analysis, understanding kind of work. Yeah, so I think it definitely led to this maybe like division of labour between industry and academia.

ALEXIS (PhD student at large academic group):

> Yeah, I mean, and then this very, very narrow sense, when it comes to like, pre-trained models being developed in industry, and then being kind of consumed in academia, it's more like: in industry, you're incentivized to, yeah, push capabilities; just come up with a better generative model, it doesn't really matter what you use it for. You just need to like, see better samples or better likelihood scores and so on. And then in academia . . . I guess the value is creativity in terms of developing applications on top of that, or . . . pointing out which problems can now be solved, thanks to these incremental improvements in capabilities. It sounds kind of funny when I put it that

way. Because I'd usually expect the opposite to happen: academia comes up with, you know, like foundational advances, and an industry actually goes and applies it. But if anything, it's kind of backwards in this intellectual sense.

FRANKIE (researcher who was in the process of transitioning from a non-profit lab to an industry lab):

I mean, there's.... *I* did this. I focused a lot more on thinking about data and creating data. What are the problems in existing data? How do we create data better? Largely focused on evaluation, because you can't really do this much with pre-training, because the datasets are so large. So anyway, like, there's that whole, like, how should we think about data in this new world? There's: how do we think about analysis? As you brought up. There's also like, given… let's say, we've done some analysis, and we know what the capabilities of a particular model is at whatever generation we're at; how do we think about, for instance, reducing problems that we care about to problems that GPT-3 can solve? That perhaps there are, like simple reductions that make things that look initially like they're hard for GPT-3 into actually something that's easy for GPT-3?

Tweet:

The ML community is screwed if we keep insisting that scientific inquiry about known algorithms isn't "novel" . . . but that engineering yet another new, incremental algorithm that we know nothing about is great.

Tweet:

If big companies are dominating leaderboards with superior computing resources in #NLProc, what would academia do? Not necessarily a bad thing: we can now focus on answering the WHY questions. We got a v large # of subs of model analysis and interpretability papers at #emnlp2020.

Tweet:

Of course you need an enormous amount of computation and resources. But if your can't make a GPT3 there is loads of critical science to do using and understanding it

Tweet:

I really like [Jakob Foerster's] quote about compassion speed in research – roughly along the lines of "You should work on ideas that will not be crushed by the compute steamroller"

I think a good area is interpreting ML for science–we can't solve this simply by adding more compute!

Appendix II: Five randomly selected BIS from NeurIPS 2020

Forget About the LiDAR: Self-Supervised Depth Estimators with MED Probability Volumes[346]

[346] Juan Luis GonzalezBello and Munchurl Kim, 'Forget About the LiDAR: Self-Supervised Depth Estimators with MED Probability Volumes', in *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates, Inc., 2020), 12626–37, https://proceedings.neurips.cc/paper/2020/hash/951124d4a093eeae83d9726a20295498-Abstract.html.

In this paper, we presented FAL-net, a method to "forget about the LiDAR" for the learning of monocular depth from stereo images. Our approach incorporates our proposed mirrored exponential disparity (MED) probability volumes and a two-stage learning strategy with our novel mirrored occlusion module (MOM). Our MOM computes very realistic occlusion masks to filter out invalid regions due to parallax. Our FAL-net showed superior performance and reduced number of parameters and inference times than the SOTA fully-, semi-, and self- supervised methods. Even though we focused on learning single image depth estimation (SIDE) from stereo pairs, our method can be easily extended when learning from monocular videos. Our MOM can be adopted as long as the network incorporates a disparity probability volume in its output layers and the relative camera poses are known or estimated. The camera-pose information can be integrated into the warping operation g(·) in Eq. (4) to obtain the mirrored occlusions for the corresponding frame pair. What could be at stake here is the exponential quantization, as inverse depths in structure-from-motion (SFM) are defined up to an unknown and inconsistent scale. The ambiguous scale could prevent the network from taking advantage of all disparity levels. A turn-around for this issue is to incorporate velocity supervision, as introduced in PackNet [11], or consistent SFM [27] to fully exploit the exponential quantization. Being depth estimation a low-level computer vision task, we authors do not consider that any ethical implication is involved in our research. However, we believe it is crucial to know if the network consistently under or overestimates depth. The second is considered more critical in robotics systems, in particular, self-driving cars. In this regard, our FAL-net seems to be on the safer side. We measured this by computing the mean median-scaling factor [35] between the GT and our depth estimates. We obtained a mean scale factor of 1.016, indicating that our network detects objects slightly closer than they are. Finally, we would like to remind the reader that, if one wants to use software-based depth estimators for safety-critical systems, all the necessary redundancy checks and safety norms must be followed.

Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation[347]

This work will potentially impact the community in two main ways. Our proposal to use high-accuracy AutoML ensembles followed by model distillation allows practitioners to deploy their favorite models, but obtain significantly better accuracy than they could fitting these models directly to their data in the standard fashion. Our work thus helps realize AutoML's promise of strong performance on diverse data while distilling its complexity. Furthermore, our improved model-agnostic distillation strategy can help facilitate interpretability of accurate-but-opaque predictors by choosing a simple understandable model as the student model. While the majority of enterprise ML applications today involve tabular data and tree models, empirical research on distillation has mostly focused on computer vision applications with only neural network models. Thus, this paper serves a key segment of practitioners that has been overlooked. A major difference in distillation with tabular data are the limited sample sizes of most people's datasets, which means augmentation during distillation is critical. We expect our work to have strong practical impact for these medium/small-scale problems. By allowing practitioners to deploy simpler models that retain the accuracy of their more complex counterparts, our work helps improve the cost of ML inference, the reliability of deployments (student models are less opaque), and may open up new ML applications that were once out of reach due to previously unachievable accuracy-latency limits. The second avenue for impact is theoretical. The dramatic performance of deep networks on modalities such as images, speech and text has not quite been replicated on tabular data; ensemble methods are still the go-to-methods for such data. One reason for this gap is perhaps that it is difficult to discover invariants for tabular data, in contrast to

[347] Rasool Fakoor et al., 'Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation', in *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates, Inc., 2020), 8671–81, https://proceedings.neurips.cc/paper/2020/hash/62d75fb2e3075506e8837d8f55021ab1-Abstract.html.

the pre-baked translation invariance of CNNs for natural images. In the absence of a strong architectural inductive bias, it is important to heavily augment the data to reduce the variance of fitting high-capacity models such as neural networks and handle situations with limited amounts of data. Our work identifies a simple way to achieve this augmentation, where Gibbs sampling is a natural fit that is computationally efficient (because we only need to run a few rounds) and facilitates fine-grained control over the sample-quality vs. the diversity of the augmented samples. Our study of augmentation in the distillation context is different than most existing work on augmentation for supervised learning, where a popular strategy is to use desired invariances that are known a priori to inspire augmentation strategies (since labels are not available for the augmented data in this setting, one typically has to assume each augmented example shares the same label as a real counterpart in the dataset). Concerns. General concerns regarding model distillation include its potential use in "stealing" (cloning) models hidden behind an API. We are not aware of documented occurrences of this practice beyond academic research. This paper does not enhance the capabilities of such attacks as our augmentation strategy to improve distillation requires access to the training data. Another concern is models obtained through distillation may be less reproducible as one needs to repeat both the teacher- training and the student-training exactly. This should be addressed through well-documented code and saving the augmented dataset and all teacher/student/self-attention models to file. A final concern is the role of distillation in model interpretability. Once somebody distills an opaque model into an understandable model that almost retains the average performance of the original model, they may become overconfident that they understand the operating behavior of the opaque model, even though the distilled model may be a poor approximation in certain regions of the feature space (particularly regions poorly represented in the training data due to selection bias). The data augmentation strategy proposed in this paper may actually help mitigate this issue, but is by no means intended to resolve it. For true insight, we recommend

careful analysis of the data/models as opposed to the hands-off AutoML + distillation approach presented here.

Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms[348]

In many domains of application, the collection and use of personal data are activities that can have damaging consequences. Data breaches can cause, on the one hand, major distress for the individuals concerned through identity theft and the publication of private information (such as medical or financial records), and, on the other hand, can lead to financial ruin and legal battles for the organizations that are hacked. The study and development of statistical methodology that respects the privacy of individuals has, therefore, the potential for huge impact on society. As the performance of the available methodolgy improves, the need for analysts to use outdated and unsafe procedures will decrease, leading to a positive impact on the world. However, the existence of information theoretic lower bounds proving that private procedures necessarily have worse performance than their non-private counterparts could slightly discourage the use of private methodology, on the grounds of relative inefficiency. Overall, though, this seems like a small price to pay, and a fuller understanding of the possibilites and limitations of private data analysis should be very positive. In this paper we improve upon existing methodology for locally private goodness-of-fit testing, and provide deeper knowledge on the underlying theory.

[348] Thomas Berrett and Cristina Butucea, 'Locally Private Non-Asymptotic Testing of Discrete Distributions Is Faster Using Interactive Mechanisms', in *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates, Inc., 2020), 3164–73, https://proceedings.neurips.cc/paper/2020/hash/20b02dc95171540bc52912baf3aa709d-Abstract.html.

Transferable Graph Optimizers for ML Compliers[349]

The increasing complexity and diversity of hardware accelerators has made the development of robust and adaptable ML frameworks onerous and time-consuming, often requiring multiple years of effort from hundreds of engineers. In this paper, we demonstrated that many of the optimization problems in such frameworks can be solved efficiently and optimally using a carefully designed learned approach. This has two significant benefits over a heuristic based hand-tuned approach. First, it can potentially save years worth of engineering effort needed to design and maintain the set of heuristics with each new generation of hardware. And second, the improved solutions found using a learned approach can have a multiplicative effect by improving hardware utilization and computational efficiency for all workloads. This increased efficiency may eventually lead to a reduction in the overall carbon footprint for many applications. We also want to highlight a broader effort in the community to use machine learning in the hardware design process[20]. The techniques presented in this paper can be instrumental in evaluating the behavior of benchmark workloads on new and unseen hardware architectures without requiring significant redesign of compilers. Therefore, we believe that the ideas introduced in this paper are an important step that can positively impact the larger ML for Systems research directions. One of the potential downside of the work is the loss of explainability for the choices made by the learned model. An advantage of heuristic based approaches deployed in current systems is the ability to explain the choices made by the algorithm based on the heuristics used. Often it is feasible to "fix" a poor decision by designing a customized heuristic. The current learned approaches to optimization problems including the ones presented in this paper do not lend themselves to explainable results. However, a principled approach to integrating domain specific knowledge of

[349] Yanqi Zhou et al., 'Transferable Graph Optimizers for ML Compilers', in *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates, Inc., 2020), 13844–55, https://proceedings.neurips.cc/paper/2020/hash/9f29450d2eb58feb555078bdefe28aa5-Abstract.html.

compiler developers with the learning based approach with the goal of improving explainability is an interesting direction for future research.

Neural Methods for Point-wise Dependency Estimation[350]

This paper presents methods for estimating point-wise dependency between high-dimensional data using neural networks. This work may benefit the applications that require understanding instance- level dependency. Take adversarial samples detection as an example: we can perform point-wise dependency estimation between data and label, and the ones with low point-wise dependency can be regarded as adversarial samples. We should also be aware of the malicious usage for our framework. For instance, people with bad intentions can use our framework to detect samples that have a high point-wise dependency with their of-interest private attributes. Then, these detected samples may be used for malicious purposes.

[350] Yao-Hung Hubert Tsai et al, 'Neural Methods for Point-wise Dependency Estimation', arXiv:2006.05553 [cs.LG]