February 2022

Exploring Clusters of Research in Three Areas of AI Safety

Using the CSET Map of Science

CSET Data Brief

AUTHORS
Helen Toner
Ashwin Acharya

Table of Contents

Introduction

Within the field of artificial intelligence, a growing area of focus is AI safety: research to identify, prevent, and mitigate unintended behavior in AI systems. If AI is ever to grow into its potential as a general-purpose technology and be deployed in high-stakes settings, substantial progress in tackling AI safety challenges will be required.

This brief investigates the development of AI safety as a research field (or set of research fields) by making use of the CSET Map of Science derived from CSET's research clusters and merged corpus of scholarly literature.[1] This resource was constructed by grouping research publications into citation-based "research clusters," then visualizing those clusters based on the strength of citation connections. An interactive version of the Map of Science is available online at sciencemap.cset.tech. We used it to investigate how AI safety is beginning to emerge within the larger ecosystem of machine learning research, including by exploring how research clusters related to AI safety have grown over time, how active different countries are in different areas, and which publications have been particularly influential.

What is AI Safety?

Terminology has proliferated in discussions of how—broadly speaking—to build and deploy AI systems in ways that benefit humanity: trustworthy AI, responsible AI, beneficial AI, friendly AI, ethical AI, safe AI . . . the list seems to grow longer each year.

This brief focuses on three sub-areas within "AI safety," a term that has come to refer primarily to technical research (i.e., not legal, political, social, etc. research) that aims to identify and avoid unintended AI behavior. AI safety research primarily seeks to make progress on technical aspects of the many socio-technical challenges that have come along with progress in machine learning over the past decade.

Like all of the terms in question, AI safety does not have clearly defined boundaries; instead, it is a fuzzy catch-all for a range of

different research problems and approaches. For the purposes of this brief, we searched for research clusters that relate to three types of work that are highly relevant to AI safety: robustness, interpretability, and reward learning.[2] We think of these categories as follows:

> **Robustness** research focuses on AI systems that seem to work well in general but fail in certain circumstances. This includes identifying and defending against deliberate attacks, such as the use of adversarial examples (giving the system inputs intentionally designed to cause it to fail), data poisoning (manipulating training data in order to cause an AI model to learn the wrong thing), and other techniques.[3] It also includes making models more robust to incidental (i.e., not deliberate) failures, such as a model being used in a different setting from what it was trained for (known as being "out of distribution"). Robustness research can seek to identify failure modes, find ways to prevent them, or develop tools to show that a given system will work robustly under a given set of assumptions.
>
> **Interpretability** research aims to help us understand the inner workings of machine learning models. Modern machine learning techniques (especially deep learning, also known as deep neural networks) are famous for being "black boxes" whose functioning is opaque to us. In reality, it's easy enough to look inside the black box, but what we find there—rows upon rows of numbers ("parameters")—is extremely difficult to make sense of. Interpretability research aims to build tools and approaches that convert the millions or billions of parameters in a machine learning model into forms that allow humans to grasp what's going on.[4]
>
> **Reward learning** research seeks to expand the toolbox for how we tell machine learning systems what we want them to do. The standard approach to training a model involves specifying an objective function (or reward function): typically, to maximize something, such as accuracy in labeling examples from a training dataset. This approach works well in settings where we can identify metrics and

training data that closely track what we want, but can lead to problems in more complex situations.[5] Reward learning is one set of approaches that tries to mitigate these problems. Instead of directly specifying an objective, these approaches work by setting up the machine learning model to learn not only how to meet its objective but what its objective should be.

In summary: robustness is concerned with the quality and reliability of outcomes in AI applications; interpretability helps us better predict outcomes and improve learning and accountability when poor outcomes are observed; and reward learning helps reduce the risk of disconnects between intended and observed outcomes. Note that these categories are not intended to represent an exhaustive breakdown of all types of AI safety work. Future analysis could consider additional areas.

Methodology

This brief makes use of CSET's Map of Science to investigate what AI safety research looks like in practice so far. As described above, the Map of Science consists of citation-based research clusters. We sought to locate AI safety research within the map by identifying research clusters that contained a substantial number of publications that appeared to us to be relevant to the three areas of AI safety described above. Our methodology is laid out in more detail in Appendix B. In brief, we followed four steps:

1. On the basis of three documents that summarize recent progress in AI safety,[6] as well as the papers cited therein, we identified three categories of AI safety research to search for (robustness, interpretability, and reward learning, as introduced above), then identified words and phrases used to describe research in each category, excluding terms which were too broad to be useful in a keyword search. For example, DeepMind's description of robustness research mentions that it aims to address safety problems that arise from distributional shift, adversarial inputs, and unsafe exploration. We included the latter two terms as keywords, but did not include "distributional shift" as it is used across

many areas of machine learning. See Appendix B for the full list of keywords used.

2. Using these keywords, we created a list of 42 clusters in which 5 percent or more of papers contained keywords from our list, and in which 50 percent or more of the papers were AI papers.[7]

3. Next, we manually screened each listed cluster by reading the titles of 20–30 papers in the CSET Map of Science summary for each cluster, all from the last five years: the 10 most-cited papers, 10 "core papers" (those with most citation linkages to other papers in the cluster), and up to 10 review papers (where these existed). Based on these titles, we kept clusters that appeared to be focused on a safety-related topic and excluded those that appeared to mainly contain research on other topics (including clusters primarily related to using AI for spelling correction, visual navigation, electronic health records, and a wide range of other topics). After this screening, we had a shortlist of 13 clusters.

4. Finally, we manually screened the titles and abstracts of a random sample of papers from each shortlisted cluster in order to estimate what proportion of papers in the cluster appeared to be relevant to AI safety. We discarded clusters where less than 40 percent of the papers appeared to be relevant; more details on these discarded clusters are available in Appendix B. This left us with a final list of eight clusters.

The underlying logic of this methodology is to use the CSET Map of Science to explore how research related to AI safety is emerging, in practice, as part of the larger ecosystem of scientific research. Where existing analyses of AI safety generally take a more high-level, conceptual approach to describing the space, we aim to use those conceptual descriptions as a starting point to discover where and by whom AI safety–related research is being done and how it relates to other machine learning and computer science research. Steps 1 and 2—creating a keyword list and identifying clusters where even a small share of papers include keywords from the

list—gave us a long, likely over-inclusive list of clusters to use as a starting point. Steps 3 and 4 manually filtered that list down to only those clusters that appeared to group around an AI safety–related topic. This approach allowed us to find clusters that a keyword search alone might have missed (since some clusters only included a small proportion of papers using our keywords, plus many related papers using different terminology) without having to include large numbers of irrelevant clusters (due to the hand screening). Of course, the validity of this methodology is still fairly dependent on the quality of the initial keyword search; see Appendix B for more details on keywords used.

After Step 4, we were left with eight clusters containing 15,197 papers: three clusters that appeared to contain a significant amount of robustness-related research, two for interpretability, and three for reward learning. The rest of this brief describes these eight clusters in more detail. In Appendix A, we also describe a bonus cluster we found along the way, which appears to focus on less technical aspects of AI safety and the future of AI more generally.

We note two additional points for readers to keep in mind:

- Our intention in this analysis is to explore how AI safety–related topics show up within CSET's Map of Science. This required making a range of assumptions and subjective judgement calls about what counted as "AI safety" within the categories we considered. Throughout, we aim to expose those assumptions and judgements to the reader, and we therefore encourage readers to treat this brief as an initial, exploratory look at this area rather than a definitive analysis. It is worth noting that the clusters we include do not exclusively contain AI safety research, nor are they a comprehensive sample of AI safety research. Each of our clusters includes at least some papers that we would not classify as AI safety–related, and it is likely that plenty of other relevant papers are distributed across other clusters that did not meet the criteria described above. (For instance, some work on reward learning that uses deep reinforcement learning models might fall into a cluster that is primarily

about regular deep reinforcement learning, which we would not count as an AI safety–related cluster.)

- CSET's Map of Science is a living dataset: because new papers are constantly being published, and these new papers cite older papers, any method of clustering scientific publications based on citation linkages will naturally need to be updated over time. Between the time we identified the clusters in this brief and when the brief was published, the map underwent a major update, so the AI safety clusters we describe here no longer exist in their earlier form. Fortunately, for each of our clusters there exists a corresponding cluster in the new map that is fairly similar. We therefore provide these "closest match" cluster IDs throughout for readers who wish to explore our findings in CSET's online interactive version of the CSET Map of Science (see sciencemap.cset.tech).

Paper Roadmap

In what follows, we explore the three categories of AI safety research we identified—robustness, interpretability, and reward learning. For each category, we introduce the relevant research clusters we found, describe a small number of the most-cited papers within the category (which we call "foundational papers," since it usually appears that they came to be a touchstone for work in that area), and show the growth of the area over time by region.

Robustness

Clusters Identified

We identified three research clusters that appeared to contain a substantial amount of robustness-related research. Basic information about these clusters is summarized in Table 1.

Table 1: Research clusters with a substantial amount of robustness research ("robustness-related clusters")

Original cluster	Rough description of papers in cluster	# of papers	Est. % relevant papers (+/- 10%)[8]	Most similar cluster in new Map of Science
#3735	Almost exclusively research on creating and defending against adversarial examples	4,406	96%	#2381
#48621	Mix of adversarial examples, data poisoning, backdoor attacks; also includes some work on using AI for cybersecurity applications	967	82%	#26042
#83739	A range of approaches for testing and verifying the performance of ML systems, plus some work on adversarial examples	781	79%	#32715

Source: Authors' analysis of the CSET Map of Science derived from CSET's research clusters and merged corpus of scholarly literature.[9]
Note: The first column gives the ID numbers for the clusters used in our analysis, which was based on the Map of Science before its major update in August 2021. Interested readers can use the cluster IDs in the final column to explore the most similar clusters in the Map of Science after this update.

Based on the research clusters we identified, this category appears to be dominated by research on so-called "adversarial examples," also known as "evasion attacks," for machine learning models.

Adversarial examples are inputs that are intentionally designed to cause machine learning systems to fail. The classic examples in the literature are photos that have been altered at the pixel level so that they plainly look like one thing to humans but are classified by image models as something else. Other examples include physical alterations—such as using small, innocuous-looking stickers to cause a machine learning system to misidentify a stop sign as a speed limit sign—as well as attacks on systems for speech recognition, natural language processing, and other purposes.[10]

All three of the clusters we found included some work on adversarial examples, and by far the largest cluster (#3735) contained papers almost exclusively on this topic. While the other two clusters contained a substantial amount of research on other robustness-related topics, such as data poisoning, testing, and verification, adversarial examples appear to be the hottest robustness-related topic at present.

Foundational Papers

Accordingly, the two most highly cited publications within these clusters are two of the papers that first introduced the concept of adversarial examples. Published in 2014 and 2015 respectively, the papers can be read as a pair: "Intriguing Properties of Neural Networks" (Szegedy et al., 2014) and "Explaining and Harnessing Adversarial Examples" (Goodfellow et al., 2015).[11]

As its title implies, "Intriguing Properties of Neural Networks" discusses several interesting features of deep neural networks, including a discovery that was new at the time: that imperceptible changes to input images could cause networks to classify them completely incorrectly. The authors show, for example, images of a school bus and a fluffy white dog that their image classifier confidently assigns to the category "ostrich."

"Explaining and Harnessing Adversarial Examples," written by two of the authors of the previous paper with one additional collaborator, builds further on this idea, describing the problem of adversarial examples in more detail and offering a possible explanation for why they arise. The more than 6,000 citations this

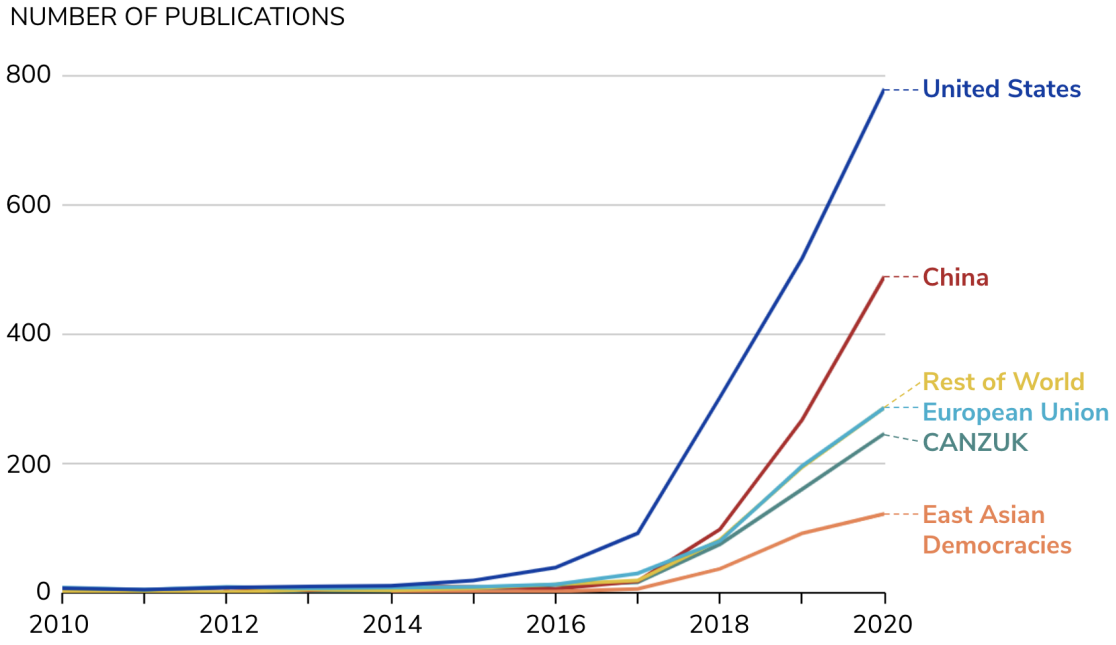
paper has received suggest that it has become the go-to summary of early work on adversarial examples.

Growth over Time

A chart of the number of papers published in these clusters (see Figure 1) shows how dramatically robustness has grown as an area over the last 10 years, with the clusters we identified containing almost no papers before 2014. As one might expect given the timeline of the two papers described above, research in these clusters began to grow slowly starting in around 2015, then really took off around 2017.

Both the United States and China saw rapid growth between 2018 and 2020, though U.S. papers still clearly outnumbered Chinese papers in the clusters we considered.

Figure 1: Papers published in robustness-related clusters over time



Source: Authors' analysis of the CSET Map of Science.
Note: The "Rest of World" category grew at an extremely similar rate to the EU, so the relevant line is hard to see in this figure. Most individual countries' publication numbers are much lower than those of the United States and China, so we use country and regional groupings to make these charts easier to interpret: "CANZUK" refers to Canada, Australia, New Zealand, and the United Kingdom; "East Asian Democracies" groups Japan, South Korea, and Taiwan.

Interpretability

Clusters Identified

We found two research clusters that contained large amounts of research we considered to be focused on interpretability.

Table 2: Research clusters with a substantial amount of interpretability research ("interpretability-related clusters")

Original cluster	Rough description of papers in cluster	# of papers	Est. % relevant papers (+/- 10%)	Most similar cluster in new Map of Science
#9891	Work using a range of techniques to improve the interpretability of machine learning models, especially neural networks	3,521	86%	#5109
#17893	Extracting decision rules from neural networks	1,483	79%	#15771

Source: Authors' analysis of the CSET Map of Science.
Note: The first column gives the ID numbers for the clusters used in our analysis, which was based on the Map of Science before its major update in August 2021. Interested readers can use the cluster IDs in the final column to explore the most similar clusters in the Map of Science after this update.

Determining what research counted as relevant here was far more of a judgment call than for robustness. We were looking for research that aimed to make it easier to understand how AI systems work, a standard both of the clusters in Table 2 clearly met.

Clusters that we shortlisted but ultimately excluded covered many different topics, including:

- Designing robots to move in ways that allow humans to infer what they'll do next

- So-called “disentanglement” research, which aims to train models that map onto the real world more clearly, and which has benefits for interpretability but is often pursued for other reasons

- Data visualization methods that make it easier to explore data and build models

One particularly unclear case was a cluster filled with research on models trained for natural language processing (NLP) tasks. This cluster included some work on how language is represented in these models, which is relevant to interpretability (because it lets us understand how the models work), as well as more general research on better ways to build and test NLP models. See Appendix B for more information on all of these shortlisted clusters.

Foundational Papers

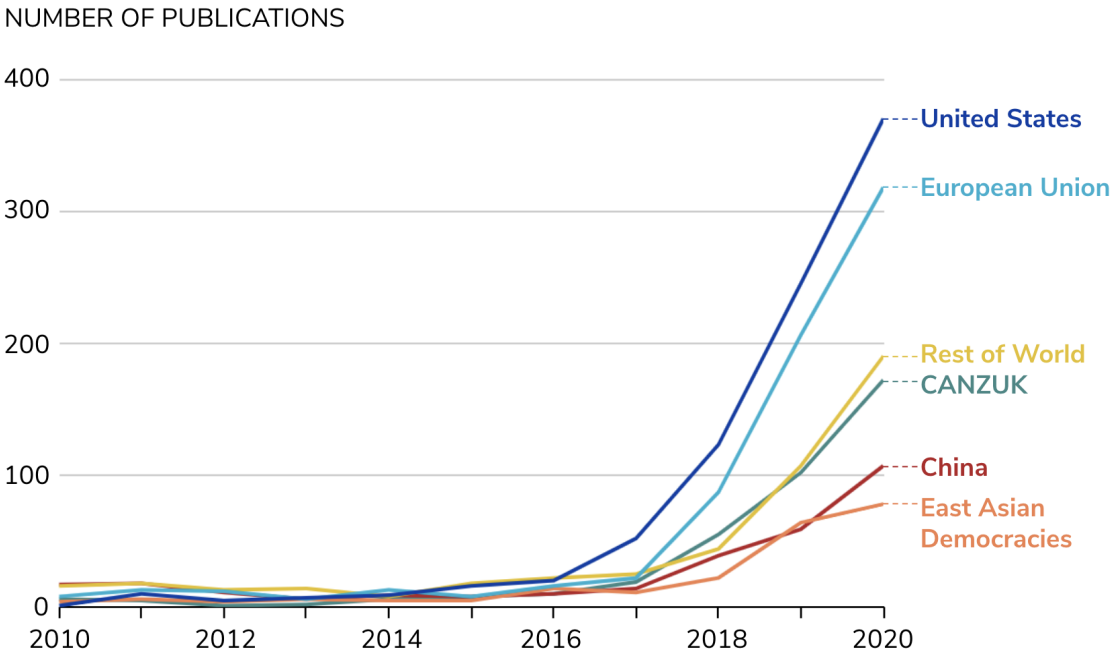
Within the interpretability-related clusters we identified, one paper stands out by far as the most cited, with more than twice as many citations as the next most cited. “Visualizing and Understanding Convolutional Networks” (Zeiler and Fergus, 2014) presents one of the earliest approaches to understanding the inner workings of convolutional neural networks.[12] At the time, convolutional networks were beginning to take off as an architecture that seemed especially well suited to image classification; they have since become near ubiquitous in computer vision. Zeiler and Fergus introduced a novel way to “see” what is happening deep inside these networks, and their paper appears to have become the go-to reference on interpretability.

Growth over Time

As with robustness, interpretability research grew rapidly during the second half of the 2010s. The pattern of growth across countries and regions, however, is different: the United States leads, but in interpretability, the European Union is a close second. One possible reason for this may come from the General Data Protection Regulation (GDPR), the European privacy law enacted in

2016, which includes provisions referring to a user's right to receive an explanation of how their data has been processed. While the legal details of how and whether a "right to explanation" applies to machine learning models have been the subject of vigorous debate,[13] the prospect that it might become a legal obligation may have accelerated European research on how to understand and explain AI decisions.

Figure 2: Papers published in interpretability-related clusters over time



Source: Authors' analysis of the CSET Map of Science.
Note: Most individual countries' publication numbers are much lower than those of the United States and China, so we use country and regional groupings to make these charts easier to interpret: "CANZUK" refers to Canada, Australia, New Zealand, and the United Kingdom; "East Asian Democracies" groups Japan, South Korea, and Taiwan.

The number of Chinese publications, on the other hand, lags much further behind the United States in this category. It appears that Chinese researchers have not jumped into interpretability research with the same gusto as in robustness research, at least in the clusters we identified.

Reward Learning

Clusters Identified

We identified three clusters that appeared to contain substantial amounts of research on reward learning.

Table 3: Research clusters with a substantial amount of reward learning research ("reward learning–related clusters")

Original cluster	Rough description of papers in cluster	# of papers	Est. % relevant papers (+/- 10%)	Most similar cluster in new Map of Science
#5424	Work on robots learning from humans and collaborating with humans; the majority of the work is learning from demonstrations	2,252	68%	#1764
#24156	Research on inverse reinforcement learning, learning from human feedback, learning from demonstrations, and other human-robot interactive setups	1,352	81%	#20613
#78510	A range of different ways for humans to be involved in training robots, including via teaching and giving feedback	435	62%	#29472

Source: Authors' analysis of the CSET Map of Science.
Note: The first column gives the ID numbers for the clusters used in our analysis, which was based on the Map of Science before its major update in August 2021. Interested readers can use the cluster IDs in the final column to explore the most similar clusters in the Map of Science after this update.

As with interpretability, we needed to make multiple judgment calls in order to determine what does and does not count as "reward learning" research for our purposes. All three shortlisted clusters that we considered for reward learning clearly fell within the larger space of human-robot interaction, which makes sense given that reward learning research primarily involves reinforcement learning, and robotics is the area where reinforcement learning has been used most.

To identify relevant papers, we looked for research that aimed to have the AI model learn or infer a reward function—especially in interaction with a human—rather than having the reward specified directly. This included, for example, research on "inverse reinforcement learning," a well-established paradigm in which the AI's job is to observe some behavior, then infer what reward function might have been behind it.[14] We excluded more traditional human-robot interaction research, such as work on how humans respond to different types of robots or how best to design robot facial expressions. One large category of work that we were especially unsure about was "learning from demonstrations," also called "imitation learning," in which a robot observes some behavior (typically performed by a human) and then attempts to replicate it. We ultimately included this, but it was a borderline decision, since much of the research we observed on learning from demonstrations did not necessarily appear to be pushing towards the development of richer, more nuanced reward functions and therefore may not be of significant relevance to AI safety concerns.

All in all, our impression after identifying and exploring these research clusters is that reward learning may be a more awkward fit into the concept of "AI safety" than is true for robustness or interpretability. As laid out in the introduction, we included reward learning in this brief because of its relevance to specification problems—that is, the difficulty of specifying a goal for an AI system to maximize that lines up with the operator's intentions. While some of the reward learning work in the clusters we identified does seem relevant for this, and while we are aware of individual researchers working on reward learning who are motivated by specification problems, it appears that much of the

work in these clusters is motivated by other considerations, such as making robots more responsive to individual users.

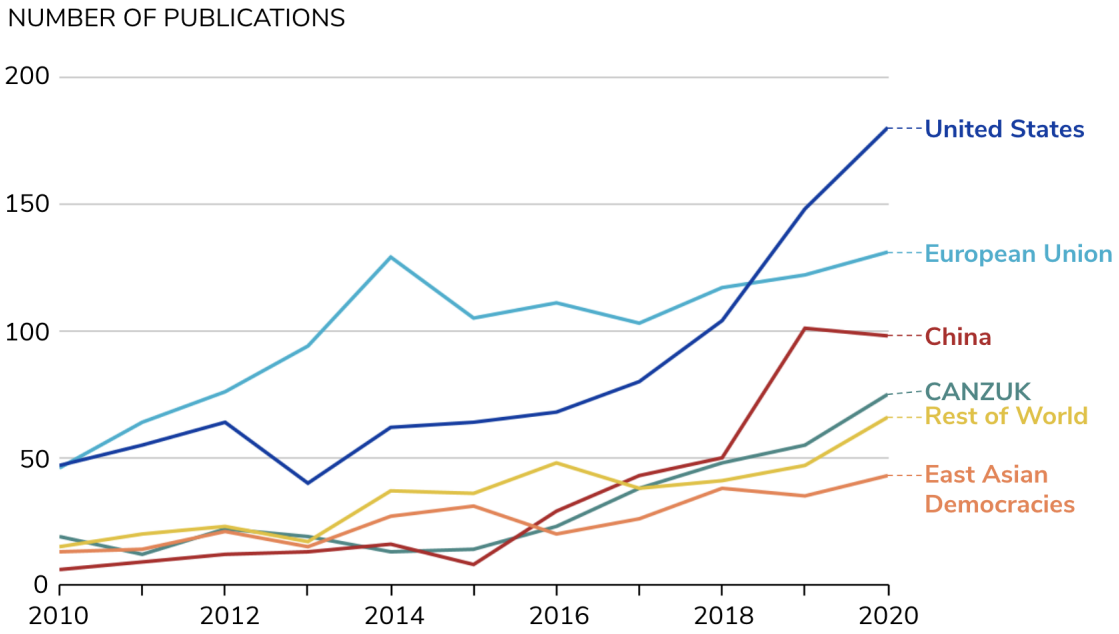
Foundational Papers

The three most highly cited works in the reward learning–related clusters were all somewhat older than the foundational papers described for previous categories. Most highly cited was a 2009 review paper by Brenna D. Argall et al., "A Survey of Robot Learning from Demonstration."[15] As the title indicates, this paper gives an overview of approaches to learning from demonstrations, covering a range of design choices that determine what kind of information a machine learning system takes in and how it processes that information. The next two most highly cited papers were both early publications on inverse reinforcement learning: "Apprenticeship Learning via Inverse Reinforcement Learning" (Abbeel and Ng, 2004) and "Algorithms for Inverse Reinforcement Learning" (Ng and Russell, 2000).[16] The relatively older age of these three papers is consistent with the steady rate of publications in the early 2010s shown in Figure 3.

Growth over Time

Similar to the previous two categories, the reward learning–related research clusters we identified saw substantial growth in the second half of the 2010s. Relative to robustness and interpretability, however, growth in this category was less explosive, with a larger number of publications being released in earlier years. This may reflect the fact that, whereas deep learning saw a sudden boom during the 2010s, interest in robotics has been somewhat steadier over time. As with interpretability, the European Union is second to the United States in this category, though in this case with relatively little growth in recent years. China, on the other hand, saw significant growth between 2015 and 2019 and is now third among our country and regional groupings.

Figure 3: Papers published in reward learning–related clusters over time



Source: Authors' analysis of the CSET Map of Science.
Note: Most individual countries' publication numbers are much lower than those of the United States and China, so we use country and regional groupings to make these charts easier to interpret: "CANZUK" refers to Canada, Australia, New Zealand, and the United Kingdom; "East Asian Democracies" groups Japan, South Korea, and Taiwan.

Conclusion

Taken together, we believe the research clusters described in this brief give a good initial view into research progress on topics related to AI safety. It appears that robustness research forms the clearest and most cohesive category, with a large fraction of the work in that category focused on adversarial examples. Interpretability and reward learning appear to be somewhat fuzzier categories, though the clusters we identified do contain a significant amount of research in those areas.

Based on the research clusters we identified, it appears that work in all these areas is growing at a significant pace worldwide. The United States appears to lead in every area, with China showing substantial growth in robustness research and the EU producing a large amount of interpretability work. Notably, despite the significant growth of the AI safety–related clusters we describe here, they still represent only a tiny fraction of total worldwide research on AI in general. We identified eight clusters containing a little over 15,000 papers; if we remove the "safety" component and consider all research clusters that contain more than 50 percent AI papers, we instead find nearly 2,000 clusters containing over 1.9 million papers. These numbers imply that safety research may make up less than 1 percent of AI research overall.

The clusters we identified show some promising progress on safety-related problems and may also point to some gaps. For instance, one could see the heavy emphasis on adversarial examples in robustness clusters in a negative light; perhaps the perceived prestige of that topic has led related problems to be relatively neglected.

Future work could further explore such hypotheses and refine our findings here, for instance by accounting for additional technical approaches within the categories we identified or adding whole new categories of AI safety work. We encourage interested readers to explore the interactive CSET Map of Science on the CSET website. We see this as an area of growing importance, so we would be particularly glad to hear from readers with expertise in AI safety with ideas about how to expand on this analysis.

Authors

Helen Toner is director of strategy at CSET, where Ashwin Acharya was a research analyst.

Acknowledgments

The authors are grateful to Ilya Rahkovsky and Autumn Toney for valuable data support, to Igor Mikolic-Torreira and Catherine Aiken for helpful suggestions and comments throughout, and to Larry Lewis and Rohin Shah for insightful comments on an earlier version. Melissa Deng and Alex Friedland provided editorial support. All remaining defects are the authors' responsibility alone.

Document Identifier: doi: 10.51593/20210026

Appendix A: Bonus Cluster on AI Risks and Futures

In addition to the research clusters described so far relating to robustness, interpretability, and reward learning, one additional cluster stood out during our exploratory research, so we wanted to include it here. This cluster, #54141, does not primarily contain technical research on a particular type of AI safety, and therefore did not meet the search criteria we used to identify relevant clusters.[17] However, it was apparent even at a glance that it contained some highly relevant works that focus on the idea of "AI safety" itself, or more generally on potential risks from AI and artificial general intelligence (AGI). In addition to these higher-level discussions of AI risks and related ideas for a broad audience, the cluster also includes a small amount of technical research on AI safety as well as less-relevant work, such as papers on the philosophy of intelligence, public perceptions of AI, or futurism more generally.

Table A1: Bonus cluster

Original cluster	Rough description of papers in cluster	# of papers	Est. % relevant papers (+/- 10%)	Most similar cluster in new Map of Science
#54141	A grab bag of research on different topics relating to AI futures, including more philosophically inclined work on AI safety and AI risk, AGI, superintelligence, the singularity, etc., as well as some technical research on AI safety	1,268	41%	#44272

Source: Authors' analysis of the CSET Map of Science.
Note: The first column gives the ID numbers for the clusters used in our analysis, which was based on the Map of Science before its major update in August 2021. Interested readers can use the cluster IDs in the final column to explore the most similar clusters in the Map of Science after this update.

Foundational Papers

The four most-cited publications in the cluster provide an informative sample of its contents. Most cited by far was Ray Kurzweil's 2005 book *The Singularity is Near*. Written for a popular audience, this book forecasted that computation would continue to progress on an exponential curve, leading to the "Singularity"—the moment when machine intelligence surpasses human intelligence—within a few decades.

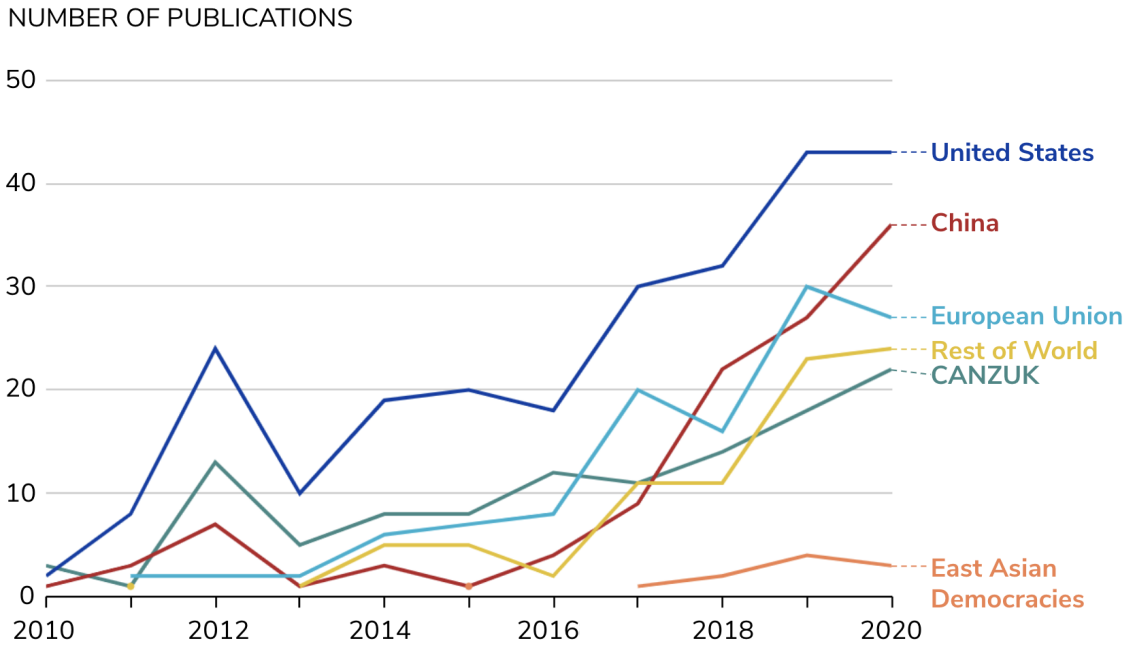
Kurzweil's book, which is viewed with skepticism by many in the AI community, stands in contrast to the next most-cited publication in the cluster, the paper "Concrete Problems in AI Safety" (Amodei et al. 2016).[18] Unlike *The Singularity is Near*, "Concrete Problems" was written by AI researchers for AI researchers, aiming to convert philosophical concerns about potential risks from AI into concrete, technical problems that researchers could sink their teeth into.

The next two most-cited publications are both back in the category of high-level, book-length discussions of AI: philosopher Nick Bostrom's *Superintelligence* (2014) and roboticist Hans Moravec's *Mind Children* (1988). Bostrom introduced the idea of a superintelligence—an AI vastly smarter than humans—to a wide audience, and discussed some of the potential risks its development might bring. *Mind Children* is best known for what came to be called "Moravec's paradox": in the author's original formulation, this was the idea that cognitive tasks we think of as difficult (such as abstract thought) seem to require few computational resources, while tasks we think of as simple (such as perception or locomotion) are highly computationally demanding. It seems likely that many of the publications that cited *Mind Children* did so in order to refer to this paradox, which became widely known in simplified form: in AI, "the hard problems are easy and the easy problems are hard."[19]

Growth over Time

This cluster has seen steady growth since 2010, perhaps reflecting an overall increase in interest in AI and related questions over that period.

Figure A1: Papers published in cluster #54141 over time



Source: Authors' analysis of the CSET Map of Science.
Note: Most individual countries' publication numbers are much lower than those of the United States and China, so we use country and regional groupings to make these charts easier to interpret: "CANZUK" refers to Canada, Australia, New Zealand, and the United Kingdom; "East Asian Democracies" groups Japan, South Korea, and Taiwan.

Appendix B: Methodology

We summarize the methods we used to identify the clusters covered in this brief under "Methodology," above. Here, we provide additional details on our methods and findings.

Keywords and Definitions

The notional definition and keyword lists we used for each category of AI safety research are as follows:

- Robustness: ensuring that AI systems stay within safe limits, regardless of the conditions encountered
    - "adversarial learning"
    - "safe exploration"
    - "machine learning security"
    - "adversarial example"
    - "safety margin"
    - "error correction"
    - "graceful degradation"
    - "failsafe"
    - "fail-safe"
    - "fail safe"
- Interpretability: understanding how AI systems work and how they may fail
    - "interpretab*"
    - "explainab*"
    - "transparency"

- Reward learning: methods to learn or infer a reward function in interaction with a human, rather than having the reward specified directly
    - “value learning”
    - “reward learning”
    - “inverse reinforcement learning”
    - “imitation learning”
    - “human feedback”
    - “scalable supervision”
    - “corrigibility”
    - “wireheading”
    - “specification gaming”
- General keywords:
    - “artificial general intelligence”
    - “superintelligence”
    - “AI alignment”

Note that * allows for any completion of a word, i.e., “interpretab*” would capture “interpretability,” “interpretable,” etc.

As an informal sensitivity check, we also ran some searches using Chinese translations of these keywords to see if any additional clusters would show up with a high proportion of our keywords in Chinese. We did not identify any such clusters, suggesting that Chinese work on these topics is either published in English, or—even if published in Chinese—cites a sufficient amount of English-language research to be included in the clusters we identified.

Shortlisted Clusters

As described above, we used keyword searches to create a list of 42 clusters that met two criteria: 5 percent or more of papers included a keyword from our list, and 50 percent or more of papers were AI papers.[20] To further investigate this list, we then manually screened the titles of 20–30 publications that the CSET Map of Science displays for each cluster, all from the last five years: the 10 most-cited papers, 10 "core papers" (those with most citation linkages to other papers in the cluster), and up to 10 review papers (where these existed). Based on these titles, we kept clusters that appeared to be focused on a safety-related topic and excluded those that appeared to contain research on other topics. This left us with a shortlist of 13 clusters, described in Table A2.

Table A2: Shortlisted clusters

Original cluster	Rough description of papers in cluster	# of papers	Est. % relevant papers (+/- 10%)	Most similar cluster in new Map of Science	Included in brief?
Robustness					
#3735	Almost exclusively research on creating and defending against adversarial examples	4,406	96%	#2381	Yes
#48621	Mix of adversarial examples, data poisoning, backdoor attacks; also includes some work on using AI for cybersecurity applications	967	82%	#26042	Yes
#83739	A range of approaches for testing and verifying the performance of ML systems, plus some work on adversarial examples	781	79%	#32715	Yes

Interpretability					
#9891	Work using a range of techniques to improve the interpretability of machine learning models, especially neural networks	3,521	86%	#5109	Yes
#17893	Extracting decision rules from neural networks	1,483	79%	#15771	Yes
#48185	Some interpretability work, also plenty on data visualization, interactive interfaces for machine learning, and exploratory data analysis	864	28%	#23422	No
#76510	Primarily about human-robot interactions, including how to design robot motions that allow humans to infer what the robot will do next	539	16%	#55552	No
#86537	Some interpretability papers, but the cluster did not appear to have a coherent theme that we could identify; it included papers on negotiation, English classes, after-action review, and other topics	485	15%	#71862	No
#112039	Primarily work on natural language processing; some work clearly focuses on performance (not interpretability), but the cluster also includes work on how languages are	548	16%	#1193	No

	represented in neural networks, with which it is hard to draw a definitive line between “relevant” and “not relevant”; “% relevant papers” is therefore especially tentative here				
#118744	Primarily work on “disentanglement,” i.e., learning statistical representations that map well onto real-world concepts; hard to categorize as relevant or not given that disentangled representations are generally more interpretable, but interpretability is often not the primary aim of the research	339	12%	#7095	No
Reward learning					
#5424	Work on robots learning from humans and collaborating with humans; the majority of the work is learning from demonstrations	2,252	68%	#1764	Yes
#24156	Research on inverse reinforcement learning, learning from human feedback, learning from demonstrations, and other human-robot interactive setups	1,352	56%	#20613	Yes

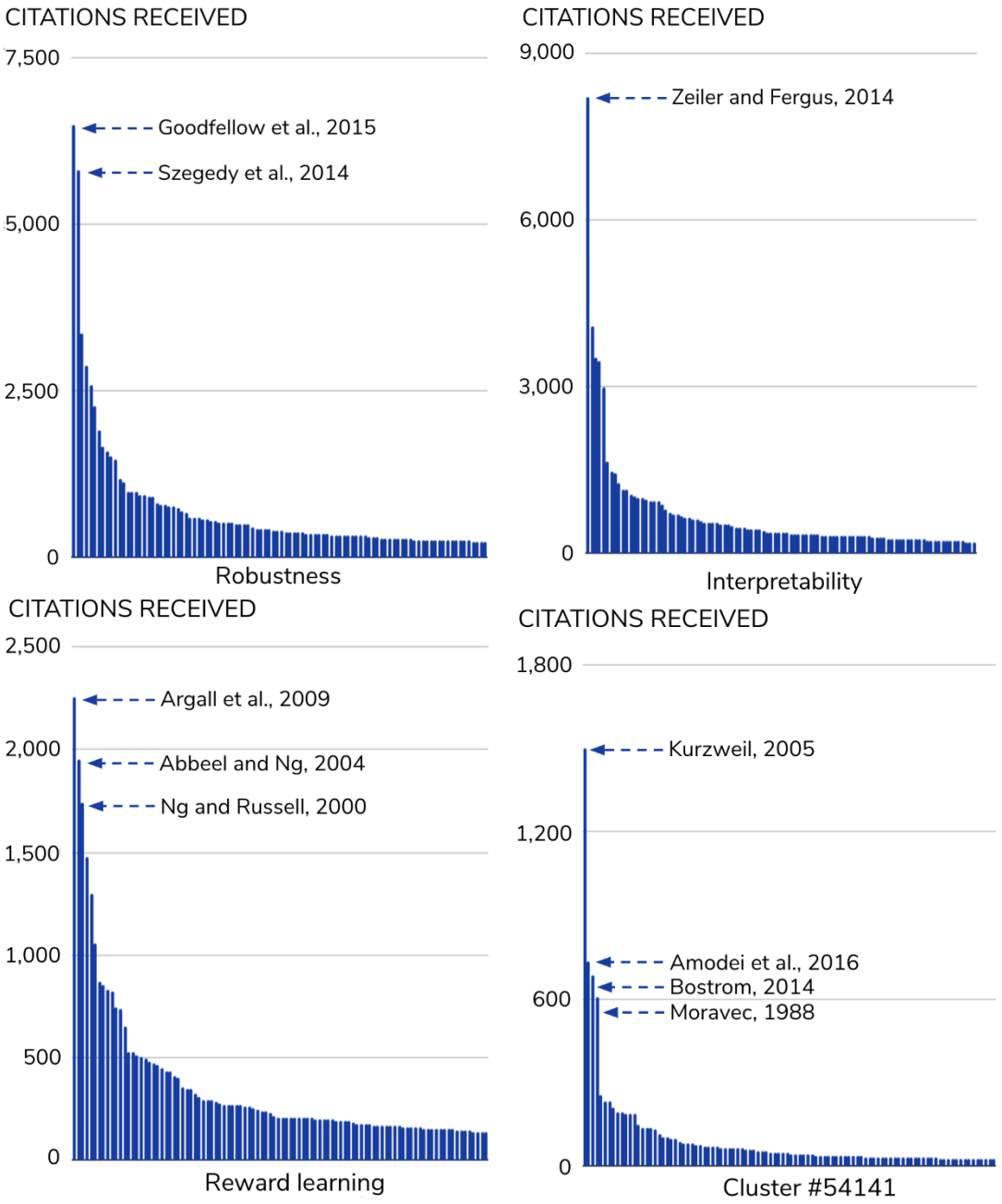
#78510	A range of different ways for humans to be involved in training robots, including via teaching and giving feedback	435	59%	#29472	Yes

Source: Authors' analysis of the CSET Map of Science.

Appendix C: Foundational Papers

For reference, we present here the distribution of the top 100 most-cited papers in each category, with the papers we highlighted in the main text indicated.

Figure A2: The citation distribution of the 100 most-cited papers in research clusters of each category



Source: Authors' analysis of the CSET Map of Science.

Endnotes

[1] CSET Map of Science is derived from CSET's research clusters and merged corpus of scholarly literature. CSET's merged corpus of scholarly literature includes Digital Science's Dimensions, Clarivate's Web of Science, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Papers With Code. All China National Knowledge Infrastructure content furnished for use in the United States by East View Information Services, Minneapolis, MN, USA. For more on the methodology used to create the research clusters, see Ilya Rahkovsky et al., "AI Research Funding Portfolios and Extreme Growth," Frontiers in Research Metrics and Analytics, April 6, 2021, https://www.frontiersin.org/articles/10.3389/frma.2021.630124/full.

[2] This breakdown is derived from the categories used in Pedro A. Ortega and Vishal Maini et al., "Building safe artificial intelligence: specification, robustness, and assurance," DeepMind Safety Research (*Medium*), September 27, 2018, https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1 and "Example Research Topics," in "The Open Phil AI Fellowship," Open Philanthropy, https://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence/the-open-phil-ai-fellowship.

[3] For an overview of methods to attack machine learning systems, see Andrew Lohn, "Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity" (Center for Security and Emerging Technology, December 2020), https://cset.georgetown.edu/publication/hacking-ai/.

[4] Interpretability is sometimes also referred to as "explainability." For a nice account of the complexity of what constitutes an "explanation" in the real world, see Finale Doshi-Velez and Mason A. Kortz, "Accountability of AI Under the Law: The Role of Explanation" (Berkman Klein Center Working Group on Explanation and the Law, 2017), https://dash.harvard.edu/handle/1/34372584.

[5] Tim G. J. Rudner and Helen Toner, "Key Concepts in AI Safety: Specification in Machine Learning" (Center for Security and Emerging Technology, December 2021), https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-specification-in-machine-learning/.

[6] Ortega and Maini et al., "Building safe artificial intelligence"; "Example Research Topics," Open Philanthropy; Tom Everitt, Gary Lea, and Marcus Hutter, "AGI Safety Literature Review," arXiv preprint arXiv:1805.01109v2 (2018), https://arxiv.org/abs/1805.01109.

[7] The methodology we use to estimate what proportion of papers in a cluster are AI papers is described in detail in Autumn Toney, "Creating a Map of Science

and Measuring the Role of AI in it" (Center for Security and Emerging Technology, June 2021), https://cset.georgetown.edu/publication/creating-a-map-of-science-and-measuring-the-role-of-ai-in-it/.

[8] This column gives an estimate of what proportion of papers in the cluster are related to the relevant category of AI safety (in this case, robustness). This estimate was created by randomly selecting a sample of papers from the cluster that had English-language titles and abstracts, then manually evaluating whether each of those papers appeared to be relevant to the category or not. The number of papers in the random sample was chosen to give an estimate of the 95% confidence interval +/- 10%.

[9] CSET's merged corpus of scholarly literature includes Digital Science's Dimensions, Clarivate's Web of Science, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Papers With Code. All China National Knowledge Infrastructure content furnished for use in the United States by East View Information Services, Minneapolis, MN, USA.

[10] Kevin Eykholt, Ivan Evtimov, and Earlence Fernandes, "Robust Physical-World Attacks on Deep Learning Visual Classification," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2018, https://www.researchgate.net/publication/329750836_Robust_Physical-World_Attacks_on_Deep_Learning_Visual_Classification; Yao Qin et al., "Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition," *Proceedings of the 36th International Conference on Machine Learning*, 2019, http://proceedings.mlr.press/v97/qin19a.html; Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, Chenliang Li, "Adversarial Attacks on Deep-learning Models in Natural Language Processing: A Survey," ACM Transactions on Intelligent Systems and Technology 11, no. 3 (June 2020): 1-24, https://dl.acm.org/doi/10.1145/3374217.

[11] Christian Szegedy et al., "Intriguing properties of neural networks," 2nd International Conference on Learning Representations," April 2014, https://nyuscholars.nyu.edu/en/publications/intriguing-properties-of-neural-networks; Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv preprint arXiv:1412.6572v3 (March 2015), https://www.semanticscholar.org/paper/Explaining-and-Harnessing-Adversarial-Examples-Goodfellow-Shlens/bee044c8e8903fb67523c1f8c105ab4718600cdb.

[12] Matthew D. Zeiler and Rob Fergus, "Visualizing and Understanding Convolutional Networks," *Computer Vision – ECCV 2014*, 2014, 818-833. https://link.springer.com/chapter/10.1007/978-3-319-10590-1_53

[13] Andrew Burt, "Is there a 'right to explanation' for machine learning in the GDPR?," *IAPP*, June 1, 2017, https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/.

[14] Technically, IRL can also involve inferring the reward function from a given policy, not only a given behavior trajectory. Andrew Y. Ng and Stuart J. Russell, "Algorithms for Inverse Reinforcement Learning," *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, June 2000, 663-670, https://dl.acm.org/doi/10.5555/645529.657801.

[15] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning, "A survey of robot learning from demonstration," Robotics and Autonomous Systems 57, no. 5 (May 2009): 469-483, https://www.sciencedirect.com/science/article/abs/pii/S0921889008001772.

[16] Pieter Abbeel and Andrew Y. Ng, "Apprenticeship learning via inverse reinforcement learning," ICML '04: Proceedings of the Twenty-First International Conference on Machine Learning, July 2004, https://dl.acm.org/doi/10.1145/1015330.1015430; Ng and Russell, "Algorithms for Inverse Reinforcement Learning."

[17] See Appendix B for more details on our methodology. This cluster did not meet the criterion that at least 50 percent of the papers in the cluster should be AI papers.

[18] Dario Amodei et al., "Concrete Problems in AI Safety," arXiv preprint arXiv:1606.06565 (June 2016). https://arxiv.org/abs/1606.06565.

[19] Steven Pinker, *The Language Instinct* (New York, NY: William Morrow and Company, 1994).

[20] The methodology we use to estimate what proportion of papers in a cluster are AI papers is described in detail in Toney, "Creating a Map of Science and Measuring the Role of AI in it."