On the Opportunities and Risks of Foundation Models


Digest (of the introduction) by Samuel Albanie, June 2022

Disclaimer: Not all views expressed in the report are held by all authors

Slow description

This report is the result of a distributed writing effort (spanning many disciplines)
Outline

• What is a foundation model?
• Social impact and ecosystem
• Norms, incentives and the role of academia
• Stanford report on foundation models
What is a foundation model?

A foundation model is a model trained at broad scale that can adapted to a wide range of downstream tasks.

Examples: BERT, GPT-3, CLIP

The technology is not new:
Self-supervised learning with neural networks

What is new?
Scale and the ability to perform tasks beyond training

Foundation models

Two key ideas underpin the significance of foundation models:

Emergence
- system behaviour is implicitly induced rather than explicitly constructed
- cause of scientific excitement and anxiety of unanticipated consequences

Homogenisation
- consolidation of methodology for building machine learning system across many applications
- provides strong leverage for many tasks, but also creates single points of failure

Image credits/References
(GPT-3) T. Brown et al., “Language models are few-shot learners”, NeurIPS (2020)
(CLIP) A. Radford et al., “Learning transferable visual models from natural language supervision”, ICML (2021)
Emergence and homogenisation

Modern systems targeting AI tend to use machine learning.
The ideas behind machine learning (ML) have been discussed for a long time (Turing, 1948; Samuel, 1959).
Machine learning really began to rise in popularity in 1990s.
It represented a shift in how AI systems were built.
Machine learning does not specify how to solve a task.
Instead, the "how" emerges from the learning process.
Machine learning also represents a step towards homogenisation:
Many applications can be powered by the same learning algorithm.
Complex tasks in NLP/computer vision still required domain experts.
Feature engineering (e.g. SIFT) needed to achieve good performance.

Deep Learning

Slightly more than a decade ago, there was a resurgence of Deep Learning.
Beyond the original algorithms, key factors included:
• GPUs
• Increased data availability
These produced breakthrough results like AlexNet.
Deep learning also represented a shift towards homogenisation:
Instead of hand-crafting features, the same architecture could be used widely.

References
A. M. Turing, "Intelligent Machinery" (1948)
D. Lowe, "Object recognition from local scale-invariant features" ICCV (1999)
Foundation models - origin story

Foundation models are enabled by transfer learning (Bozinovski, 1976)
Take knowledge from one task (e.g. object recognition in images) apply it to another task (e.g. activity recognition in videos)
In deep learning, the dominant paradigm is pretraining:
• train a model on a surrogate task
• adapt by fine-tuning on the task of interest
Foundation models are powerful transfer learners due to their scale
Ingredients of scaling:
• computer hardware improvements (e.g. GPUs and memory)
• Transformer architectures (leverage parallelism, expressivity)
• Availability of training data

Transfer learning saw initial success from supervised pretraining (e.g. ImageNet)
However, annotation cost limited its benefits
In self-supervised learning, the pretraining task is derived from the data
Example: BERT is trained to predict a masked word from its context

Self-supervised tasks are more scalable than supervised tasks (no labelling cost)
Also potentially a richer learning signal: the model predicts part of inputs (which can be very diverse), rather than a label space (which is typically more limited)

References
(Transformers) A. Vaswani et al., “Attention is all you need”, NeurIPS (2017)
## Foundation models - NLP developments

### Self-supervision in NLP

Self-supervision has been particularly productive in NLP. Word embeddings associate words with context-independent vectors:

- **Word representations** (Turian et al., 2010)
- **word2vec** (Mikolov et al., 2013)
- **GloVe** (Pennington et al., 2014)

**Autoregressive** language modelling (contextual representations):

- **seq2seq pretraining** with a language model (Dai et al., 2015)
- **GPT** (Radford et al., 2018)
- **ELMo** (Peters et al., 2018)
- **ULMFit** (Howard et al., 2018)

**Transformers**: BERT, GPT-2, RoBERTa, T5, BART

### BERT as an inflection point

Prior to 2019, self-supervised learning was essentially a sub-area in NLP. After 2019, self-supervised language models became a substrate of NLP, with use of BERT becoming the norm.

This acceptance of the use of a single model for a wide range of tasks marks the start of the **foundation model era**.

Foundation models produce massive levels of homogenisation.

Almost all SotA NLP models are adapted from a handful of sources (BERT, T5 etc.).

**Benefit**: this provides very high leverage.

Improvements in the foundation model yield gains across much of NLP.

It also represents a **liability**.

All systems can inherit the biases of a few foundation models.

### References

- J. Pennington et al., “Glove: Global vectors for word representation”, EMNLP (2014)

- M. Lewis et al., “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”, ACL (2020)
Foundation models - homogenisation

Homogenisation across research communities

Beyond NLP, increasing homogenisation across communities

Transformer-based sequence models are applied to:
- text (BERT)
- images (ViT)
- speech (Mockingjay)
- tabular data (TaBERT)
- protein sequences (ESM-1b)
- organic molecules (C5T5)
- reinforcement learning (Decision Transformer)

A future of unified tools across modalities?

Homogenisation of models

Homogenisation of individual models across research communities (multimodal models)

Examples in vision and language such as CLIP (Radford et al., 2021), DALL-E (Ramesh et al., 2021)

For domains like healthcare, data is naturally multimodal, encouraging multimodal foundation models

Image credits/References

(ViT) A. Dosovitskiy, et al. “An image is worth 16x16 words: Transformers for image recognition at scale”, ICLR (2021)
(TaBERT) P. Yin et al., “TaBERT: Pretraining for joint understanding of textual and tabular data”, arxiv (2020)
(ESM-1b) A. Rives et al., “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”, PNAS (2021)
(C5T5) D. Rothchild et al., “C5T5: Controllable generation of organic molecules with transformers”, arxiv (2021)
(DT) L. Chen et al., “Decision transformer: Reinforcement learning via sequence modeling”, NeurIPS (2021)
(CLIP) A. Radford et al., “Learning transferable visual models from natural language supervision”, ICML (2021)
Foundation models - risks and naming

### Risks of scale, homogenisation and emergence

**Scale**: has played a key role in the emergence of new abilities.

- **GPT-3** (175B params) enables in-context learning by providing a prompt - emergent property not observed in **GPT-2** (1.5B params).

**Homogenisation and emergence** can interact in an unsettling way.

- Homogenisation can bring gains where task-specific data is limited.
- The risk is that flaws are inherited by all adapted models.
- The power of foundation models comes from emergent properties.
- They are thus hard to understand/have unexpected failure modes.
- Since emergence generates uncertainty over capabilities and flaws, aggressive homogenisation is particularly risky.

**Derisking** is the central challenge in developing these models from an ethical and AI safety perspective.

### The naming of Foundation Models

Bommasani et al. introduce the name "foundation models" to describe the recent paradigm shift.

<table>
<thead>
<tr>
<th>Rationale</th>
<th>pre-trained model</th>
<th>self-supervised model</th>
</tr>
</thead>
<tbody>
<tr>
<td>describes technical attributes of these models</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

However, these names do not capture the paradigm shift.

<table>
<thead>
<tr>
<th>language model</th>
<th>general-purpose model</th>
<th>multi-purpose model</th>
</tr>
</thead>
<tbody>
<tr>
<td>is too narrow: foundation models go beyond language capture their flexibility but do not capture their unfinished character and need for adaptation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>task-agnostic model</th>
</tr>
</thead>
<tbody>
<tr>
<td>captures the manner of training, but not the implications for downstream tasks</td>
</tr>
</tbody>
</table>

"Foundation" is chosen to describe the emerging paradigm.

- "Foundation" describes the role these models play: a foundation model is incomplete but serves as a common building block for task-specific models constructed through adaptation.
- "Foundation" also implies the significance of architectural stability, safety and security:
  - *well-constructed foundations* are a solid bedrock for future applications
  - *poorly-constructed foundations* are a recipe for disaster!

At present, little is known about the nature/quality of the foundation that foundation models provide.

**Critical problem** for researchers, foundation model providers, application developers (who build atop foundation models), policymakers and society at large to address.

---

References

R. Bommasani et al., "On the opportunities and risks of foundation models", arxiv (2021)
Outline

• What is a foundation model?
• Social impact and ecosystem
• Norms, incentives and the role of academia
• Stanford report on foundation models
Social impact

Foundation models are scientifically interesting due to their impressive capabilities. But what makes them critical to study is their integration into real-world products. Google search uses models like BERT as a signal.

What is the nature of the social impact of foundation models? How can we responsibly anticipate and tackle ethical/societal considerations? Note: it is often easiest to reason about specific deployments to specific users.

Reasoning about social impact of foundation models in general is challenging.

Research vs Deployment

It is useful to distinguish between:
- Research on foundation models
- Deployment of foundation models

Most public knowledge of foundation models comes through model research:
- academic papers
- demonstrations
- progress on leaderboards

Direct social impact is driven by deployment (private data/proprietary practices):
- Deployments can arise through new products
- They can also arise through upgrades to existing products (e.g., Google search)

Research models are typically not extensively tested:
- They may have unknown failure modes (warning labels can be provided)

Deployed foundation models that affect people’s lives should be more rigorously audited and tested.

GitHub Copilot (OpenAI Codex)

(OpenAI Codex) M. Chen et al., "Evaluating large language models trained on code", arxiv (2021)
Ecosystem

The foundation model ecosystem

To understand the impact of foundation models (both research and deployment), consider the full ecosystem.
The foundation model itself is only one component of an AI system.

Data Creation

- Heavily human-centric process - most data implicitly about people
- Often created by people for other people (emails, photos etc.), measurements of people (e.g. genome), measurements of environments people live in (e.g. satellite images)
- All data has an owner and is created with a purpose
- The purpose may or may not be to train foundation models...

Data Curation

- Data is curated into datasets
- There is no single "natural distribution" (selection and filtering)
- Ensuring data relevance/quality with legal/ethical compliance is important but often challenging (appreciated in industry)

Training

- The celebrated centerpiece of AI research

Adaptation

- Adaptation creates a system that performs some task starting from a foundation model
- It may combine many modules, rules (e.g. restrictions on output space), classifiers (e.g. for toxicity) etc.
- A model that generates toxic content may be tolerable if appropriate precautions are taken downstream

Deployment

- Direct social impact occurs through deployment to people
- There may be value in permitting harmful models in research to advance scientific understanding (with appropriate caution)
- Staged deployments may partially mitigate harms

Image credits/references:
Think ecosystem, act model

<table>
<thead>
<tr>
<th>Ecosystem abstractions</th>
</tr>
</thead>
<tbody>
<tr>
<td>The social impact of foundation models depends on the whole ecosystem.</td>
</tr>
<tr>
<td>However, it is important to reason about the implications of a single model.</td>
</tr>
<tr>
<td>Many researchers and practitioners’ domain of focus is restricted to the model training stage.</td>
</tr>
<tr>
<td>It is difficult to reason about model training in isolation because foundation models are unfinished, intermediate objects.</td>
</tr>
<tr>
<td>By their nature, they can be adapted to downstream tasks (sometimes in unforeseen ways).</td>
</tr>
<tr>
<td>Two things can help:</td>
</tr>
<tr>
<td>• Surrogate metrics for a representative set of downstream evaluation tasks.</td>
</tr>
<tr>
<td>• Documenting these metrics (e.g. via Model Cards).</td>
</tr>
<tr>
<td>Characterising the full potential downstream social impact of foundation models is challenging.</td>
</tr>
<tr>
<td>It requires a deep understanding of both the technological ecosystem and of society itself.</td>
</tr>
</tbody>
</table>

Image credits/references:
(Model Cards) M. Mitchell et al., “Model cards for model reporting”, FAccT (2019)
Outline

• What is a foundation model?
• Social impact and ecosystem
• Norms, incentives and the role of academia
• Stanford report on foundation models
The future of foundation models

Professional norms - "The Ethos of Science" (Merton)

As noted by Merton in a 1942 essay, "science" is often used to describe:

- a set of characteristic methods by which means of knowledge is certified
- a stock of accumulated knowledge stemming from the application of these methods
- a set of cultural values and mores governing the activities termed "scientific" (or any combination of the above)

The "ethos of science" is the complex of values and norms held to be binding on the scientist.

Four institutional imperatives are taken to comprise the ethos of modern science:

1. **Universalism**: truth-claims, whatever their source, are to be subjected to pre-established impersonal criteria.
   - The acceptance or rejection of claims is not to depend on the personal/social attributes of their originator.
   - Pasteur: "The scientist has a homeland, science does not."

2. **Communism**: the findings of science are assigned to the community.
   - An eponymous law does is not the exclusive possession of the discoverer.

3. **Disinterestedness**: scientists are objective and impartial.
   - Often attributed to personality, can also be understood through the lens of institutional incentives.

4. **Organised skepticism**: temporary suspension of judgement and detached scrutiny of beliefs.
   - This is both a methodological and institutional mandate (often bringing science into conflict with other institutions).

References:
R. Bommasani et al., "On the opportunities and risks of foundation models", arxiv (2021)
R. K. Merton, "The Normative Structure of Science" (1942)
# The role of academia and incentives

<table>
<thead>
<tr>
<th>Industry/Academia</th>
<th>Incentives</th>
</tr>
</thead>
<tbody>
<tr>
<td>The technology behind foundation models is based on decades of research</td>
<td>The political economy in which foundation models are developed creates an incentive structure for decision-making at each stage</td>
</tr>
<tr>
<td>This research spans machine learning, NLP, optimisation, computer vision etc.</td>
<td>Market-driven commercial incentives can align well with social benefit</td>
</tr>
<tr>
<td>Contributions have come from both academic and industry labs</td>
<td>However, they can also lead to underinvestment where shareholders cannot capture the value produced by innovation</td>
</tr>
<tr>
<td>Research on building foundation models: almost exclusively in industry</td>
<td>The Gates foundation states that in a previous generation, the market for vaccines worked well for wealthy countries, but not for low-income countries</td>
</tr>
<tr>
<td><strong>A potential role for academia</strong></td>
<td>Commercial incentives can ignore social externalities (Reich et al. 2021):</td>
</tr>
<tr>
<td>The high pace of progress and possibility of centralisation raises issues that may benefit from humanists and social scientists in addition to technologists</td>
<td>• the health of the informational ecosystem for democracy</td>
</tr>
<tr>
<td>Post-hoc audits of ethical/social consequences after design and deployment are insufficient</td>
<td>• environmental cost of computing resources</td>
</tr>
<tr>
<td>Ethical design could instead be infused into technological development from the start</td>
<td>There may be little incentive for companies to create an open ecosystem for developing foundation models that encourages broad participation</td>
</tr>
<tr>
<td>Academic institutions typically host the widest set of disciplines under one roof</td>
<td>By contrast, the research mission of universities is the production and dissemination of knowledge/creation of global public goods (Kerr, 2001)</td>
</tr>
<tr>
<td>They bring together computer scientists with economists, legal scholars, ethicists etc.</td>
<td>Academia can help to ensure that the development of foundation models is aligned with social benefit that may not be incentivised commercially</td>
</tr>
<tr>
<td>Academia may therefore have an important role to play in developing foundation models</td>
<td></td>
</tr>
<tr>
<td>This role could include: promoting social benefit, mitigating harms, determining boundaries</td>
<td></td>
</tr>
</tbody>
</table>

**References:**
- R. Reich et al., "System error: Where big tech went wrong and how we can reboot", Hodder & Stoughton (2021)
- C. Kerr, "The uses of the university", Harvard University Press (2001)
Resource accessibility

<table>
<thead>
<tr>
<th>Trends in machine learning research</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Academia</strong> has not participated fully in the development of foundation models</td>
</tr>
<tr>
<td>Deep learning has benefited tremendously from increased open science/reproducibility</td>
</tr>
<tr>
<td>Public releases of codebases and datasets have become the norm</td>
</tr>
<tr>
<td>Open frameworks such as TensorFlow and PyTorch enabled easier sharing of code</td>
</tr>
<tr>
<td>Foundation models may roll back this trend</td>
</tr>
<tr>
<td>Models may not be released at all (or are restricted to limited API access)</td>
</tr>
<tr>
<td>Training of models may be unavailable to AI researchers due to compute costs</td>
</tr>
<tr>
<td>Some small scale research feasible thanks to smooth scaling laws (Kaplan et al., 2020)</td>
</tr>
<tr>
<td>However, some abilities (like in-context learning) have only been demonstrated at scale</td>
</tr>
<tr>
<td>The study of pretrained models can be useful, and has been productive in NLP</td>
</tr>
<tr>
<td>But this may be insufficient to address limitations of models arising from design/training</td>
</tr>
<tr>
<td>There are community efforts such as EleutherAI and BigScience (hugging face)</td>
</tr>
<tr>
<td>But the gap between private/community efforts is likely to grow, rather than shrink</td>
</tr>
<tr>
<td>For technologies (as search) centralisation/barrier to entry are potent (K. Radinsky, 2015)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Public infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>It may be possible to close the gap through public infrastructure</td>
</tr>
<tr>
<td>We can draw inspiration from:</td>
</tr>
<tr>
<td>• Hubble Space Telescope (16B USD in 2021 terms, according to NASA)</td>
</tr>
<tr>
<td>• Large Hadron Collider (budget of 9B USD, as of 2010)</td>
</tr>
<tr>
<td>There is a US National Research Cloud initiative underway</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Volunteer computing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donated compute from volunteers across many nodes can be effective</td>
</tr>
<tr>
<td>Folding@home illustrated value for protein dynamics (Beberg et al., 2009)</td>
</tr>
<tr>
<td>Learning@home is exploring similar ideas for foundation models</td>
</tr>
<tr>
<td>This approach faces major technical challenges (latency, bandwidth)</td>
</tr>
</tbody>
</table>

References:
- M. Abadi et al., “[TensorFlow]: a system for [Large-Scale] machine learning”, OSDI (2016)
- J. Kaplan et al., “Scaling laws for neural language models”, arxiv (2020)
  (EleutherAI) [https://www.eleuther.ai/](https://www.eleuther.ai/)
  (BigScience HuggingFace) [https://bigscience.huggingface.co/](https://bigscience.huggingface.co/)

(Hubble Space Telescope cost estimate) [https://www.nasa.gov/content/about-facts-hubble-faqs](https://www.nasa.gov/content/about-facts-hubble-faqs)
(LHC cost estimate) [https://en.wikipedia.org/wiki/Large_Hadron_Collider#Cost](https://en.wikipedia.org/wiki/Large_Hadron_Collider#Cost)
(US Research cloud) [https://hai.stanford.edu/policy/national-research-cloud](https://hai.stanford.edu/policy/national-research-cloud)
A. Beberg et al., “Folding@ home: Lessons from eight years of volunteer distributed computing” (2009)
Outline

• What is a foundation model?
• Social impact and ecosystem
• Norms, incentives and the role of academia
• Stanford report on foundation models
A Report on Foundation Models

Overview of the Stanford report

In March 2021, an informal community at Stanford of students, faculty and researchers with interest in some element of foundation models was created. Not just AI researchers, but also experts in healthcare, law, ethics, economics etc. Led to the founding of the Center for Research on Foundation Models (CRFM) at Stanford.

Given gaps in mutual understanding and existing literature, that goal was to:

- provide a fuller picture of foundation models
- identify opportunities and risks of foundation models
- establish a constructive vision for future responsible development of foundation models

The report writing was an experiment with over 100 people from different backgrounds. Much of the report is a survey of existing work that is unified to highlight connections.

The report focuses on four themes relating to foundation models:

<table>
<thead>
<tr>
<th>capabilities</th>
<th>applications</th>
<th>technology</th>
<th>society</th>
</tr>
</thead>
</table>

Caveats: field is evolving rapidly, so much of the coverage is inevitably incomplete. Many applications are discussed, but others (e.g. music, agriculture, finance, natural sciences) are omitted. Other directions (applications to neuroscience, psychology etc.) form possible future work.

Disclaimer: Not all views expressed in the report are held by all authors.

References:
R. Bommasani et al., "On the opportunities and risks of foundation models", arxiv (2021)
## Capabilities

### Language
- Foundation models dominate NLP benchmarks
- There is still a gap between current abilities and humans
- Gap can be studied through lens of linguistic variation
- Variation includes different styles, dialects, languages
- Children more sample efficient than foundation models
- Multimodal signals/grounding may bridge the gap

### Vision
- Computer vision led adoption of deep learning
- Demonstrated benefits of pretraining (e.g. ImageNet)
- CLIP showed major gains from internet scale image+text
- Multimodal/embodied data may enable further progress
- Key challenges in modelling (e.g. videos) & evaluation
- Applications (e.g. healthcare and society) (surveillance)

### Robotics
- Longstanding goal: "generalist robots" for many tasks
- Robotics is anchored to the physical world
- Key challenge: sufficient data of the right form
- Foundation models may allow easier specification and learning of tasks by robots
- Applications (e.g. household); robustness and safety

### Reasoning and search
- Theorem providing/program synthesis - classic problems
- Combinatorial search space means traditional search-based methods are typically intractable
- AlphaGo shows deep networks can guide search space
- Humans also efficiently transfer knowledge across tasks
- Foundation models may help close this gap

### Interaction
- Foundation models lower difficulty threshold for prototyping and building AI applications
- They raise the ceiling for novel user interaction
- This suggests a synergy:
  - developers can provide applications that better fit the user’s needs and values
  - also introduce more dynamic interaction/feedback

### Philosophy
- What could a foundation model understand about the data it is trained on?
- For natural language, different positions can be taken
- Tentative conclusion: skepticism about the capacity of future models to understand language may be premature

---

**Image credits/References:**
- R. Bommasani et al., "On the opportunities and risks of foundation models", arxiv (2021)
- ImageNet
- A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)
Applications

Healthcare and biomedicine

Many tasks require expert knowledge that is costly:
- Healthcare tasks (e.g. disease treatment)
- Biomedical research (e.g. discovery of new therapies)

Foundation models may be able to learn from data across modalities (images, text, molecules)
Could yield benefits in improved sample efficiency
May also allow improved interface design
This could allow patients/providers to interact with AI
Generative abilities of foundation models have potential for open-ended research problems (e.g. drug discovery)
Foundation models also bring risks (e.g. exacerbating historical biases in medical datasets/trials)

Challenges: data sources and privacy (sociotechnical)
Model interpretability and explainability; regulation

Law

Attorneys devote effort to producing coherent narratives and deciphering ambiguous legal standards
Foundation models may provide benefits:
- Legal documents provide data to train on
- Generative abilities could map to generative legal tasks

However, major progress is needed to enable:
- reliable reasoning over multiple sources of information
- generation of truthful long-form documents

Sample efficiency is valuable due to cost of legal experts
Could enable reallocation of resources to justice/service
As with healthcare, privacy will be a key concern

Fundamental advances will be required with respect to:
- provenance of behaviour
- guarantees for factuality of generation

Education

Effective teaching requires reasoning about student cognition and must reflect the learning goals of students
Models may use external information and modalities (textbooks, diagrams, videos) to assist learning:
- generative tasks (problem generation)
- interactive tasks (feedback to teachers)

Sample efficiency may enable adaptive/personalised learning content
Student privacy will then become a key issue
Other factors also become more critical:
- unequal access to technology in education
- technology-aided plagiarism

Image credits/References:
R. Bommasani et al., "On the opportunities and risks of foundation models", arxiv (2021)
### Technology

#### Modelling
5 key attributes underly foundation model architectures:
- **Expressivity**: ability to assimilate real-world information
- **Scalability**: handling large quantities of high dim. data
- **Multimodality**: consume/produce over modalities
- **Memory**: effective knowledge storing and retrieval
- **Compositionally**: generalisation to novel scenarios

#### Adaptation
Foundation models are "unfinished" assets
Adaptation strategies: fine-tuning, prompting
Adaptation can go beyond task specialisation:
- alleviate deficiencies (temporal adaptation)
- constraints (e.g. right to be forgotten under GDPR)
Expansive evaluation protocols will be required

#### Systems
Computer systems are a bottleneck in scaling up data/model size, which appear to correlate with performance
The next generation of foundation models will require co-design of hardware, software, models and algorithms
Co-design is emerging (e.g. retrieval-based architectures)
Practical deployment requires efficient inference

#### Training
Status quo for training: modality-specific objectives
Masking text (BERT); augmented images (SimCLR)
Future training objectives may involve:
- principled selection (a systematic approach)
- domain generality (unified training across sources)
Key design trade-offs (generative/discriminative); goals

#### Evaluation
Foundation models: 1 step removed from specific tasks
New paradigms for evaluation may consider:
- directly measuring inherent capabilities
- adaptation when by controlling for access/resources
- broader evaluation (robustness, fairness, efficiency, environmental impact)

#### Data
Training data is integral to foundation model abilities
The criticality of data is emphasised in data-centric AI
Models have not had much transparency
One path forwards data hub for foundation models
Consideration must be given to selection, curation, documentation, access, visualisation, quality, regulation

---

Image Credits/References:
R. Bommasani et al., "On the opportunities and risks of foundation models", arxiv (2021)
(SimCLR) T. Chen et al., “A simple framework for contrastive learning of visual representations”, ICML (2020)
C. Ré, The Road to Software 2.0 or Data-Centric AI [https://hazyresearch.stanford.edu/data-centric-ai] (2021)
## Technology

### Security
- Foundation models may form a **single point of failure**
- Discovered **security vulnerabilities** (adversarial triggers)
- Privacy risks (e.g. memorisation of training data)
- Generality poses risks for **function creep** (unintended use)
- One view: foundation models as **operating systems**
- Privacy: public data may reduce need for sensitive data

### Robustness to distribution shifts
- Typical ML models are highly sensitive to **distribution shift**
- Foundation models trained on broad data collections appear to offer greater **robustness** to distribution shifts
- However, they are not a **panacea** for robustness
- Key challenges include: extrapolation across time and spurious correlations derived from the training data

### AI Safety and Alignment
- In **deployment**, it is more important that models are:
  - **reliable**
  - **robust**
  - **interpretable**
- **Task**: align models to avoid **misspecified goals/values**
- **Task**: Forecast emergent behaviours (ability to deceive)

### Theory
- The study of foundation models is largely **empirical**
- **Supervised theory** is inadequate, due to the discrepancy between training/adaption phases
- Advances in theory to address this **discrepancy** may yield useful insights

### Interpretability
- Most interpretability methods focus on explaining the behaviour of **task-specific models**
- Foundation models **span tasks** (introducing challenges)
- One lens: the one **model-many models** paradigm
- **Goal**: find extent that the one model (foundation) and its many models (adapted) share decision-making blocks

#### Key concepts for interpretability:
- **explainability** (validity of post hoc explanations)
- **mechanisms** that drive model behaviour
- It is also valuable to consider the **societal impact** of interpretability and non-interpretability

---

**Image credits/References:**
R. Bommasani et al., "On the opportunities and risks of foundation models", arxiv (2021)
**Society**

### Inequity and fairness

ML can contribute to and amplify social **inequity**

For **foundation models**, it is useful to separate:

- **intrinsic biases** (properties in the foundation model)
- **extrinsic harms** (harms in specific applications)

Source tracing to understand ethical/legal responsibility

Mitigations: proactive interventions/reactive recourse

### Environment

Foundation models involve significant training/ emissions

One perspective: amortised cost over re-use

Several factors would be **beneficial** to consider:

- compute-efficient models, hardware, energy grids
- environmental cost as a factor for evaluation
- greater documentation and measurement

### Economics

Foundation models may have **economic impact** due to:

- novel capabilities
- potential applications in wide array of industries

Initial analyses have been conducted to understand implications for productivity, wage inequality, concentration of ownership

### Ethics of scale

Widespread adoption of foundation models poses ethical, political and social concerns

Ethical issues related to scale:

- homogenisation
- concentration of power

How can norms and release strategies address these?

---

**Image credits/References:**

R. Bommasani et al., "On the opportunities and risks of foundation models", arxiv (2021)
# Responses/critiques

<table>
<thead>
<tr>
<th>Blodgett &amp; Madaio - &quot;risks in education&quot;</th>
<th>Malik - &quot;castles in the air&quot;</th>
<th>Marcus &amp; Davis - &quot;A new foundation?&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foundation models may bring benefits, but risk harms</td>
<td>Clearly, these models have been useful (e.g. BERT)</td>
<td>Foundation: bedrock on which something complex is built</td>
</tr>
<tr>
<td>Four risks in an educational setting:</td>
<td>The pretrain and fine-tune paradigm has merits</td>
<td>Programmers can build on an OS with reliability below</td>
</tr>
<tr>
<td><strong>Risks of educational technologies at scale</strong></td>
<td>There are big risks with training on uncurated data</td>
<td>A foundation for AI should provide something similar: reliable use of information, reliable reasoning etc.</td>
</tr>
<tr>
<td>arguments for student benefit often motivate surveillance historically, scaling has not benefited all learners</td>
<td>The name &quot;foundation models&quot; suggests that these models provide a template for all of AI research</td>
<td>But we have stochastic parrots (Bender et al., 2021)</td>
</tr>
<tr>
<td>Technology (e.g. TV, computers) has second-order effects</td>
<td>Subscribes to embodiment hypothesis (cognitive science)</td>
<td>Good at mimicry, but lack depth of understanding</td>
</tr>
<tr>
<td><strong>Risks of homogenisation</strong></td>
<td>&quot;..intelligence emerges in the interaction of an agent with an environment and as the result of sensorimotor activity...&quot; (Smith et al., 2005)</td>
<td>Five serious concerns:</td>
</tr>
<tr>
<td>Homogenisation of pedagogy, ideology, content</td>
<td>Not arguing for only following human development</td>
<td>• Unjustified renaming of pretrained language models</td>
</tr>
<tr>
<td>Data may dictate ideology about what is valuable</td>
<td>But interaction, grounding, acting in a physical world etc. are important parts of AI</td>
<td>• Limited scientific argument (lack of concrete proposals)</td>
</tr>
<tr>
<td><strong>Risks of limited roles of stakeholders in design</strong></td>
<td>Foundation models at present are &quot;castles in the air&quot;</td>
<td>• &quot;Not invented here&quot; attitude</td>
</tr>
<tr>
<td>At odds with educational philosophy (learners' interests shape teachers' choices about what and how to teach)</td>
<td>Strategy: avoid over-investing in current paradigm</td>
<td>• little discussion of work from relevant fields</td>
</tr>
<tr>
<td><strong>Risks of totalising visions of models in education</strong></td>
<td></td>
<td>• other machine learning ideas in the &quot;scrap heap&quot;</td>
</tr>
<tr>
<td>Formalising learning such that it is legible to these models</td>
<td></td>
<td>• Actual impact of foundation models so far is modest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Actively promotes tunnel vision (Bender, 2021)</td>
</tr>
</tbody>
</table>

**References:**
- S. L. Blodgett and M. Madaio., "Risks of AI foundation models in education", arxiv (2021)
- Malik https://crfm.stanford.edu/commentary/2021/10/18/malik.html
- https://crfm.stanford.edu/commentary/2021/10/18/marcus-davis.html
- Marcus & Davis - "A new foundation?" (Bender, 2021) https://twitter.com/emilymbender/status/1430944351358648324
Responses/critiques

**Sastry - "beyond release/not release"

Discussion has focused on release vs not release
There may be other options in the release design space
Different APIs take different approaches:
- (OpenAI) text in & text out
- (Cohere) access to text embeddings
Exposing model guts brings risks and requires trust
Increases the risk of model stealing attacks
This could defeat the goal of constraining access
For healthy governance, we need ways to:
- audit models
- audit the audits
Research into release design space would be valuable

**Steinhardt - "risks of emergent behaviour"

Push emergence/homogenisation further to logical conclusions
The report's use of "emergence" fits self-organising systems
Different definition: qualitative changes arising from quantitative parameter change ("More is different", Anderson 1972)
Applies to both self-organising systems and physical systems
Phase changes: behaviour manifests quickly at thresholds
We should expect behaviour to emerge routinely (and suddenly)
Capabilities like hacking may emerge with little time to respond
Misaligned objectives: deceptive behaviour may also emerge
Homogenisation contributes to inertia, slowing responses
Institutions can take years/decades to respond to technology
When problems are clear, we will be fixing a rocket as it takes off
Alternative: fix rocket while it's on the launchpad (think ahead)
Forecasting AI (Steinhardt, 2021) - can help build a picture
Mitigation strategies alignment (Hendrycks et al., 2021)

References:
https://crfm.stanford.edu/commentary/2021/10/18/sastry.html
(Steinhardt) https://crfm.stanford.edu/commentary/2021/10/18/steinhardt.html
P. Anderson, "More is different: broken symmetry and the nature of the hierarchical structure of science", Science (1972)
(Steinhardt forecasting) https://bounded-regret.ghost.io/ai/forecasting/ (2021)
(Steinhardt ML safety) D. Hendrycks et al., "Unsolved problems in ml safety", arxiv (2021)
## Summary

A *foundation model* is a model trained at broad scale that can be adapted to *a wide range* of downstream tasks.

**Characteristics:** *emergence* and *homogenisation*

These models may have significant *societal impact*.

The *professional norms* are not yet fully formed.

Development has been led by *industry* rather than *academia*.

## Further resources

- The full Stanford report on *foundation models*
- Workshop with *talks* and *panel discussions* on:
  - Opportunities and Responsibilities
  - Technological Foundations
  - Industry and Applications
  - Harms and Societies

---

### References:

R. Bommasani et al., "On the opportunities and risks of foundation models", arxiv (2021)

Workshop on Foundation Models: (Welcome and Introduction): [https://www.youtube.com/embed/RLrjKGN89Fc](https://www.youtube.com/embed/RLrjKGN89Fc)

Workshop on Foundation Models Session I: (Opportunities and Responsibilities): [https://www.youtube.com/embed/Iux1MExMIAk](https://www.youtube.com/embed/Iux1MExMIAk)

Workshop on Foundation Models Session II: (Technological Foundations): [https://www.youtube.com/embed/PNTbvwEqBk](https://www.youtube.com/embed/PNTbvwEqBk)

Workshop on Foundation Models Session III: (Industry and Applications): [https://www.youtube.com/embed/du1YiyHtwKs](https://www.youtube.com/embed/du1YiyHtwKs)

Workshop on Foundation Models Session IV: (Harms and Society): [https://www.youtube.com/embed/7Ze6Y37EAG0](https://www.youtube.com/embed/7Ze6Y37EAG0)