The Artificial Intelligence Patent Dataset (AIPD) 2023 update

Nicholas A. Pairolero,[1,*] Alexander V. Giczy,[1,2] Gerard Torres,[1] Tisa Islam Erana,[3] Mark A. Finlayson,[1,3] and Andrew A. Toole[1]

Abstract

The 2023 update to the Artificial Intelligence Patent Dataset (AIPD) extends the original AIPD to all United States Patent and Trademark Office (USPTO) patent documents (i.e., patents and pre-grant publications, or PGPubs) published through 2023, while incorporating an improved patent landscaping methodology to identify AI within patents and PGPubs. This new approach substitutes BERT for Patents for the Word2Vec embeddings used previously, and uses active learning to incorporate additional training data closer to the "decision boundary" between AI and not AI to help improve predictions. We show that this new approach achieves substantially better performance than the original methodology on a set of patent documents where the two methods disagreed—on this set, the AIPD 2023 achieved precision of 68.18 percent and recall of 78.95 percent, while the original AIPD achieved 50 percent and 21.05 percent, respectively. To help researchers, practitioners, and policy-makers better understand the determinants and impacts of AI invention, we have made the AIPD 2023 publicly available on the USPTO's economic research web page.

Introduction

The Artificial Intelligence Patent Dataset (AIPD) was publicly released by the United States Patent and Trademark Office (USPTO) in 2021 (Giczy et al. 2022). Since its release, the AIPD has significantly contributed to the understanding of artificial intelligence (AI) invention by encouraging research into both its determinants and impacts (e.g., Toole et al. 2020; Chattergoon and Kerr 2022; Chowdhury et al. 2022; Gaske 2023; Gomes et al. 2023; Liu et al. 2023; Park 2024; Giczy et al. 2024; Gao et al. 2024; Rathi et al. 2024), and calling attention to the significant challenges associated with identifying AI within patent documents (Hotte et al. 2022; Grashof et al. 2023, Montobbio et al. 2023). Subsequent to the creation of the AIPD, researchers have developed several new methodologies for identifying technologies disclosed in patent documents (Krestel et al. 2021; Choi et al. 2022; Pujari et al. 2022; Yoo et al. 2023; Islam Erana

[1] United States Patent and Trademark Office, 600 Dulany St, Alexandria, VA 22314. The views expressed are those of the individual authors and do not necessarily reflect official positions of the Office of the Chief Economist or the U.S. Patent and Trademark Office.
[*] Corresponding author, Nicholas.Pairolero@uspto.gov
[2] Addx Corporation
[3] Florida International University, Knight Foundation School of Computing and Information Sciences, CASE Building Room 362, 11200 S.W. 8th Street, Miami, FL 33199

and Finlayson 2024; Pelaez et al. 2024), allowing us to improve our approach while extending the original dataset to include all patent documents published through 2023.

The 2023 update of the AIPD (hereinafter called the "AIPD 2023") identifies which of 15.4 million U.S. patent documents (patents and pre-grant publications, or PGPubs) published from 1976 through 2023 contain AI (separately identified for the eight AI component technologies from the AIPD, including machine learning, vision, natural language processing, speech, evolutionary computation, AI hardware, knowledge processing, and planning and control).[4] The update includes an additional 2.2 million patent documents published since January 2021 that were not included in the original 2021 release,[5] and has been publicly released on the USPTO's economic research webpage (https://www.uspto.gov/ip-policy/economic-research/research-datasets/artificial-intelligence-patent-dataset).

The AIPD 2023 was created from the original AIPD framework and incorporates several improvements from the recent patent landscaping literature. For example, we now incorporate BERT for Patents (Devlin et al. 2018; Srebrovic and Yonamine 2020) into our machine learning architecture (originally based on Abood and Feltenberger 2018 and extended in Giczy et al. 2022 and Islam Erana and Finlayson 2024). Additionally, we overcome a limitation of the Abood and Feltenberger (2018) "expansion method" used to create the training dataset for the original AIPD (Giczy et al. 2022) by including training observations closer to the "decision boundary" of AI and not AI, thereby enabling the model to learn from patent documents that are more difficult to classify.[6] These observations were manually labeled and selected via an active learning model that sampled patent documents from close to the 50 percent prediction threshold (i.e., from the set of observations where the model was most uncertain). Islam Erana and Finlayson (2024) shows the benefits of adopting these new approaches within the original AIPD framework.

Given the large number of differences between the AIPD 2023 and the original approach, we carefully analyzed the set of "disagreements" between the models. Overall, the number of

[4] As described in Giczy et al. 2022 and Toole et al. 2020, our definition of AI is broad and encompasses earlier and more general technologies beyond the deep learning and large language models that are currently most associated with AI.

[5] The AIPD 2023 dataset does not include 14,140 patent documents that were in the previous AIPD 2021 release. These documents were not included for several reasons, including that some were granted patents and PGPub that have since been withdrawn. See Appendix C for information regarding withdrawn patents and PGPubs.

[6] Given an input consisting of likely true positive observations, the "expansion method" finds negative observations by randomly sampling from those that are far away from the true positives (i.e., to increase the likelihood that the observations are actually true negatives), leaving little interior training data from which the model can learn the true decision boundary.

disagreements varied across the AI component technologies, ranging from 123,810 patent documents in speech to 264,618 in machine learning and 809,066 in AI hardware. Notably, the disagreements far outnumbered those documents where both models agreed that the inventions contained AI. For example, 70.21 percent of the documents where at least one of the models predicted machine learning were disagreements. To better understand which model was "right" more often, we compared predictions for 1,000 patents documents published in 2019 and manually reviewed and labeled 229 documents that differed in at least one AI technology component. In all but one of the eight components the AIPD 2023 achieved higher precision and recall on the manually reviewed documents, leading to greater F1 scores. When considered at the aggregate AI level (i.e., disagreements in at least one AI technology component), the AIPD 2023 achieved precision of 68.18 percent and recall of 78.95 percent, while the original AIPD achieved 50 percent and 21.05 percent, respectively.

Even though the AIPD 2023 component technology models have better F1 scores than the original approach, the AIPD 2023 produces a substantially greater number of AI predictions each year, which is consistent with training and evaluation metrics for each component technology that favor higher recall at the expense of precision. For researchers seeking greater continuity with the original AIPD, or those that prefer greater precision at the expense of recall (Grashof et al. 2023), we show that increasing the AIPD 2023 prediction threshold for determining AI can produce an AI prediction volume that closely matches the original approach.[7] Further, we identify a threshold estimate in the AIPD 2023 for balancing precision and recall, which when used produces a more accurate estimate of the volume of AI. The AIPD 2023 release contains the raw model prediction scores, allowing researchers to choose the prediction threshold and thus the level of precision and recall that is most appropriate for their application. In addition, the dataset includes binary variables for several thresholds, including 50 percent, the threshold for balancing precision and recall, and the estimate that best reproduces the volume of AI from the original AIPD's 50 percent threshold.

As with the original dataset, our testing revealed that the AIPD 2023 is better for certain AI technology components than others. For example, the new predictions for evolutionary computation are substantially worse than those for the other AI component technologies (as revealed by model training metrics), a feature of the dataset that has not changed since the original AIPD. There are likely too few patent documents containing evolutionary computation in our training dataset to produce reliable predictions. Additionally, the AIPD 2023 model for AI

[7] Given a patent document, each AIPD component technology model produces a prediction, which can be interpreted as a probability between 0 and 1 (with 1 indicating AI and 0 indicating not AI). The prediction threshold is the probability for which all predictions above it will be labeled AI and all below will be labeled not AI.

hardware (i.e., hardware that is specifically designed to improve AI computation) achieved both worse precision and recall than the original AIPD in our manual evaluation. Although there are many potential reasons for this, one possibility is that our new training dataset contains annotations from several different reviewers, and labeling patent documents in AI hardware is difficult. AI software inventions are often described as being embedded in a physical hardware system, and general-purpose hardware improvements may improve AI computation as well as computation more generally. These nuances could make it more difficult for humans to consistently label AI hardware, and therefore reduce the overall quality of the predictions.

The article proceeds as follows: first, we provide a brief overview of the model used to produce the original AIPD, and describe the literature that uses this dataset or relies on the article that describes it, Giczy et al. (2022). Second, we identify the differences between the approach used for the AIPD 2023 relative to the original AIPD model, followed by a description of the evaluation sample and the performance results obtained from it. Next, we provide several extensions, which include an analysis of the impact of adjusting the prediction threshold for AI, and more information on the set of disagreements between the new and original approaches. We conclude by describing several practical challenges associated with implementing a machine learning approach such as the one we used, as well as highlighting potentially promising areas for future research in this area. More information on the dataset, including how it may be used with publicly available patent data from PatentsView,[8] is available in the Appendices.

Background

Original AIPD methodology

The original AIPD was created using a multi-step deep learning approach based on the automated patent landscaping methodology of Abood and Feltenberger (2018). In the first step, patent classification/keyword queries were created to identify patent documents within each of the eight AI component technologies. These documents formed the positive example "seed sets." Next, the expansion method of Abood and Feltenberger (2018) was used to identify negative example "anti-seed" documents. This expansion approach used technology classifications, citations, and patent family relationships to find documents that were "far enough away" from the seed documents to be likely true negatives. The seed and anti-seed sets formed the training datasets for each AI component technology model.

The deep learning architecture was based on the best performing model of Abood and Feltenberger (2018), consisting of long short-term memory (LSTM) neural networks for patent

[8] PatentsView is a publicly accessible data visualization platform supported by the USPTO's Office of the Chief Economist that contains several research datasets on U.S. patents and PGPubs (see https://patentsview.org/).

application claims and abstracts (using Word2Vec for text embedding) and a dense neural network to process patent citations (which were one-hot encoded as inputs). The outputs of these layers were then combined using several additional neural network layers. Giczy et al. (2022) showed that this approach achieved superior performance relative to alternatives in the literature on a holdout set of 368 patent documents that were manually annotated by USPTO patent examiners. More information on the methodology and evaluation of the original AIPD is available in Giczy et al. (2022).

Use of the AIPD

Since the release of the AIPD in 2021, the dataset has been downloaded 5,226 times, and the article describing the AIPD (Giczy et al. 2022) has been referenced over 50 times by a variety of studies in the economics, management, computer science, and legal literatures.[9] Some of these studies have used the AIPD directly, while others have used information in Giczy et al. (2022) to inform their research methods or as a resource for supporting material. Table A1 in Appendix A shows that these uses are the primary ways the AIPD has been used, with the addition of several articles that benchmark the AIPD against other AI classification methods to assess how input datasets on AI affect applied results (e.g., like the degree to which AI is a "general purpose technology" or GPT) (Hötte et al. 2022; 2023; 2024).

Beyond scientific impact, the AIPD has been used broadly to stimulate policy discussions between the U.S. Federal Government and various stakeholders, including at several events associated with the USPTO's AI and Emerging Technology Partnership,[10] as well as the Office of the Director of National Intelligence Science and Technology Partnership.[11] Additionally, the AIPD was used in the USPTO's 2022 report to Congress on patent eligible subject matter in the United States (Vidal 2022) to document how recent changes in patent law might affect upstream AI investments that support invention, as well as downstream innovation and commercialization opportunities in AI (Toole et al. 2020; Frumkin et al. 2024). The release of the AIPD 2023 should continue to support this research and policy activity by improving the quality of the underlying dataset and extending it through the end of 2023.

AIPD 2023 methodology

To create the AIPD 2023, we used the same machine learning approach as the original release but incorporated several improvements. First, the machine learning models now use BERT for

[9] As of May, 2024.
[10] For more information, see https://www.uspto.gov/initiatives/artificial-intelligence/ai-and-emerging-technology-partnership-engagement-and-events.
[11] For more information, see https://www.dni.gov/index.php/who-we-are/organizations/policy-capabilities/in-step-the-intelligence-science-technology-partnership.

Patents (Srebrovic and Yonamine 2020) rather than Word2Vec text embeddings. In a 2018 article, researchers at Google showed that BERT out-performed existing approaches, including Word2Vec, on several natural language processing benchmark tasks (Devlin et al. 2018). Within the patent landscaping context, Islam Erana and Finlayson (2024) shows that BERT for Patents achieves superior performance over Word2vec by a significant margin when incorporated into the machine learning architecture used to produce the AIPD (Abood and Feltenberger 2018; Giczy et al. 2022).[12,13]

The second major improvement over the original AIPD is updated training data. We used the original AIPD training data as a base, but extended it by: (1) adding newly labeled data closer to the "decision boundary" between AI and not-AI, (2) adding patent documents that were manually labeled by USPTO patent examiners when evaluating the original AIPD, and (3) adding AI patent documents published after 2019. The "decision boundary" documents were selected using active learning via a support vector machine (SVM) supervised machine learning model to identify and annotate documents that were close to the 50% prediction threshold between AI and not-AI (i.e., those documents for which the active learning model was the most uncertain) (Islam Erana and Finlayson 2024). These documents were from years 1976-2018, with 90% of the documents being from 2018. Graduate students in AI at Florida International University (FIU) annotated this data set, resulting in 1,147 documents across the eight AI component technologies.[14] The number of patent documents used for training from this source is shown in Table 1 in the "Decision Boundary" columns.

We also included in the training data the 800 patent documents that were previously annotated by USPTO examiners during the original AIPD evaluation. These documents were randomly sampled from the AIPD: 200 from the original seed training set, 200 from the original anti-seed training set, and 400 from all patent documents not in the seed or anti-seed sets. USPTO patent examiners specialized in AI labeled which of these 800 documents contained each

[12] BERT has also been shown to improve model performance across a variety of other patent related tasks, including prior art search (Vowinckel and Hahnke 2023; Chikkamath et al. 2024) and citation prediction (Ghosh et al. 2024).

[13] Additionally, we removed the citation part of the deep learning model architecture: the one-hot encoding was previously set to a maximum of 50,000 citations, and this number is too small to be of impact. Islam Erana et al. (2023) also excluded citations in its comparisons. It did, however incorporate Cooperative Patent Classification (CPC) codes of cited patent documents, a feature which we did not include in our models.

[14] Using the training dataset from the original AIPD and a small set of positive and negative examples labeled by FIU researchers to initiate the active learning model, the SVM was retrained every 10 new annotations selected near the 50 percent prediction threshold (using the uncertainty sampling method of Lewis and Gale 1994) to continually improve its understanding of the decision boundary. More information on this procedure is available in Lewis and Gale (1994) and Islam Erana and Finlayson (2024).

of the eight AI component technologies (more information on this process is available in Giczy et al. 2022). For training the updated AIPD models we selected only those patent documents where two annotators agreed on whether the document was AI in the component technology or not, i.e., we did not include those patent documents that required adjudication by a third examiner. The number of training documents from this source is shown in the "Examiner Annotated" columns of Table 1.

While incorporating the decision boundary annotations into the training data improves model performance (Islam Erana and Finlayson 2024), the most recent document in this training set was published in 2018. To capture the new ways AI has been used in invention since then, we added additional positive observations published from 2019 through 2023 to the seed set. These documents were obtained from search queries updated from the ones used for the original AIPD.[15] The queries were designed to be narrow, i.e., with very high precision. Moreover, to be conservative and not to overwhelm the previous training data, each seed set was increased by only 10 percent.[16] Table B1 in Appendix B provides the queries used to add these additional documents to the seed sets, and Table B2 shows how many were added using this approach.

Table 1: Number of documents and sample weights for each source of training data

AI component technology	Metric	Seed/Anti-seed		Decision Boundary		Examiner Annotated	
		Seed	Anti-seed	Positive	Negative	Positive	Negative
Machine learning	Number	1045	14957	31	1116	103	598
	Sample weight	14.3	1.0	482.5	13.4	145.2	25.0
Evolutionary computation	Number	101	14964	35	1112	2	797
	Sample weight	148.2	1.0	427.5	13.5	700.0	18.8
Natural language processing	Number	1182	14956	19	1128	54	709
	Sample weight	12.7	1.0	787.2	13.3	277.0	21.1
Vision	Number	879	14958	59	1088	24	751
	Sample weight	17.0	1.0	253.5	13.7	623.2	19.9
Speech	Number	828	14964	19	1128	24	762
	Sample weight	18.1	1.0	787.6	13.3	623.5	19.6
Knowledge processing	Number	725	14966	76	1071	60	626
	Sample weight	20.6	1.0	196.9	14.0	249.4	23.9

[15] The original queries used to define the seed sets in the AIPD were not directly re-useable due to substantial changes in the Cooperative Patent Classification (CPC) system for classifying AI inventions. The updated queries followed the same approach however, identifying likely true positives using several classification systems, including CPC, the United States Patent Classification (USPC), the International Patent Classification (IPC), and Derwent World Patent Index classification system.

[16] Due to the small number of seed documents in evolutionary computation, all 2019-2023 documents from the updated query were added.

AI component technology	Metric	Seed/Anti-seed		Decision Boundary		Examiner Annotated	
		Seed	Anti-seed	Positive	Negative	Positive	Negative
Planning and control	Number	1587	14960	287	860	50	551
	Sample weight	9.4	1.0	52.1	17.4	299.2	27.2
AI hardware	Number	2885	14955	49	1098	21	756
	Sample weight	5.2	1.0	305.2	13.6	712.1	19.8

Notes: A single document may be classified in more than one AI component technology. Seed documents include those added from 2019-2023 (see Appendix B). We edited the raw training data to exclude overlapping documents such that, for the same AI component technology, the document remained in only one set, with an order of precedence of: examiner annotated, decision boundary, and seed/anti-seed. Additionally, we removed documents without both abstract and claims text following text pre-processing.

The training data summarized in Table 1 is imbalanced across several dimensions. First, there were differing numbers of positive and negative observations in each AI component technology. Second, the decision boundary documents and examiner annotations were far outnumbered by the seed/anti-seed, even though the former two may contain more information about how to classify AI. We accounted for these imbalances by weighting the observations during training, as specified in Table 1, so that each group of documents (regardless of the number of documents in them) received approximately equal weight.[17]

As with the original AIPD, we trained one model for each AI component technology.[18] To estimate performance, each model was trained five times using an 80/20 train/test split.[19] We averaged the resulting "test" performance metrics over the five runs by training epoch (i.e., the number of complete passes through the data during training) and used this information to determine the optimal number of training epochs to use for the final models (using all the training data).[20] Table 2 shows both validation (Panel a) and final model (Panel b) metrics (accuracy, precision, recall, and the F1 measure), as well as the number of epochs used to train the final models ("Epoch" column in Panel b). As expected, the validation F1 scores (Panel a) are usually lower than the final model scores (Panel b) but the differences are not generally

[17] The weights were set such that, for each AI component, the number of documents times the weight approximately equaled the number of anti-seed documents, i.e., the largest number among all the training data groups. The weight was capped at about 700 to reduce unnecessary influence from any set having a very small number of documents.

[18] Additional methodological details are provided in Appendix C.

[19] In TensorFlow and Keras, "test" is referred to as "validation," i.e., the subset of data withheld during model training and used to evaluate model performance after each training epoch.

[20] We used stratified splits for the 80/20 training runs (where the strata were seed, anti-seed, positive decision boundary, negative decision boundary, positive examiner annotated, and negative examiner annotated) and trained the models for a maximum of 40 training epochs for each run. We selected the number of epochs to use based on how the average F1 score (over 5 runs) changed, picking the number of epochs that approximately maximized F1 (so as to avoid overfitting). For the final models we used all the training data (i.e., no 80/20 split) with the selected number of epochs from the previous step.

substantial. In most component technologies, such as machine learning and natural language processing, final model F1 scores are within two standard deviations of the validation metrics, but in some others, such as planning and control, the differences are larger.

Table 2: Training metrics for each AI component technology model

(a) Validation metrics from 5x 20/80 split training for final number of training epochs

AI component	Validation accuracy		Validation precision		Validation recall		Validation F1	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
Machine learning	0.984	0.005	0.831	0.049	0.935	0.036	0.865	0.020
Evolutionary computation	0.986	0.006	0.208	0.033	0.274	0.021	0.225	0.022
Natural language processing	0.993	0.003	0.906	0.028	0.965	0.019	0.929	0.019
Vision	0.977	0.009	0.725	0.085	0.888	0.028	0.776	0.054
Speech	0.987	0.007	0.796	0.100	0.942	0.023	0.845	0.060
Knowledge processing	0.958	0.015	0.568	0.131	0.843	0.021	0.648	0.080
Planning and control	0.927	0.014	0.628	0.060	0.802	0.037	0.686	0.027
AI hardware	0.943	0.004	0.786	0.025	0.856	0.032	0.811	0.011

(b) Final model training metrics

AI component	Epochs	Accuracy	Precision	Recall	F1
Machine learning	29	0.987	0.831	0.973	0.884
Evolutionary computation	27	0.976	0.267	0.406	0.306
Natural language processing	30	0.991	0.877	0.977	0.912
Vision	32	0.992	0.870	0.976	0.908
Speech	31	0.998	0.940	0.963	0.947
Knowledge processing	28	0.975	0.668	0.945	0.755
Planning and control	25	0.968	0.848	0.954	0.890
AI hardware	30	0.962	0.741	0.986	0.832

Notes: In sub-table (a), an 80/20 train/test split was used five times, and validation metrics were averaged across all five runs for the number of training epochs used in each of the final models. In sub-table (b), all training data was used to train a final model without a train/test split; hence, the table shows only the training metrics. All metrics are based on a 50% threshold between AI (positive result) and not AI (negative result) in each AI component.

As seen in Panel b of Table 2, the final training F1 scores ranged from a high of 0.947 for speech, to a low of 0.306 for evolutionary computation. Similar to the original AIPD, evolutionary computation continues to be a challenging component technology to identify in patent data. Giczy et al. (2022) suggests this may be the result of too few positive observations in the training data, a characteristic that has not changed since the original analysis. A final observation is that precision is lower than recall in each component technology, suggesting that the models may favor returning relevant AI documents at the expense of higher false positive rates when using a

prediction threshold of 50 percent. We provide more discussion on the tradeoff between precision and recall in the "Extensions and discussion" section below.

Evaluation

Given the large number of methodological differences between the original AIPD and the 2023 update, we conduct a series of analyses to identify how these changes affect the predictions. Our first analysis focuses on the "disagreements" between the two approaches, or the set of documents predicted as AI by either the AIPD 2023 or the original AIPD but not both. Table 3 shows for each AI component the number of disagreements of two types: AI predicted in the 2023 AIPD but not the original AIPD, and AI predicted in the original AIPD but not the 2023 update. In addition, the table includes the total number of disagreements, the total number of AI predictions (either AI in the AIPD 2023 or original AIPD, or both), and the percentage of disagreements relative to the total number of AI predictions in each component. Notably, the percentage of disagreements is substantial, nearly two thirds or higher in each component technology, ranging from a low of 62.59 percent in planning and control, to a high of 95.35 percent in evolutionary computation. In machine learning, 70.21 percent of the positive predictions from both models are disagreements.

Table 3: Summary statistics on the "disagreements" between the AIPD 2023 and the original AIPD

	AI in AIPD 2023 (86%) but not original AIPD (35%)	Not AI in AIPD 2023 (86%) but AI in original AIPD (35%)	Total disagreements	Total AI predictions	Percentage of disagreements out of predictions
Machine learning	181,134	83,484	264,618	376,870	70.21%
Evolutionary computation	156,115	39,413	195,528	205,069	95.35%
Natural language processing	236,017	35,715	271,732	393,485	69.06%
Vision	409,096	146,565	555,661	806,991	68.86%
Speech	86,743	37,067	123,810	178,731	69.27%
Knowledge processing	248,410	494,460	742,870	1,111,018	66.86%
Planning and control	422,654	425,267	847,921	1,354,646	62.59%
AI hardware	622,026	187,040	809,066	1,107,293	73.07%
Any AI	1,113,051	332,239	1,445,290	2,628,504	54.99%

Note: Includes all patent documents published between 1976 and 2020 and having predictions from both the updated AIPD 2023 and the original AIPD. Total disagreements are when one model (AIPD 2023 or original AIPD) predicts AI and the other does not. Total AI predictions is either model predicts AI. The difference between the total

number of AI predictions and the total disagreements in each component technology is the number of agreements (i.e., both agree AI or not AI in that component). The percentage of disagreements is relative to the total number of AI predictions in that component.

Manual evaluation

To better understand which model is more likely to be "right" on the set of disagreements, we annotated 229 documents published in 2019 from the set of documents where the two approaches disagreed.[21] We selected these documents for several reasons: first, to evaluate performance on recent data; second, to assess a period of time that experienced a rapid increase in positive AI predictions (see Figure 1 below); and perhaps the most important, to examine an area where we might expect to see an improvement in performance (since the decision boundary observations were primarily chosen from those published in 2018). However, for this last reason, any improvement in performance should be considered an upper bound, rather than a population level difference.

We split the documents among three annotators who each labeled the documents for the AI component technology source of disagreement (i.e., those that disagreed for machine learning, natural language processing, etc.), with approximately 80 total annotations for each annotator, about 10 from each AI component (229 annotations total, with almost 30 total from each AI component).[22] The objective of this analysis was to assess which of the two approaches had better performance on these "disagreements."

Table 4 shows several performance metrics, including precision, recall, and the F1 measure (wherein predictions were based on a 50% threshold), from the viewpoint of each model—the 2023 update in the top panel and the original AIPD in the bottom panel. From the perspective of the AIPD 2023, precision is the share of documents labeled as AI by the 2023 update in which this model was correct. Recall is the share of true AI documents, as determined by the annotators, in which the 2023 update was correct. In contrast, the bottom panel of Table 4 shows these metrics from the perspective of the original AIPD.[23] For every AI component technology but one (AI hardware), including "any AI" (i.e., whether the patent document is

[21] Since the 15.4M patent documents in our analysis were sorted by publication data and divided into 1,000 document subsets, we selected one of the subsets that were published in 2019. After running predictions using the new models and consolidating across all eight AI components, we compared those predictions to the ones from the original AIPD models; 229 documents had different predications using a 50% threshold for both models: 161 where the updated model predicted AI in any of the components but the original model didn't, and 36 vice versa.

[22] We attempted to label 30 disagreements total in each AI component technology, but two components in our sample—machine learning and speech—did not have 30 disagreements (at 28 and 21 disagreements, respectively).

[23] From the perspective of the original AIPD, precision is the share of documents labeled AI by the *original AIPD* in which this model was correct. Recall is the share of true AI documents, as determined by the annotators, in which the *original AIPD* was correct.

predicted as AI in at least one AI component technology), the 2023 update has higher precision and recall, which results in higher F1 measures. For example, the AIPD 2023 update achieved 100 percent precision and 65 percent recall in machine learning, while the original model achieved 81.82 percent precision, and only 34.62 percent recall. While the number of annotations in each component technology was not large, when taken together these results suggest that the AIPD 2023 provides better predictive performance than the original model.

Table 4: Performance statistics on a sample of 80 manually-reviewed documents from 229 "disagreements" in a set of 1000 patent documents published in 2019

(a) From the perspective of the AIPD 2023

AI Component	True positive	True negative	False negative	False positive	Total	Precision	Recall	F1
ML	17	2	9	0	28	1.0000	0.6538	0.7907
NLP	17	1	1	11	30	0.6071	0.9444	0.7391
Vision	21	1	2	6	30	0.7778	0.9130	0.8400
Speech	14	0	0	7	21	0.6667	1.0000	0.8000
KR	14	2	12	2	30	0.8750	0.5385	0.6667
Planning	15	2	10	3	30	0.8333	0.6000	0.6977
Hardware	5	5	7	13	30	0.2778	0.4167	0.3333
Any AI	15	4	4	7	30	0.6818	0.7895	0.7317

(b) From the perspective of the original AIPD

AI Component	True positive	True negative	False negative	False positive	Total	Precision	Recall	F1
ML	9	0	17	2	28	0.8182	0.3462	0.4865
NLP	1	11	17	1	30	0.5000	0.0556	0.1000
Vision	2	6	21	1	30	0.6667	0.0870	0.1538
Speech	0	7	14	0	21	-	0.0000	0.0000
KR	12	2	14	2	30	0.8571	0.4615	0.6000
Planning	10	3	15	2	30	0.8333	0.4000	0.5405
Hardware	7	13	5	5	30	0.5833	0.5833	0.5833
Any AI	4	7	15	4	30	0.5000	0.2105	0.2963

Notes: Sample consists of patent documents published in 2019 where the original AIPD and 2023 update disagree across at least one of the eight AI component technologies. Therefore, the estimates for precision, recall and F1 should not be considered population estimates. True and false positives and negatives are based on the perspective from the model noted above each sub table. Each document was reviewed by a single reviewer. Precision for speech in Panel b is not defined since there are no true positives or false positives.

The annotation analysis does reveal a weakness in the AIPD 2023, however—the predictions for AI hardware were generally worse than the original model, with an F1 score of only 0.333 on the sample of disagreements. Although it is difficult to understand precisely why, one possibility is that AI inventions are often described in patent documents as being executed

on or embedded within hardware. This description makes it difficult to differentiate inventions that are directed toward hardware specifically designed to improve AI systems from general AI software which also describes how the software might be implemented on computer hardware. Moreover, the distinction between hardware specifically designed to improve AI systems and hardware that can be used to more generally improve computation may be hard to distinguish. For example, quantum computers may enable faster AI training as well as improved execution of other algorithms (e.g., within cryptography), the latter not fitting into our definition of AI hardware (see Giczy et al. 2022). The new training dataset used in the AIPD 2023 included data annotated by many different reviewers, including FIU AI graduate students (for the decision boundary training set) and USPTO patent examiners (for the examiner annotated training set), potentially bringing these definitional challenges to the forefront of the analysis for more challenging components such as AI hardware.

As a final note, it is important to remember that the evaluation of the original AIPD revealed that annotating AI documents is challenging, even for human experts. In that analysis, USPTO examiners achieved 0.348 precision and 0.816 recall, resulting in an F1 score of 0.488 on a random sample of patent documents selected from outside the training set (see Giczy et al. 2022). This issue has not been resolved with the 2023 update—we used the same categorical-based definition of AI as before, as well as the same definitions for each AI component technology. Disagreements between annotators in a second manual review of 300 randomly sampled documents from some of the more difficult cases in the AIPD 2023 (i.e., those that were labeled AI in the update but not AI in the original AIPD) revealed one potential reason for this disagreement—many USPTO patent applications describe the transmission and manipulation of data through programmable logic.[24] It is challenging to identify when these processes rise to the level of AI, especially in broader components such as planning and control when the data is used to form a plan and control a system, from more basic logical processes, e.g., receiving an input signal from a device and manipulating the signal to produce a desired result.

[24] In this second manual review exercise, each of three annotators were given 100 documents randomly sampled from the documents predicted to be AI in the AIPD 2023 and not AI in the original AIPD, and each document was reviewed by only one annotator. While this analysis cannot reveal anything about recall from the perspective of the AIPD 2023 (since all selected documents were predicted as AI), it can determine precision (as the share of documents accurately predicted to be AI). Reflecting the challenges associated with identifying AI within this set of potentially more challenging documents, precision varied widely across the three reviewers, from a high of 59 percent to a low of 20 percent. Overall precision was 38.67 percent, which is very similar to the overall precision determined on the evaluation set of non-training documents in the original AIPD (i.e., 40.54 percent). As a final note, precision is different on this set than the first manual review sample drawn from 2019, as described above, because (1) the 300 documents did not include the other set of disagreements (i.e., those predicted to be AI in the original AIPD and not AI in the AIPD 2023), and (2) the annotators were not the same.

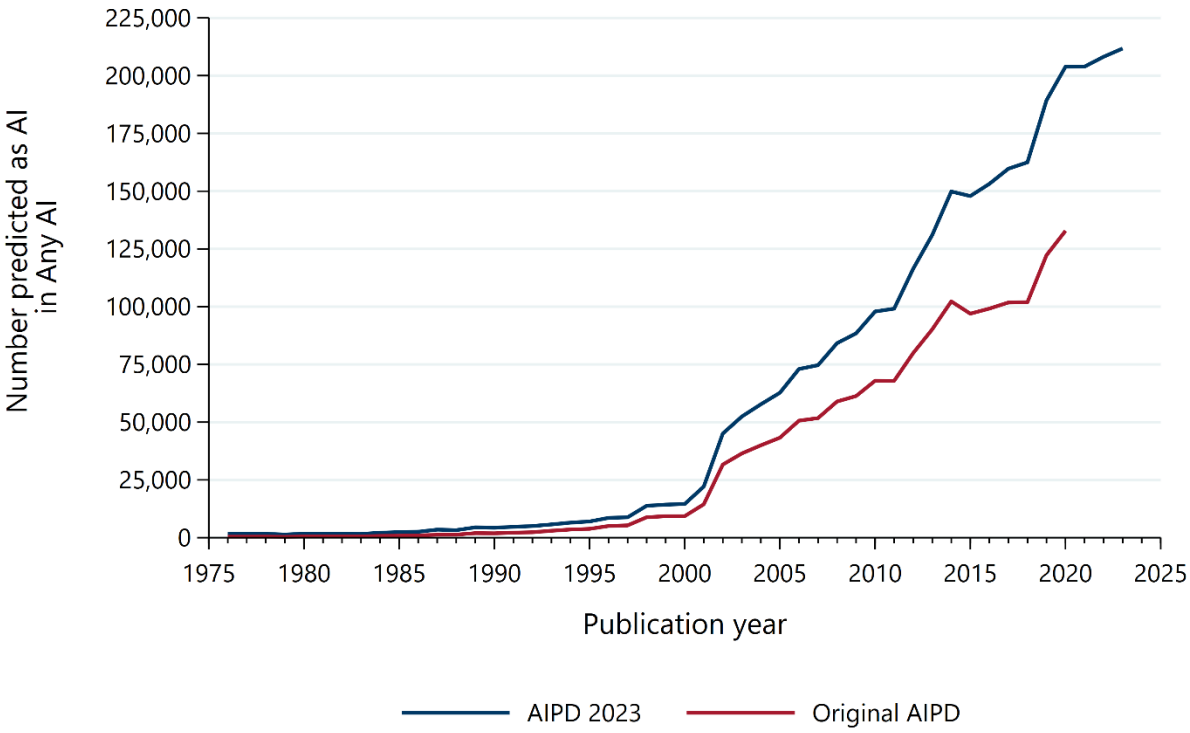
Such definitional aspects of forming a patent landscape are under-researched but are potentially very important for improving model performance. As previously discussed, we used active learning to identify training data near the decision boundary between AI and not AI for the different AI component technologies. The ultimate success of active learning depends on the ability of human annotators to consistently label documents near the boundary. The difficulty of human experts to label these cases consistently would place an upper bound on the efficacy of this approach.

Extensions and discussion

Adjusting the prediction threshold to better identify the volume of AI

Figure 1 shows the number of patent documents published in each year from 1976 to 2021 that were predicted to be AI using the 50 percent prediction threshold in the original AIPD and from 1976 to 2023 in the 2023 update (also using a 50 percent prediction threshold). Most noticeably, the number of documents predicted to be AI in the AIPD 2023 is substantially higher each year: about 50 percent higher relative to the original AIPD. The models produce similar trends however, with the exception of between 2015 and 2018, where the original AIPD is relatively flat while the AIPD 2023 increases slightly. Exploring the predictions by AI component reveals that the new models consistently predicted more AI than the original models in each component technology, except for knowledge processing where the new model predicted less AI, and planning/control where the two approaches predicted about the same.

Figure 1: The number of USPTO patent documents published each year that were predicted to be AI by the original AIPD and the 2023 AIPD update



Notes: The original AIPD runs through the end of 2020 and the AIPD 2023 update through the end of 2023. The figure uses a 50% prediction threshold. A document is predicted as "Any AI" if it is predicted as AI in any one of the eight AI component technologies.

One way to adjust the overall number of AI predictions is to change the probability threshold for determining AI. In the prior section we used a 50 percent threshold—those documents with predictions of at least 50 percent were labeled AI in that component while those strictly less than 50 percent were determined not to be AI in that component.[25] Raising the threshold generally increases precision and lowers recall since only documents reaching the new, higher probability threshold are predicted to be AI, but conversely a greater number of true AI documents with intermediate probabilities are missed. Researchers may favor greater precision or recall depending on their application, or they may seek to replicate an existing

[25] If a document is predicted as AI in any component then it is identified as being "any AI" (as in Giczy et al. 2022 and Toole et al. 2020b).

analysis with extended data from the AIPD 2023 that more closely aligns with the original AIPD.[26]

From the perspective of accurately predicting the volume of AI, one would like to balance precision and recall since:

$$N_{AI} \cdot Recall = M_{AI} \cdot Precision \quad \text{(Eq. 1)}$$

where N_{AI} is the true volume of AI and M_{AI} is the volume of AI predicted by the model (both sides of the equation are the number of true AI documents predicted by the model). If precision equals recall, then the model accurately predicts the true volume of AI. As discussed above, the AIPD 2023 had better performance overall than the original AIPD. However, recall was higher than precision for the AIPD 2023 in both the training statistics (Table 2) and the manual evaluation (Table 4, Panel a). If this relationship between recall and precision extends to the population of patent documents (i.e., $recall > precision$), then the number of documents predicted to be AI in the AIPD 2023 would be biased upward (since $M_{AI} = \frac{Recall}{Precision} \cdot N_{AI} > N_{AI}$).

From the perspective of the original AIPD, Figure 8 in Giczy et al. (2022) shows precision, recall, and F1 estimated from a holdout sample of 368 patent documents for every AI prediction threshold. In this figure, precision and recall were relatively balanced at the threshold of 50 percent (at 40.5 percent for precision and 37.5 percent for recall) and were equal at a threshold of 35 percent. Unfortunately, we cannot reproduce the analysis that adjusts the prediction threshold in Giczy et al. (2022) for the AIPD 2023 since we used the original AIPD holdout documents to train the AIPD 2023 models (i.e., the "examiner annotated" training data). However, we can more accurately determine the volume of AI with the AIPD 2023 by using the 35 percent threshold with the original dataset to determine which threshold in the AIPD 2023 would be necessary to replicate a prediction volume that balances precision and recall.[27] To accomplish this task, we analyze different thresholds for the AIPD 2023 to calibrate the prediction volume to that from the original AIPD at a 35 percent threshold.

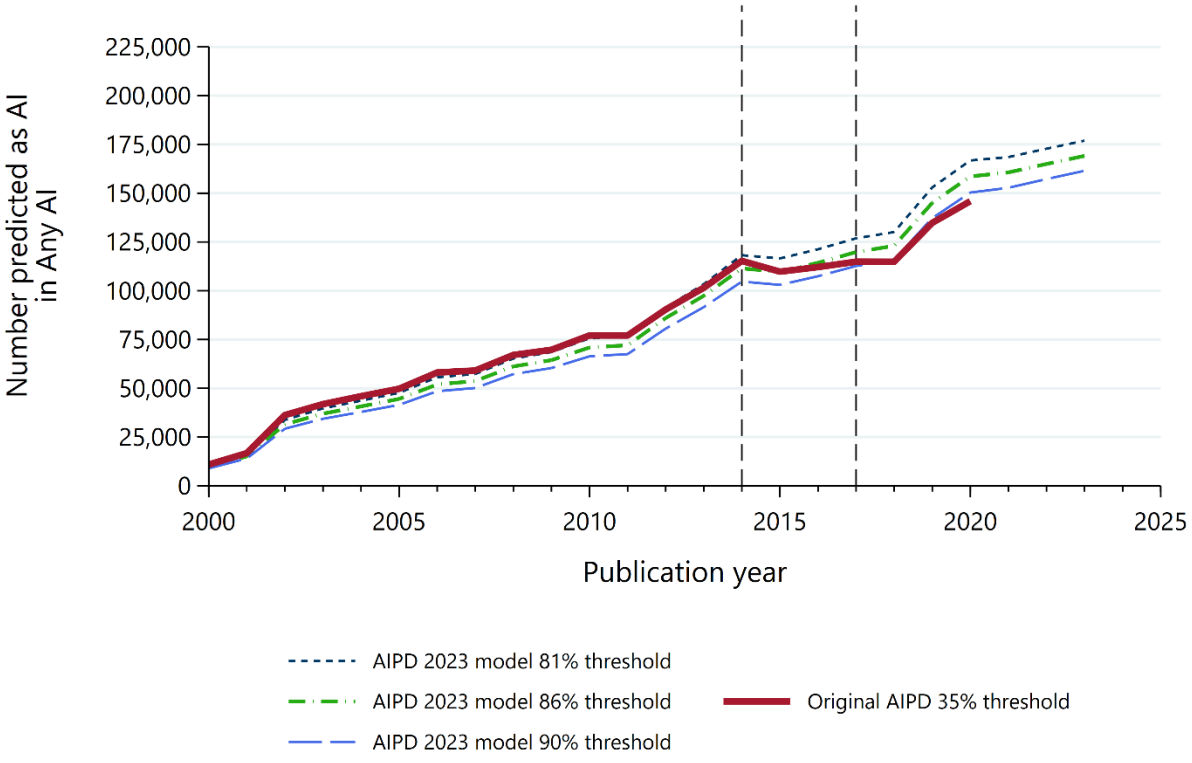
[26] For example, Grashof et al. 2023 prefers the WIPO keyword/classification approach for identifying AI invention in patent documents because of its higher precision. Rather than switching to a method like this, researchers can increase the prediction threshold for AI to increase precision.

[27] One caveat is that the precision and recall estimates provided from the original AIPD were from a random sample of patent documents outside of the training set, and therefore are not population estimates. However, given that the seed and anti-seed sets were only 0.07 and 0.88 percent of the population, respectively, a simple random sample of the size annotated for the original AIPD would have overwhelmingly contained non-training documents (i.e., if the 368 documents had been drawn randomly, the expected number of seed documents would have been 0.23 while the anti-seed would have been 3.22). Therefore, the precision and recall estimates from the non-training set in the original AIPD closely approximate the overall population estimates obtainable from a random sample of similar size.

Figure 2 summarizes this analysis. An AIPD 2023 threshold of 81% (blue dashed line) matches the original AIPD at a 35% threshold from 2000 to 2014, while an AIPD 2023 threshold of 90% (bright blue long dash line) matches 2017 and after. An AIPD 2023 threshold of 86% is the midpoint between these two threshold estimates (green dash-dot line) and appears to split the difference. Thus, researchers could select one of these AIPD 2023 thresholds, either the upper bound (81%), lower bound (90%) or midpoint (86%) to obtain a prediction volume that more closely balances precision and recall. Importantly however, while modifying thresholds does adjust the volume of AI predicted by the models in aggregate, i.e., for "any AI," it does not identify the same U.S. patent documents as AI.

In addition, researchers who would like to replicate the prediction volumes from the original AIPD at a 50% threshold might select a threshold of 93% for the AIPD 2023 (see Figure C1 in Appendix C).

Figure 2: The number of USPTO patent documents published each year between 1976 and 2023 that were predicted to be AI comparing the 2023 updated with varying prediction thresholds to the original AIPD at a 35% threshold

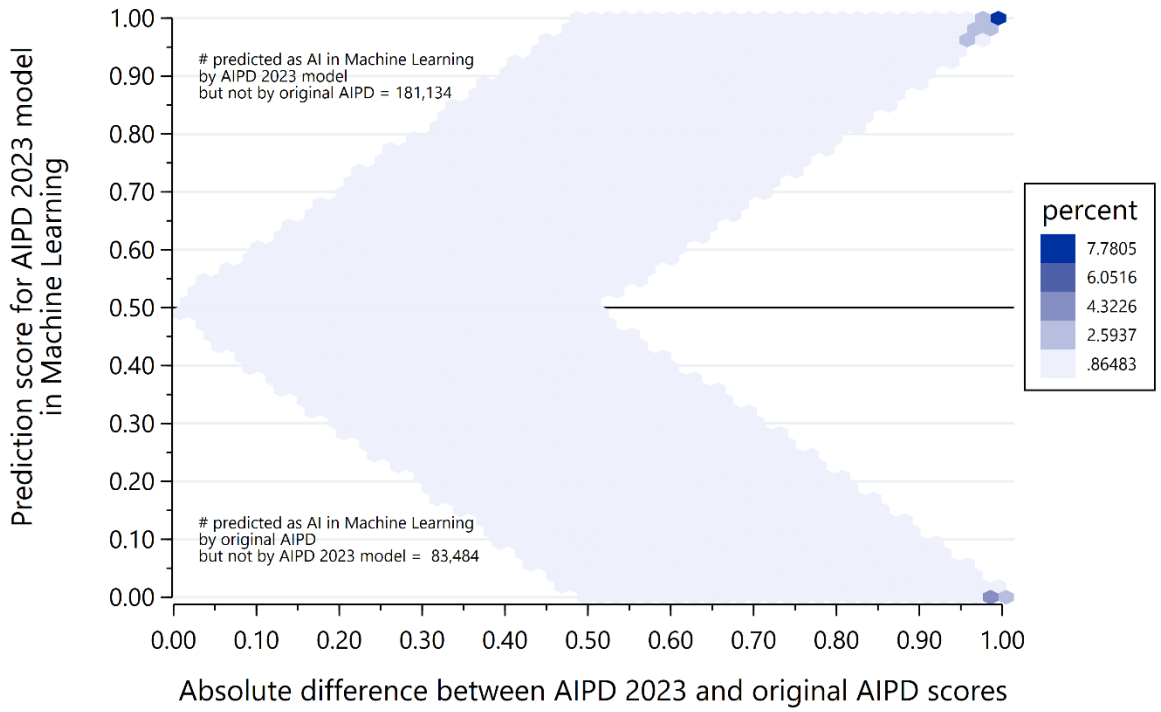


More information on the "disagreements" with the original AIPD

The manual evaluation discussed above compared the AIPD 2023 to the original AIPD on a set of patent documents published in 2019 where the two approaches disagreed (using 50 percent

thresholds for both), finding that the 2023 update had better performance (see Table 4). To better understand these differences, we further analyzed how the predicted probabilities from both models differed on the set of disagreements. Figure 3 shows machine learning prediction scores from the AIPD 2023 on the y-axis relative to the absolute difference between the prediction scores of the updated 2023 and original model on the x-axis for those documents where there is a disagreement at the 50% threshold. The figure reveals that when the two models disagree, they disagree substantially (the figures for the other AI component technologies are similar, and are available upon request). For example, the greatest density of disagreements occurs when the AIPD 2023 predicts AI with near certainty (i.e., close to 1.0), and the original AIPD predicts not AI with near certainty (i.e., near 0.0, thus resulting in an absolute difference close to 1.0), and vice versa. In other words, the models are not disagreeing most where one or the other is uncertain (i.e., where one or both models predict near 50 percent), but where they are each almost completely certain on the outcome (which is wrong for one of the models).

Figure 3: Differences between machine learning predicted probabilities for the original AIPD and 2023 update on the set of patent documents where the two approaches disagree at the 50 percent threshold



Notes: The figure includes all patent document having predictions from both the AIPD 2023 model and its corresponding original AIPD model where the two models differed in AI versus not AI (in that AI component) based

on a 50% threshold for each model. In an ideal situation the largest percentage of observations should be clustered around an AIPD 2023 model score of 0.50 (y-axis) and a difference in absolute scores close to zero (x-axis), i.e., near left tip of the "arrow" in the figure.

Figure 3 also illustrates that there is a large degree of variability in the relative predictions, since almost all combinations of valid values are present (thereby forming a completely filled in "arrow" shape). Table 6 provides more information on this variability by presenting the percentage of documents that are in each of four sections of the arrow in Figure 3: (1) upper right, which represents positive AI predictions in the AIPD 2023 at 0.90 or higher while the original AIPD predicts AI at 0.10 or lower;[28] (2) lower right, which represents the opposite; (3) the "tip" of the arrow figure, where both models have a relatively high degree of uncertainty, i.e., the AIPD 2023 predicts at between 0.40 and 0.60 and the original model is within 0.20 of the AIPD 2023 prediction; and (4) the remainder of the figure not in (1), (2) or (3).

Columns (1) and (2) in Table 6 quantify our observation that a small area of Figure 3, i.e., those disagreements where both models were very certain in their predictions, form a substantial share of the overall disagreements. For example, this small area accounts for nearly 50 percent of the disagreements for speech, and about 42 percent for machine learning. In AI hardware, the share is smaller but still large, at about 20 percent. Further, the area where the AIPD 2023 was most uncertain and the original AIPD was also generally uncertain, i.e., the arrow tip in Column (3), contains very few disagreements (ranging from a high of 1.12 percent to a low of 0.20 percent).

Table 6. Distribution of prediction scores between the AIPD 2023 and original AIPD models

AI Component	Zone	(1) Upper right	(2) Lower right	(3) Arrow tip	(4) All others
	AIPD 2023 prediction score	0.9 and above	0.1 and below	Between 0.6 and 0.4	Remaining
	Absolute difference AIPD 2023 and original AIPD	0.9 and above	0.9 and above	0.2 and below	Remaining
Machine learning		28.80%	13.22%	0.29%	57.69%
Evolutionary computation		22.01%	1.14%	0.37%	76.48%
Natural language processing		35.81%	1.56%	0.38%	62.25%
Vision		24.69%	5.63%	0.60%	69.09%
Speech		36.81%	11.24%	0.20%	51.75%
Knowledge processing		8.25%	27.64%	0.52%	63.59%
Planning and control		16.21%	17.51%	0.49%	65.79%

[28] Since in the first zone the absolute difference between the AIPD 2023 and original predictions are 0.90 or higher, if the AIPD 2023 predicts AI at 0.90 and above, then the original AIPD must predict AI at between 0.0 and 0.10.

AI hardware	17.31%	2.96%	1.12%	78.62%

Notes: See Figure 3 for a visualization of the regions provided in Columns (1)-(4) for machine learning.

These findings, combined with the overall large number of disagreements observed in Table 3, show that the models are highly sensitive to the underlying data used for training and the approach used to embed the text (the two major differences between the original AIPD and the AIPD 2023). Augmenting the training data by including annotated observations where the active learning model was most uncertain (as well as the examiner annotations and the new AI publications since 2019) and using BERT for Patents instead of Word2Vec dramatically moved the decision boundary, resulting in a new model that disagreed substantially with the previous approach. Despite these large changes, the performance improvement revealed in Table 4 emphasizes the importance of selecting an appropriate embedding approach and generating high quality training data; for example, by using active learning to generate data that allows the model to better learn the location of the decision boundary.

Comparison to other AI patent datasets

Giczy et al. (2022) benchmarked the original AIPD against several alternatives in the literature, including the Cockburn et al. (2019) and WIPO (2019) patent classification and keyword approaches, finding that the AIPD model significantly outperformed these other methods. The key finding was that these other approaches achieved high precision by specifying narrow queries to identify AI but suffered disproportionately in recall, thereby achieving relatively low F1 scores. By comparison, the original AIPD had lower precision, but disproportionately higher recall, resulting in a higher F1 score that, although not as high as that achieved by USPTO examiners, was much closer than the other approaches.

Since the publication of the original AIPD in 2021, an influential AI patent dataset produced by the Center for Security and Technology (CSET) (Thomas and Murdict 2020) has been used in several policy analyses, including Stanford's AI Index (Zhang et al. 2022; Maslej et al. 2024) and the National Science Foundation's Invention, Knowledge Transfer and Innovation report (Robbins 2024).[29,30] CSET's approach for identifying AI patents differs from ours in two significant ways. First, CSET's definition of AI relies on the Association for Computing Machinery approach that categorizes AI along 35 dimensions, including AI techniques (e.g., machine learning and logic programming), functional applications (e.g., language processing and computer vision), and application fields (e.g., life sciences and banking/finance). Our definition of AI overlaps significantly with the ACM taxonomy, but we do not use the same AI categories

[29] Code for implementing the CSET approach has been made available on GitHub at https://github.com/georgetown-cset/1790-ai-patent-data.

[30] The CSET AI data is also used by Our World in Data (https://ourworldindata.org/grapher/artificial-intelligence-granted-patents-by-industry) and articles in the popular media, including by Axios (e.g., https://www.axios.com/local/san-francisco/2024/04/03/silicon-valley-patents-ai-chatgpt).

and do not classify directly into AI application fields, preferring to use our algorithm to find AI wherever it exists across technologies (see Toole et al. 2020).

The second major difference is that CSET uses patent classifications and keywords to identify AI, similar to the approaches used by WIPO (2019) and Cockburn et al. (2019). Therefore, we might expect CSET's approach to favor precision over recall, and as a result underreport the true volume of AI. Figure 1 in Thomas and Murdict (2020) reveals this to be the case, finding that just over 10,000 patents and around 65,000 applications were published *worldwide* in 2020. By comparison, the original AIPD at the threshold of 35 percent (to balance precision and recall) has nearly 150,000 U.S. PGPubs and patents published in 2020 from the USPTO alone. Despite this fact, the CSET approach has at least one major advantage—it's easily extendable to worldwide patent datasets, whereas the AIPD is significantly more computationally intensive and is currently only available for USPTO publications.

Practical challenges associated with patent landscaping

We faced several practical challenges when updating the AIPD, which we hope by discussing here will lower the barriers for other economics, legal and business researchers considering using these or similar methods. First, we required significant computational resources to train our models and execute predictions. The server we used had 112 CPUs, 1.47 TB RAM, and eight NVIDIA A100-40GB GPUs. Due to the amount of data required to train the models, where we processed BERT for Patents text sub-word tokens through long short-term memory (LSTM) neural networks, we used only the CPUs for training. Training the final models took an average of approximately 1.3 hours for each AI technology component model for a total time of approximately 10.6 hours. To run model predictions, we divided the approximately 15.4M patent documents into "shards" of 1000 documents each, and then used one GPU to execute the predictions for groups of 200-800 shards at a time. A group of 400 shards took about 24 hours to run in a single GPU enabled process, and we used up to four GPUs and processes simultaneously. The predictions took a total of about 11 calendar days of near constant processing using this parallel approach. A significant time-consuming portion of the process was converting text into BERT for Patents embeddings; since we used all the sub-word tokens of the embedding, these files were very large and could not be reasonably kept beyond their immediate use, particularly for predictions.[31]

[31] For example, the files for training each AI component technology model were approximately 68-79 GB in size. Running predictions for a shard of 1000 patent documents required approximately 2 GB for each of abstract text and claims text (4 GB total); we kept the embeddings for a shard in memory and ran predictions for all eight classification models before discarding them. Given 15.4M documents, saving all embedding files to disk would have required over 200 TB.

A second challenge is associated with evolving patent classification systems, especially for emerging technologies such as AI. New classification symbols may be created, to include new symbols split from old ones or the creation of subordinate symbols, and some symbols may be retired. It is thus important to distinguish whether the classification symbols in patent data are from when the document was published (or granted as a patent) or have been updated to the most recent classification schema and symbols. In our analysis, this challenge affected how we updated the training data beyond 2019, requiring us to modify the classification-based seed set queries originally used to identify documents likely to contain AI (see Appendix B). Many of the original classifications had been replaced, while new symbols had been added, making it challenging to update the training dataset in a way consistent with the scope of the original AIPD queries.

Conclusion

The AIPD 2023 extends the AIPD to all USPTO patent and PGPubs through 2023, while also improving the underlying methodology used for identifying AI patent documents. The major methodological changes include the use of BERT for Patents to embed patent document abstracts and claims, as well as new training data selected through active learning to better identify where the decision boundary exists between AI and non-AI. In addition to results supporting this method from existing AI patent landscaping research (Islam Erana and Finlayson 2024), our manual evaluation shows that this new method performs better than the original AIPD approach.

Our study reveals several important insights beyond this overall finding. First, identifying AI in patent documents remains difficult, even for human experts. The research community and policymakers would benefit from greater exploration into the sources of these difficulties; for example, do we need better definitions of AI or better guidelines for human annotators when creating training datasets? Improved annotation would translate directly into improved landscaping performance. From the perspective of the AIPD 2023, the model for AI hardware performed substantially worse than the original model, perhaps because it may be an especially challenging area of technology to identify and our training dataset included annotations from many different reviewers. Beyond AI, researchers could create strategies to employ when developing training datasets for technology areas of various annotation difficulty.

In addition, a promising area of future research would be to explore model performance with abstracts and titles alone, as these are readily available in European Patent Office's (EPO) Worldwide Patent Statistical Database (PATSTAT)[32] and therefore would allow researchers to

[32] Patent application titles are included in PATSTAT table TLS202, and abstracts in TLS203. See PATSTAT Global Data Catalog, available at https://www.epo.org/en/searching-for-patents/business/patstat.

extend this approach to patent documents published worldwide. Importantly, this model should require fewer computational resources than our current approach since it would only rely on abstract/title text and not claims. Relatedly, the machine learning architecture could be simplified by taking advantage of text summarization embedding vectors, e.g., [CLS] tokens from BERT for Patents, as opposed to using individual sub-word tokens in complex LSTM networks (see Ghosh et al. 2024 for an implementation of such an approach based on Bert for Patents[33]). By characterizing these tradeoffs, researchers could make better decisions regarding the costs and benefits of different approaches to patent landscaping.

Our analysis revealed the importance of selecting an appropriate prediction threshold for a given application. From the perspective of accurately predicting the volume of AI, researchers should try and balance precision and recall as much as possible. However, this threshold may not be appropriate for other applications, e.g., when assessing diffusion, a researcher might be more concerned about increasing recall at the expense of precision to better assess the reach of a given technology. To the best of our knowledge, very little applied research exists that explores the impact of adjusting precision and recall within applied applications. While likely highly dependent on each application and therefore difficult to characterize, greater exploration into this issue would improve the evidence base derived from the identification of specific technologies within patent data.

Finally, in recent years, the economics, management, and legal research communities have begun using generative AI within the research process itself (see Korinek 2023). While we did not use generative AI to update the AIPD, these methods appear promising but also introduce new challenges. For example, how might researchers ensure the generative AI system uses a given technology definition, and if documents are labeled at different times, ensure the system consistently uses the same definition of technology? As we described earlier, these problems also exist with human labelers, but they are perhaps harder to solve with generative AI as it can be difficult to assess the reasons for its decision-making.

To help researchers, practitioners, and policy-makers better understand the determinants and impacts of AI invention, we have made the AIPD 2023 publicly available on the USPTO's economic research web page (https://www.uspto.gov/ip-policy/economic-research/research-datasets/artificial-intelligence-patent-dataset). More information on the dataset is available in Appendix D, and Appendix E provides helpful information on how researchers may link the AIPD 2023 to other patent data fields, such inventors, assignees, and their locations, using publicly available data from the USPTO-sponsored PatentsView data platform (www.patentsview.org).

[33] Ghosh et al. (2024) used the mean of the output layer embedding tokens, finding it outperformed the BERT [CLS] token.

References

Abood, A. and Feltenberger, D. 2018. Automated patent landscaping. Artificial Intelligence and Law, 26(2), pp.103-125.

Azoulay, P., Krieger, J., and Nagaraj, A. 2024. Old Moats for New Models: Openness, Control, and Competition in Generative AI1 (No. c15002). National Bureau of Economic Research.

Beliveau, S. and Ma, J. 2022. Recent Developments in AI and USPTO Open Data. arXiv preprint arXiv:2207.05239.

Bickley, S. 2023. Bridging complexity and behavioural economics: The constrained methods matrix (Doctoral dissertation, Queensland University of Technology, Brisbane, Australia).

Cao, S., Goldstein, I., He, J., and Zhao, Y. Feedback on Emerging Corporate Policies.

Charmanas, K., Georgiou, K., Mittas, N., and Angelis, L. 2023. Classifying the Main Technology Clusters and Assignees of Home Automation Networks Using Patent Classifications. Computers, 12(10), p.211.

Chattergoon, B. and Kerr, W.R. 2022. Winner takes all? Tech clusters, population centers, and the spatial transformation of US invention. Research Policy, 51(2), p.104418.

Chikkamath, R., Rastogi, D., Maan, M., and Endres, M. 2024. Is your search query well-formed? A natural query understanding for patent prior art search. World Patent Information, 76, p.102254.

Choi, S., Lee, H., Park, E., and Choi, S. 2022. Deep learning for patent landscaping using transformer and graph embedding. Technological Forecasting and Social Change, 175, p.121413.

Chowdhury, F., Link, A.N., and van Hasselt, M., 2022. Public support for research in artificial intelligence: a descriptive study of US Department of Defense SBIR Projects. The Journal of Technology Transfer, 47(3), pp.762-774.

Clarivate. 2024. Derwent World Patents Index (DWPI) classification system. https://clarivate.com/dwpi-reference-center/dwpi-classification-system/.

Cockburn, I.M., Henderson, R., and Stern, S. 2019. The Impact of Artificial Intelligence on Innovation. The Economics of Artificial Intelligence: An Agenda, p.115.

Dacus, C. and Horn, C.C. 2022. Defensive industrial policy: Cybersecurity interventions to reduce intellectual property theft. Acquisition Research Program.

Dentamaro, V., Giglio, P., Impedovo, D., and Veneto, D. 2023, April. Matching Knowledge Supply and Demand of Expertise: A Case Study by Patent Analysis. In World Conference on Information Systems and Technologies (pp. 321-329). Cham: Springer Nature Switzerland.

Denter, N. 2022. Machine learning for patent intelligence: opportunities and challenges. (Doctoral dissertation, Universität Bremen, Bremen, Germany).

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Frumkin, J., Pairolero, N.A., Tesfayesus, A., and Toole, A.A. 2024. Patent eligibility after Alice: Evidence from USPTO patent examination. Journal of Economics & Management Strategy. https://doi.org/10.1111/jems.12592.

Gao, X., Wu, Q., Liu, Y., and Yang, R. 2024. Pasteur's quadrant in AI: do patent-cited papers have higher scientific impact? Scientometrics, pp.1-24.

Gaske, M.R. 2023. Regulation Priorities for Artificial Intelligence Foundation Models. Vand. J. Ent. & Tech. L., 26, p.1.

Ghosh, M., Erhardt, S., Rose, M.E., Buunk, E., and Harhoff, D. 2024. PaECTER: Patent-level Representation Learning using Citation-informed Transformers. arXiv preprint arXiv:2402.19411.

Giczy, A.V., Pairolero, N.A., and Toole, A.A. 2022. Identifying artificial intelligence (AI) invention: A novel AI patent dataset. The Journal of Technology Transfer, 47(2), pp.476-505. https://doi.org/10.1007/s10961-021-09900-2.

Giczy, A.V., Pairolero, N.A., and Toole, A.A. 2024. Discovering value: women's participation in university and commercial AI invention. *Nature Biotechnology* 42, 26–29. https://doi.org/10.1038/s41587-023-02075-1.

Gomes, O., Mihet, R., and Rishabh, K. 2023a. A Growth and Innovation Model of the Modern Data Economy.

Gomes, O., Mihet, R., and Rishabh, K. 2023b. Cyber Risk-Driven Innovation in the Modern Data Economy. Centre for Economic Policy Research.

Graham, S.J., Marco, A.C. and Miller, R., 2018. The USPTO patent examination research dataset: A window on patent processing. Journal of Economics & Management Strategy, 27(3), pp.554-578.

Grashof, N. and Kopka, A. 2023. Artificial intelligence and radical innovation: an opportunity for all companies? Small Business Economics, 61(2), pp.771-797.

Haessler, P., Giones, F., and Brem, A. 2023. The who and how of commercializing emerging technologies: A technology-focused review. Technovation, 121, p.102637.

Hosseinioun, M. and Tafti, A.R. 2023. Search Benefits of General Resources: Recombination Premiums of AI. Available at SSRN 4669795.

Hötte, K., Tarannum, T., Verendel, V., and Bennett, L. 2022. Exploring Artificial Intelligence as a General Purpose Technology with Patent Data--A Systematic Comparison of Four Classification Approaches. arXiv preprint arXiv:2204.10304.

Hötte, K., Tarannum, T., Verendel, V., and Bennett, L. 2023. AI technological trajectories in patent data.

Hötte, K., Tarannum, T., Verendel, V., and Bennett, L. 2024. Measuring artificial intelligence.

Islam Erana, T. and Finlayson, M. 2024. Automated neural patent landscaping in the small data regime using citations and CPC codes. arXiv:2407.08001. https://arxiv.org/abs/2407.08001

Jacobi, C., Schwens, C., and Haarhaus, T. 2024. Knowledge spillovers from artificial intelligence faculty: The impact of universities on regional innovation ecosystems. (Master's thesis, Louvain School of Management, Université Catholique de Louvain, Louvain-la-Neuve, Belgium). http://hdl.handle.net/2078.1/thesis:44340.

Korinek, A. 2023. Generative AI for economic research: Use cases and implications for economists. Journal of Economic Literature, 61(4), pp.1281-1317.

Krestel, R., Chikkamath, R., Hewel, C., and Risch, J. 2021. A survey on deep learning for patent analysis. World Patent Information, 65, p.102035.

Lee, J.S. 2024. InstructPatentGPT: training patent language models to follow instructions with human feedback. Artificial Intelligence and Law, pp.1-44.

Lewis, D.D., and Gale, W.A. 1994. A sequential algorithm for training text classifiers. In Bruce W. Croft and C. J. van Rijsbergen, editors, SIGIR '94, pages 3–12, London, 1994. Springer London. ISBN 978-1-4471-2099-5.

Li, B., Huang, N., and Shi, W. 2022. Media Coverage of Labor Issues and Artificial Intelligence Innovation. Available at SSRN 4165159.

Liu, Y., Liu, C., and Huang, T. 2023. A New Index Measuring Occupational Exposure to Artificial Intelligence. Available at SSRN 4515629.

Lopez, M. and Gonzalez, I. 2024. Artificial Intelligence Is Not Human: The Legal Determination of Inventorship and Co-Inventorship, the Intellectual Property of AI Inventions, and the Development of Risk Management Guidelines. J. Pat. & Trademark Off. Soc'y, 104, p.135.

Maslej N., Fattorini L., Perrault R., Parli V., Reuel A., Brynjolfsson E., Etchemendy J., Ligett K., Lyons T., Manyika J., Niebles J., Shoham Y., Wald R., and Clark J. 2024. "The AI Index 2024 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, California.

Messeni Petruzzelli, A., Murgia, G., and Parmentola, A. 2023. Opening the black box of artificial intelligence technologies: unveiling the influence exerted by type of organisations and collaborative dynamics. Industry and Innovation, 30(9), pp.1213-1243.

Mihet, R., Rishabh, K., and Gomes, O. 2024. Data Risk, Firm Growth and Innovation.

Montobbio, F., Staccioli, J., Virgillito, M.E., and Vivarelli, M. 2023. The empirics of technology, employment and occupations: lessons learned and challenges ahead. Journal of Economic Surveys.

Muraro, V. and Göktepe-Hultén, D. 2023, April. Assessment of academic work in AI using Bibliometrics: The Case of Lund University. In 27th International Conference on Science, Technology and Innovation Indicators (STI 2023). International Conference on Science, Technology and Innovation Indicators.

Park, J. 2023. Analyzing the Role of Governmental Organizations in Artificial Intelligence Innovation: A Patent-Based Perspective. KDI School of Pub Policy & Management Paper No. DS23-09.

Park, J. 2024 Analyzing the direct role of governmental organizations in artificial intelligence innovation. The Journal of Technology Transfer, pp.1-29.

Pelaez, S., Verma, G., Ribeiro, B., and Shapira, P. 2024. Large-scale text analysis using generative language models: A case study in discovering public value expressions in AI patents. Quantitative Science Studies, pp.1-26.

Picht, P.G., Brunner, V., and Schmid, R. 2022. Artificial intelligence and intellectual property law: from diagnosis to action. Max Planck Institute for Innovation & Competition Research Paper, (22-08).

Pujari, S., Strötgen, J., Giereth, M., Gertz, M., and Friedrich, A. 2022, December. Three real-world datasets and neural computational models for classification tasks in patent landscaping. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp.11498-11513.

Rathi, S., Majumdar, A., and Chatterjee, C. 2024. Did the COVID-19 pandemic propel usage of AI in pharmaceutical innovation? New evidence from patenting data. Technological Forecasting and Social Change, 198, p.122940.

Rezazadegan, R., Sharifzadeh, M., and Magee, C.L. 2024. Quantifying the progress of artificial intelligence subdomains using the patent citation network. Scientometrics, pp.1-23.

Robbins C.,2024. Invention, Knowledge Transfer and Innovation. National Science FoundationShan, C., Wang, J., and Zhu, Y. 2023. The Evolution of Artificial Intelligence in the Digital Economy: An Application of the Potential Dirichlet Allocation Model. Sustainability, 15(2), p.1360.

Shi, J. and Wang, Y. 2024. Prerequisites for the innovation performance of artificial intelligence laboratory: A fuzzy-set qualitative comparative analysis. IEEE Transactions on Engineering Management.

Sidorov, N. and Szabó, I.G. 2023. Technology and the Future of Work: Comparing the Potential Impact of AI Industrialisation on Labour Markets in Liberal and Coordinated Market Economies. (Master's thesis, Central European University, Vienna, Austria).

Spulber, D.F. and Wang, X. 2023. Knowledge as Output and as Input: Artificial Intelligence and Quantum Computing. Available at SSRN 4382480.

Srebrovic, R. and Yonamine, J. 2020. Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery. White paper.

Straub, J. 2021. Gradient descent training expert system. Software Impacts, 10, p.100121.

Thomas, P. and Murdick, D. 2020. Patents and Artificial Intelligence: A Primer. CSET Data Brief (September 2020). Center for Security and Emerging Technology (CSET), Georgetown University, Washington, District of Columbia, United States. DOI, 10, p.20200038.

Toole A.A. and Pairolero N.A. 2020a. Adjusting to Alice: USPTO patent examination outcomes after Alice Corp. v. CLS Bank International. United States Patent and Trademark Office. IP Data Highlights. Number 3, AprilToole, A., Pairolero, N., Giczy, A., Forman, J., Pulliam, C., Such, M., and Rifkin, B. 2020b. Inventing AI: Tracing the diffusion of artificial intelligence with US patents. United States Patent and Trademark Office. IP Data Highlights, Number 5, October.

Tu, S.S., Cyphert, A., and Perl, S. 2024. Artificial Intelligence: Legal Reasoning, Legal Research and Legal Writing. Artificial Intelligence: Legal Reasoning, Legal Research and Legal Writing (May 5, 2024). Minn. JL Sci. & Tech. (Forthcoming).

Vidal, K. 2022. Report to Congress, Patent eligible subject matter: Public views on the current jurisprudence in the United States. United States Patent and Trademark Office. https://www.uspto.gov/sites/default/files/documents/USPTO-SubjectMatterEligibility-PublicViews.pdf.

Vowinckel, K. and Hähnke, V.D. 2023. SEARCHFORMER: Semantic patent embeddings by siamese transformers for prior art search. World Patent Information, 73, p.102192.

Wang, J. 2023. Combinatorial Inventions in Artificial Intelligence: Empirical Evidence and Implications for Science, Technology, and Organizations (Doctoral dissertation, Arizona State University, Tempe, Arizona).

WIPO (World Intellectual Property Office), 2019. WIPO Technology Trends 2019 – Artificial Intelligence. Geneva, Switzerland: World Intellectual Property Organization.

Wu, L., Lou, B., and Hitt, L.M. 2024. Innovation Strategy after IPO: How AI Analytics Spurs Innovation after IPO. Available at SSRN.

Yoo, Y., Jeong, C., Gim, S., Lee, J., Schimke, Z., and Seo, D. 2023. A Novel Patent Similarity Measurement Methodology: Semantic Distance and Technological Distance. arXiv preprint arXiv:2303.16767.

Zhang D., Maslej N., Brynjolfsson E., Etchemendy J., Lyons T., Manyika J., Ngo H., Niebles J., Sellitto M., Sakhaee E., Shoham Y., Clark J., and Perrault R. 2022. "The AI Index 2022 Annual Report," AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, Sanford, California.

Appendix A – Summary of using or citing literature

Table A1: Categorization of literature that uses the AIPD or cites Giczy et al. (2022)

Research category	References
AIPD utilized in research dataset	Cao et al. (2024), Chattergoon and Kerr (2022), Gao et al. (2024), Giczy et al. (2024), Gomes et al. (2023a; 2023b), Hosseinioun and Tafti (2023), Jacobi et al. (2024), Lee (2024), Li et al. (2022), Liu et al. (2023), Mihet et al. (2024), Park (2023), Park (2024), Rathi et al. (2024), Rezazadegan et al. (2024), Sidorov and Szabó (2023), Spulber and Wang (2023), Wu et al. (2024), Yoo et al. (2023)
AIPD informs research method or provides background material	Azoulay et al. (2024), Beliveau and Ma (2022), Bickley (2023), Charmanas et al. (2023), Chowdhury et al. (2022), Dacus and Horn (2022), Dentamaro et al. (2023), Denter (2022), Gaske (2023), Grashof and Kopka (2023), Haessler et al. (2023), Hyun and Kim (2024), Lopez and Gonzalez (2024), Petruzzelli et al. (2023), Montobbio et al. (2023), Muraro and Göktepe-Hultén (2023), Palaez et al. (2024), Picht et al. (2022), Pujari et al. (2022), Shan et al. (2023), Shi et al. (2024), Straub (2021), Tu et al. (2024)
AIPD utilized in comparison of landscaping methods	Hötte et al. (2022; 2023; 2024)

Appendix B – Training data

Updated seed set search queries

Table B1 summarizes the search queries we used to update the AI component technology seed sets beyond 2018. All queries, except for AI hardware and planning and control, were broken up into two steps. Step 1 involved querying Cooperative Patent Classification (CPC), U.S. Patent Classification (USPC), and International Patent Classification (IPC) codes, and step 2 further limited the results of step 1 with Derwent World Patents Index (DWPI) classification codes.[34] Both steps included only those documents published since 2019. The queries were separated into two steps since the Derwent search database was not updated with the most current classification codes, e.g., documents in the Derwent database were classified at publication, rather than being updated to reflect CPC scheme changes. Step 1 was executed using the USPTO examiner search tool, PE2E search, to avoid this limitation. The AI hardware component was further broken into a third part (consistent with the AI hardware query from Giczy et al. 2022), where the first part was as described above, the second part did not have a Derwent search component, and the third part combined the first and second parts using the "or" operator. For planning and control we combined steps 1 and 2 into a single query to overcome a technical problem encountered when using the examiner search tool.

The number of documents returned by each query and the number of documents added to the seed sets are summarized in Table B2. We limited the number added to approximately 10% of the previous AIPD seed set except for evolutionary computation, for which all results were added due to the small size of the original seed set for that category.

Table B1: CPC, IPC, USPC, and Derwent queries used to update seed training data from 2019 and beyond

AI Component	Step 1: CPC, USPC, and IPC Query (called Q1 in Step 2)	Step 2: Derwent Query
Machine learning	(G06N3/02 OR G06N3/04$ OR G06N3/08$ OR G06N3/09$ OR G06N3/10$ OR G06N20/$ OR G06N7/00 OR G06N7/01 OR G06N7/02 OR G06N7/023 OR G06N7/08).cpc. AND (706/12 OR 706/14-19 OR 706/20 OR 706/22 OR 706/25).cor. AND (G06N3/02 OR G06N3/04$ OR G06N3/08$ OR G06N3/09$ OR G06N3/10$ OR G06N20/$ OR G06N7/00 OR G06N7/01 OR G06N7/02 OR G06N7/08).ipcr,cipg,cicl,cips. AND @py>="2019"	(Q1) AND (T01-J16C1$ OR T01-J16C2$ OR T01-J16C6$).EMCD,CMCD. AND US.pfpc. AND @py>="2019"

[34] See https://clarivate.com/dwpi-reference-center/dwpi-classification-system/.

AI Component	Step 1: CPC, USPC, and IPC Query (called Q1 in Step 2)	Step 2: Derwent Query
Evolutionary computation	(G06N3/086 OR G06N3/12$ OR G06N3/00$).cpc. AND 706/13.cor. AND (G06N3/086 OR G06N3/12$ OR G06N3/00$).ipcr,cipg,cicl,cips. AND @py>="2019"	(Q1) AND (T01-J16C4$).EMCD,CMCD. AND US.pfpc. AND @py>="2019"
Natural Language Processing	(G06F40/$).cpc. *AND* (704/? *OR* 704/10).cor. *AND* (G06F40/$).ipcr,cipg,cicl,cips. AND @py>="2019"	(Q1) AND (T01-J16C3$ OR T01-J14$).EMCD,CMCD. AND US.pfpc. AND @py>="2019"
Vison	(G06V10/$ OR G06V20/$ OR G06V30/$ OR G06V40/$ OR G06T9/$ OR G06T2211/441).cpc. AND (382/$).cor. AND (G06V10/$ OR G06V20/$ OR G06V30/$ OR G06V40/$ OR G06T9/$).ipcr,cipg,cicl,cips. AND @py>="2019"	(Q1) AND (T01-J10B$ OR T04-D$).EMCD,CMCD. AND (T01-J16$).EMCD,CMCD. AND US.pfpc. AND @py>="2019"
Speech	(G10L15/$ OR G10L17/$ OR G10L21/$ OR G10L25/$ OR G10L13/$).cpc. AND (704/2??$).cor. AND (G10L15/$ OR G10L17/$ OR G10L21/$ OR G10L25/$ OR G10L13/$).ipcr,cipg,cicl,cips. AND @py>="2019"	(Q1) AND (T01-C08A$ OR W04-V$).EMCD,CMCD. AND (T01-J16$).EMCD,CMCD. AND US.pfpc. AND @py>="2019"
Knowledge processing	(G06N5/$).cpc. AND (706/45 OR 706/46 OR 706/47 OR 706/48 OR 706/49 OR 706/5? OR 706/60 OR 706/61).cor. AND (G06N5/$).ipcr,cipg,cicl,cips. AND @py>="2019"	(Q1) AND (T01-J16$).EMCD,CMCD. AND US.pfpc. AND @py>="2019"
AI Hardware, PART 1	(G06F9/$ OR G06T1/20 OR G06T1/60 OR H04N19/42$ OR H04N19/43$).cpc. AND (708/$ OR 712/$ OR 326/$ OR 257/$ OR 365/$ OR 711/$).cor. AND (G06F9/$ OR G06T1/20 OR G06T1/60 OR H04N19/42$ OR H04N19/43$).ipcr,cipg,cicl,cips. AND @py>="2019"	(Q1, Part 1) AND (T01-J16$).EMCD,CMCD. AND US.pfpc. AND @py>="2019"
AI Hardware, PART 2	(G06N3/06$ OR G06N7/04$).cpc. AND (G06N3/06$ OR G06N7/04$).ipcr,cipg,cicl,cips. AND @py>="2019"	(Q1, Part 2)
AI Hardware, PART 3		(Q1, Part 1) OR (Q1, Part 2)
	Combined Step 1 and Step 2: CPC, USPC, IPC, and Derwent query	
Planning and control	(G06Q10/$ OR G05B13/$ OR G05B17/$ OR G06N3/008).cpc. AND (G06Q10/$ OR G05B13/$ OR G05B17/$ OR G06N3/008).ipcr,cipg,cicl,cips. AND (T01-J16$ OR T06-A05A$).EMCD,CMCD. AND US.pfpc. AND @py>="2019"	

Notes: Queries for current CPC classification codes use ".cpc."; current USPC use ".cor."; current IPC use ".ipcr,cipg,cicl,cips." (IPC, primary and secondary; IPC group; IPC class; and IPC secondary, respectively); Derwent uses ".EMCD,DMCD.", where "US.pfpc." limits results to U.S. publications; '@py>="2019"' limits results to publication year 2019 and after; "$" and "?" are wildcards. U.S. patents and PGPubs may be searched within the USPTO Patent Public Search web-based application (which does not include public access Derwent databases); see https://www.uspto.gov/patents/search/patent-public-search.

Table B2: Results of 2019+ queries and additions to the training data seed set

AI Component	Modified previous AIPD seed set (note 1)	Query results (2019 and after)	Added to seed from 2019 query (note 2)	Updated baseline AIPD seed set (note 3)
Machine learning	959	2326	96	1055
Evolutionary computation	82	20	20	102
Natural Language Processing	1083	664	109	1192
Vison	803	1710	81	884
Speech	763	613	77	840
Knowledge processing	661	402	67	728
Planning and control	1451	7476	146	1597
AI hardware	2658	4389	266	2924

Notes: (1) The previous AIPD seed set was first modified by ensuring only one publication was included per application, and that publication was updated to be the latest, e.g., if a previous seed set PGPub was eventually granted a patent then the patent was used instead of the PGPub. (2) The number added was randomly selected from the 2019+ query results so as to increase the modified previous AIPD seed set by approximately 10% (except for evolutionary computation, to which all 2019+ query results were added). (3) The baseline seed set is further processed to remove duplicates between it and other training data; hence the numbers in this column may not match those in Table 1.

Training data construction

We constructed the training data by combining the documents from the updated seed/anti-seed set, decision boundary set, and examiner annotated sets. In order to not under or over-weight a given document during training, we ensured the same document was not in multiple sets, keeping only one with an order of precedence of: examiner annotated, decision boundary, and seed/anti-seed. Additionally, we removed documents without abstract and claims text following text pre-processing.[35]

[35] Some text was not present in our source text data, and during pre-processing we removed formulas (along with other things) which may result in no claims text, e.g., for compositions of matter where only the chemical formula is claimed (see Giczy et al. 2022 for more information on this data cleaning step).

Appendix C – Machine learning model and results

This appendix provides supplemental information regarding the machine learning model we used for the AIPD 2023, along with additional analysis of the results using alternative prediction thresholds.

Additional information about the AIPD 2023 methodology

Table C1 summarizes the methodological differences between the original AIPD and the AIPD 2023.

Table C1: Methodology changes between the original AIPD to AIPD 2023 update

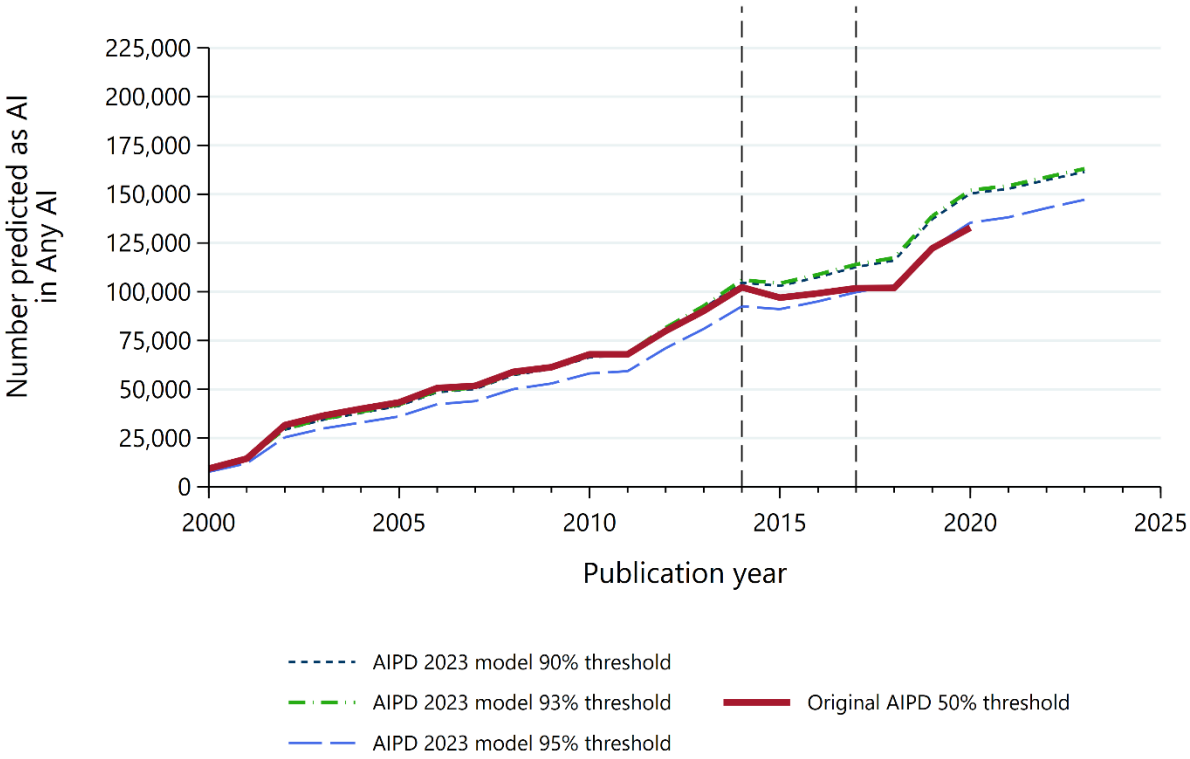
	Original AIPD	AIPD 2023 Update
Scope of patent landscape	• U.S. patents and PGPubs from 1976 to 2020, inclusive	• U.S. patents and PGPubs from 1976 to 2023, inclusive
Training Data	• Expansion method per Abood & Feltenberger (2018)	• Copied and updated/cleaned the training data from the original AIPD and increased seed sets by using classification queries for documents published 2019 and after • Decision boundary data from Florida International University • Examiner annotations from original AIPD evaluation • Sample weights to balance training data
Inputs	• Abstract text • Claims text • Citations	• Abstract text • Claims text
Word embedding	• Word2Vec with separate embedding for abstract text and claims text • Embedding vector for each word token	• BERT for Patents (limited to 512 sub-word tokens per BERT) used for both abstract and claims • Embedding vector for each word or sub-word token •
Citation embedding	• One-hot encoding (50,000 dimensions)	• N/A
Classification model	• LSTM neural network for text, one branch each for abstracts and claims • Neural network branch for citations • Neural network to combine text and citation branches	• LSTM neural network for text, one branch each for abstracts and claims • Neural network to combine text branches

Notes: Not reflected in the table are python and TensorFlow version changes.

Alternative AIPD 2023 thresholds for original AIPD at 50%

Our previous analyses, e.g., Toole et al. (2020b), Giczy et al. (2022) and Giczy et al. (2024), used the original AIPD with a 50 percent threshold to identify AI invention. For researchers who would like to match the number of AI predictions each year with the original AIPD at a 50% threshold, Figure C1 below shows that a threshold of 93 percent is a reasonable estimate. The 93 percent threshold was obtained using an identical calibration exercise as that used earlier to determine the 86 percent threshold which matched the original AIPD at 35 percent (i.e., the threshold that balances precision and recall).

Figure C1: Number of USPTO patent documents published each year between 2000 and 2023 that were predicted to be AI comparing the 2023 updated with varying prediction thresholds to the original AIPD at a 50% threshold



Document disagreements from alternative thresholds

The following tables, similar to Table 3, provide a summary of the number of document prediction disagreements when using different predictions thresholds for the AIPD 2023 and the original AIPD. Table C2 uses an 86% threshold for the AIPD 2023 and a 35% threshold for the original AIPD and corresponds to Figure 2. Table C3 uses a 93% threshold for the AIPD 2023 and a 50% threshold for the original AIPD, corresponding to Figure C1 above. Both tables show that

a significant number of disagreements remain between the two models at these alternate threshold comparisons.

Table C2: Summary statistics on "disagreements" between the AIPD 2023 using an 86% prediction threshold and the original AIPD using a 35% threshold

	AI in AIPD 2023 (86%) but not original AIPD (35%)	Not AI in AIPD 2023 (86%) but AI in original AIPD (35%)	Total disagreements	Total AI predictions	Percentage of disagreements out of predictions
Machine learning	227,527	115,972	343,499	445,597	77.09%
Evolutionary computation	98,887	71,824	170,711	180,128	94.77%
Natural language processing	222,176	59,348	281,524	411,329	68.44%
Vision	403,249	234,389	637,638	886,911	71.89%
Speech	95,178	50,166	145,344	200,141	72.62%
Knowledge processing	236,868	669,501	906,369	1,178,042	76.94%
Planning and control	451,632	597,015	1,048,647	1,491,631	70.30%
AI hardware	518,820	366,926	885,746	1,164,760	76.05%
Any AI	581,584	608,393	1,189,977	2,296,424	51.82%

Notes: Includes all patent documents published between 1976 and 2020 and having predictions from both the updated AIPD 2023 and the original AIPD. Total disagreements are when one model (AIPD 2023 or original AIPD) predicts AI and the other does not. Total AI predictions is either model predicts AI. The difference between the total number of AI predictions and the total disagreements in each component technology is the number of agreements (i.e., both agree AI or not AI in that component). The percentage of disagreements is relative to the total number of AI predictions in that component.

Table C3: Summary statistics on "disagreements" between the AIPD 2023 using a 93% prediction threshold and the original AIPD using a 50% threshold

	AI in AIPD 2023 (93%) but not original AIPD (50%)	Not AI in AIPD 2023 (93%) but AI in original AIPD (50%)	Total disagreements	Total AI predictions	Percentage of disagreements out of predictions
Machine learning	98,839	100,233	199,072	294,575	67.58%
Evolutionary computation	61,493	42,906	104,399	110,447	94.52%

Natural language processing	143,403	44,809	188,212	300,871	62.56%
Vision	246,138	180,951	427,089	644,033	66.31%
Speech	59,749	41,045	100,794	151,737	66.43%
Knowledge processing	101,259	617,948	719,207	963,867	74.62%
Planning and control	222,729	524,647	747,376	1,154,721	64.72%
AI hardware	327,106	267,459	594,565	812,373	73.19%
Any AI	602,713	514,475	1,117,188	2,118,166	52.74%

Notes: Includes all patent documents published between 1976 and 2020 and having predictions from both the updated AIPD 2023 and the original AIPD. Total disagreements are when one model (AIPD 2023 or original AIPD) predicts AI and the other does not. Total AI predictions is either model predicts AI. The difference between the total number of AI predictions and the total disagreements in each component technology is the number of agreements (i.e., both agree AI or not AI in that component). The percentage of disagreements is relative to the total number of AI predictions in that component.

Appendix D – Dataset description

The predictions data file, **ai_2023_model_predictions**, is a patent document-level dataset that contains the AIPD 2023 model predictions for each granted patent published from 1976 to 2023 and PGPubs from 2001 to 2023, excluding those that were withdrawn. The dataset is structured identically to the original AIPD predictions file, **ai_model_predictions**, but with additional binary variables for the various prediction thresholds as discussed in the "Extensions and discussion" section above. For each AI component technology, including "any_ai" (i.e., the model predicts AI in at least one AI component technology), we provide binary variables for whether the model predicts AI at the 50 percent, 86 percent, and 93 percent prediction thresholds. In addition to these binary variables, we provide the raw predictions scores, as well as selected meta data on the patent documents themselves, including patent application numbers, document publication dates, and an indicator for whether each document is a patent.

Table D1 summarizes each variable in the prediction dataset, including variable names, types, and descriptions. As with the original dataset, the data file is a comma-separated csv file where string data types are double-quote delimited. The variable **doc_id** is the primary key and is formatted to be compatible with patent and PGPub identifiers in PatentsView data tables.[36] Publication dates and application numbers are sourced from PatentsView.

As a final note, the original AIPD predictions file contained fields for whether each document was included in a training set for each of the AI technology component models, and a separate file recording whether each document was a positive or negative example in these training datasets. This information will be released with the publication of Islam Erana and Finlayson (2024), and we will update the USPTO's AIPD webpage as new information about this dataset becomes available (https://www.uspto.gov/ip-policy/economic-research/research-datasets/artificial-intelligence-patent-dataset).

Table D1: Variables in the AIPD 2023 predictions file, "ai_2023_model_predictions"

Variable name	Type	Brief description
doc_id	str	Document number: 7 or 8 digits for utility patents, "RE" followed by 5 digits for reissue patents, and 11 digits for PGPubs (4-digit year followed by number without intermediate slash)
flag_patent	int	Patent flag: 1 for patent, 0 for PGPub
pub_dt	str	Document publication date in YYYY-MM-DD format; equivalent to issue date for granted patents; in the Stata .dta file, this variable is in Stata date format %td_CY-N-D (displayed as YYYY-MM-DD)

[36] See "Datasets" at https://patentsview.org/. More information about using the AIPD 2023 with PatentsView is available in the next appendix.

Variable name	Type	Brief description
appl_id	str	Patent application number: 2-digit series code (to include a leading zero for series below 10) followed by a 6-digit serial number; excludes intermediate slash between series code and serial number
predict50_any_ai	int	AI prediction in any of the eight AI technology components based on 50% threshold: 1 if AI in any component, 0 if not AI in all components
predict86_any_ai	int	AI prediction in any of the eight AI technology components based on 86% threshold: 1 if AI in any component, 0 if not AI in all components
Predict93_any_ai	int	AI prediction in any of the eight AI technology components based on 93% threshold: 1 if AI in any component, 0 if not AI in all components
predict50_"X"	int	AI prediction in AI technology Component "X" based on 50% threshold: 1 if AI in component "X", 0 if not AI in component "X"
predict86_"X"	int	AI prediction in AI technology Component "X" based on 86% threshold: 1 if AI in component "X", 0 if not AI in component "X"
predict93_"X"	int	AI prediction in AI technology Component "X" based on 93% threshold: 1 if AI in component "X", 0 if not AI in component "X"
ai_score_"X"	float	AI technology component "X" model score, from 0.0 (not AI in technology component "X") to 1.0 (AI in technology component "X")

Notes: The variable **doc_id** is the primary key in the data table and is formatted to be compatible with the patent and PGPub identifiers in PatentsView. Components "X" are "ml," "evo," "nlp," "speech," "vision," "planning," "kr," and "hardware" for machine learning, evolutionary computation, natural language processing, speech, vision, planning and control, knowledge processing, and AI hardware, respectively.

Appendix E – Linking the AIPD 2023 with PatentsView

PatentsView is a data visualization, dissemination, and analysis platform sponsored by the USPTO's Office of the Chief Economist (www.patentsview.org). It makes USPTO patent and PGPub data available via an application programming interface (API) and bulk download tables. We encourage researchers to use PatentsView to add additional data fields to the AIPD 2023, including information on inventors, assignees, their locations, and technology areas. Although Giczy et al. (2022) provided several use cases in its Online Supplementary materials, the structure of the underlying PatentsView tables have changed since that article was published. Therefore, we update these use cases below to better assist users who are unfamiliar with the newly restructured PatentsView datasets. We also provide additional information about the use cases and add new ones.

General

The PatentsView data tables (see "Datasets" on the PatentsView website) are divided into two sets, one for granted patents and another for PGPubs. Data for granted patents are organized by variable **patent_id**, and PGPubs by variable **pgpub_id**. Both of these variables are strings and do not contain country or kind codes, nor do they include commas or slash characters. The variable **patent_id** is formatted without any leading zeros (for patents prior to U.S. Patent No. 10,000,000), and reissue patents are formatted with "RE" followed by 5 digits. The variable **pgpub_id** is formatted as a 4-digit year followed by 7 digits, to include leading zeros following the year. The AIPD 2023 can be combined with PatentsView data by merging the AIPD 2023 variable **doc_id** with PatentsView variable **patent_id** or **pgpub_id**, as applicable.

PatentsView primarily uses USPTO patent and PGPub full text data (see https://bulkdata.uspto.gov/) and thus reflects as-published data. However, some information such as CPC symbols, are both released as published and updated to current values. Additionally, PatentsView includes entity disambiguation for inventors, assignees, and their locations, and also performs inventor gender attribution.[37]

De-duplicate utility patent documents by application number

A given patent application may correspond to multiple patent document-level observations in the AIPD 2023. For example, an application may be first published as a PGPub 18 months after filing and then subsequently published as a granted patent. To avoid double counting, documents can be de-duplicated by application number, preserving the last published PGPub if the application is still pending or abandoned (e.g., not granted a patent) or the patent if the application was granted. The steps below are expanded beyond what was previously presented

[37] For additional information, see the PatentsView website (https://patentsview.org/), in particular "Methods & Sources" and the Data Dictionaries.

in Giczy et al. (2022). Steps 1 ("Remove withdrawn patents and PGPubs") and 2 ("Accounting for reissue patents") can be ignored for less detailed analyses, as these steps have a very small effect in practice.

Remove withdrawn patents and PGPubs

The first step removes withdrawn patents and PGPubs. While PatentsView includes a variable to identify withdrawn patents, it does not include one for withdrawn PGPubs. However, the most current list can be downloaded from the USPTO website.

- Withdrawn patents numbers: Patent Grant Authority Files at https://www.uspto.gov/patents/search/patent-document-authority-files (direct link: https://www.uspto.gov/sites/default/files/documents/authority.zip)
- Withdrawn PGPubs: Pre-Grant Authority Files at https://www.uspto.gov/patents/search/patent-document-authority-files (direct link: https://www.uspto.gov/sites/default/files/documents/pgpubauthority.zip)

Note that these files need additional processing. In the Authority Files withdrawn documents are identified with a "W" in one column, and patent and PGPub numbers need to be reformatted for compatibility with the AIPD 2023 variable **doc_id** (and with PatentsView).[38]

Accounting for reissue patents

Reissue patents present a challenge since they are given a new application number from their parent patent.[39] Moreover, a reissue patent may be a continuation of another reissue patent, and we should treat this continuation as a separate application. In Giczy et al. (2022), we suggested that researchers may wish to drop all reissue patents because of these complexities. Below we provide a reasonable approach to incorporate reissue patents into the de-duplication process, although some reissues will still be dropped due to lack of necessary or inaccurate data regarding the parent application.

The USPTO Patent Examination Research Dataset (PatEx; Graham et al. 2018)[40] includes application parent-child relationships in the **continuity_parents** data table. These relationships include reissues (REI), continuations (CON), divisionals (DIV), and continuations-in-part (CIP), along with others (where variable **continuity_type** refers how the child relates to the parent).

[38] E.g., numbers include the "US" country codes, and patents are listed as 8-character strings with leading zeros (for reissue patents, the leading zero(s) are after the "RE" characters). See file layout descriptions on the webpage.
[39] See USPTO MPEP 1401.
[40] PatEx is created by the USPTO Office of the Chief Economist from data in the USPTO Patent Examination Data System (PEDS); see documentation available at https://www.uspto.gov/ip-policy/economic-research/research-datasets/patent-examination-research-dataset-public-pair.

The basic approach is to identify the parent application number of the (child) reissue patent and to use the parent application identifier for de-duplication. However, the PatEx continuity data includes all parents of a given application and not just the immediate parent that we are interested in. Hence, we keep only the last-filed parent utility application (excluding national state entries of international applications), which are identified by the highest numbered parent utility application number (i.e., beginning with a number less than 29[41]) since application numbers are assigned sequentially. We then use PatentsView data table **g_application** to get the patent numbers for each parent application (PatEx data table **application_data** could also be used for this step). If the continuity type from the parent to child relationship is "REI" and the parent is a non-reissue utility patent, then the (child) reissue patent's application number is replaced by that of the parent. If, however, the parent to child relationship is "CON," "DIV," or "CIP," then keep the (child) reissue application number as-is, since we consider all other continuation applications to be separate applications (i.e., not de-duplicated with their parents).

The above approach is straightforward in theory, but in practice there are complications due to incomplete or inconsistent data, particularly since the data results in a given application having several patents (non-reissue and/or reissue). Thus, to simplify we suggest keeping only reissue patents where the **continuity_type** is "CON," "DIV," or "CIP," and dropping those with "REI." The resulting dataset will include the first granted patent of an application instead of any subsequent reissue. This simplified approach is summarized in the pseudo-code below.

Pseudo-code (simplified approach):

```
subset the AIPD data to keep if doc_id begins with "RE"

open PatEx data table continuity_parents

   drop if continuity_type == "NST" (since these are PCT
   national stage entry applications)

   by application number, keep the highest numbered utility
   parent application number (i.e., the first two digits less
   than "29," since utility patents are numbered below 29)

   keep if continuity_type == "CON", "DIV", or "CIP"
```

[41] U.S. patent applications are numbered with a 2-digit series code plus a 6-digit number, where utility, plant, and reissue applications begin with series code 01 and above, and design patents use series code 29 (see MPEP 503; note the MPEP does not reflect that series code 18 is currently being used as of the date of publication). Thus, we use series code less than 29, and since a child reissue patent application would be of the same patent type as its parent, we will have only reissue utility applications after merging with the continuity data.

```
drop reissue patents where the filing date is prior to the
parent filing date or the reissue application ID is smaller
than the parent application ID (these observations have data
errors)
```

De-duplicate documents by application number

To create the de-duplicated dataset, we find for each distinct application ID (**appl_id)** the date of the first publication in that application and then keep the last published document for that application. An application may have several PGPubs (to include republications or corrections) and, if granted, a patent. If we include reissue patents that are of continuity type CON, DIV, or CIP (as described above), that patent may be a reissue.[42] However, the data contains inconsistencies, and even if we remove withdrawn patents (as described above) there remains applications that have more than one granted patent. The approach, then, is to identify these applications and create an ID to separate the two. The updated pseudo-code from Giczy et al. (2022) is provided below.

Pseudo-code:

```
merge PatentsView tables to get application filing_dt for each
document:

   copy doc_id as patent_id_if flag_patent==1 and merge on
   patent_id using table g_application

   copy doc_id as pgpub_id if flag_patent==0 and merge on
   pgpub_id using table pg_published_application

drop if doc_id begins with "RE", or identify and keep only
reissue patents that are CON, DIV, or CIP as described in the
simplified approach above

by appl_id, count the number of PGPubs, (regular) patents, and
reissue patents: n_pgpub_appl, n_patent_appl, n_reissue_appl

sort applications by appl_id, filing_dt, and pub_dt; assign a
numbered index i_appl (1… n) for each publication of an
appl_id (i.e., restart index at 1 for each application)

copy appl_id to variable appl_id2

modify appl_id2 for duplicate patent(s) by appending "_i_appl"
of the duplicate to appl_id2; duplicates are identified by:[43]
```

[42] Since the simplified approach, as previously described above, does not include reissue patents of continuity type REI, applications will not have a non-reissue patent and a reissue patent.

[43] The algorithm is based on inspecting the data, e.g., by two tabulation of the by-application PGPub, patent, and reissue patent counts; other data sets may require another approach.

```
    (for multiple patents, no PGPub, no reissues):
    n_patent_appl>1, n_pgpub_appl==0, n_reissue_appl==0 and
    i_appl>1

    (for multiple reissue patents, no PGPub, no regular
    patents): n_patent_appl==0, n_pgpub_appl==0,
    n_reissue_appl>1 and i_appl>1

    (for one patent, one PGPub, one reissue): n_patent_appl==1,
    n_pgpub_appl==1 and n_reissue_appl==1 and doc_id begins
    with "RE" (i.e., change appl_id2 for the reissue patent
    since it is supposed to be a distinct CON, DIV, CIP per
    simplified approach)

    (for all other cases): examine the data and manually
    identify the duplicate

by appl_id2:

    identify the earliest pub_dt among all document and copy
    this date across all observations of appl_id2

    identify the last pub_dt document

keep the last published document of appl_id2
```

Inventors for machine learning patents

Next, we describe how to identify all inventors associated with machine learning patents in the AIPD 2023. The first step is to keep all document level observations in the AIPD 2023 predictions file where **flag_patent** == 1 (i.e., keep all patent documents). In the second step, limit the set of patent document observations to those that are classified as AI in machine learning. The AIPD 2023 provides three different thresholds for predicting AI: 50 percent, 86 percent, and 93 percent. In this use case, we use the 86 percent threshold that balances precision and recall (keep if **predict86_ml** == 1) (more information on these alternative thresholds is provided in the "Extensions and discussion" section and in Appendix C). Researchers may also set their own threshold for determining machine learning by using the **ai_score_ml** variable. Next, merge the machine learning patents with PatentsView inventor information using either the **g_inventor_disambiguated** or **g_inventor_not_disambiguated** data tables on **doc_id** (AIPD 2023) and **patent_id** (PatentsView). The first table, **g_inventor_disambiguated**, collapses the same inventor across patent document records into a single inventor id, while the second table, **g_inventor_not_disambiguated**, provides the raw inventor data (i.e., as printed on the patent).[44] For each inventor, PatentsView contains the first name, last name, and an inventor

[44] More information on the disambiguation process is available at https://patentsview.org/disambiguation

location id. Disambiguated inventor data also contains an attributed gender variable. The revised pseudo-code from Giczy et al. (2022) is provided below.

Pseudo-code:

```
keep if flag_patent == 1

drop if doc_id begins with "RE" (or use dataset from above
where reissue patents are de-duplicated with parent patents)

keep if predict86_ml == 1

      (or keep if ai_score_ml >= threshold)

left merge g_inventor_disambiguated or
g_inventor_not_disambiguated on doc_id (left), patent_id
(right)
```

PatentsView also contains data on the inventor location: city, state (if any), and country. Location data may be disambiguated (**g_location_disambiguated**) or raw as found on the printed patent (**g_location_not_disambiguated**). The data table **location_disambiguated** also includes latitude and longitude coordinates, U.S. counties, and U.S. state and country Federal Information Processing Standards (FIPS) codes, as applicable.

Pseudo-code:

```
if g_inventor_disambiguated used for inventor data:
   merge g_location_disambigated on location_id
else if g_inventor_not_disambiguated used for inventor data:
   merge g_location_not_disambigated on rawlocation_id
```

If researchers wish to analyze inventors by application, as opposed to those on granted patents, then the de-duplicated application file described above will contain a mix of patents and PGPubs. Inventor data for PGPubs may be found in the PatentsView PGPub data tables, which use a "pg_" versus a "g_" prefix (e.g., **pg_inventor_disambiguated** and **pg_location_disambigated**). PGPubs may be identified using variable **flag_patent** == 0. Note that PatentsView disambiguation does not extend across patents and PGPubs, e.g., **location_id** in a patent data table should not be used for PGPubs and vice versa.

Owners of machine learning patents

Patent owners are referred to as "assignees." Similar to inventors, PatentsView contains data tables for assignees (e.g., **g_assignee_disambiguated** or **g_assignee_not_disambiguated**), and assignees may be added in the manner discussed for inventors above. Likewise, the same location data tables as discussed above for inventors may also be used with assignees.

However, effective September 16, 2012, an assignee may apply for a patent (before this date only inventors, in general, may be the applicant).[45] Thus, if assignee data is missing after merging assignee data tables, then applicant data might fill in the blanks once assignee-applicants are separated from inventor-applicants. Since inventors can only be persons and not organizations, all organizations may be reasonably assumed to be assignees. Persons may be assignees, and we may identify them by PatentsView variable **applicant_type** == "applicant" and **applicant_authority** == "assignee" or "obligated-assignee". Only raw (not-disambiguated) applicant data is available with PatentsView. The pseudo code for identifying the owners of machine learning patents is below.

Pseudo-code:

```
keep if flag_patent == 1

drop if doc_id begins with "RE" (or use dataset from above
where reissue patents are de-duplicated with parent patents)

keep if predict86_ml == 1

      (or keep if ai_score_ml >= threshold)

left merge g_assignee_disambiguated or
g_assignee_not_disambiguated on doc_id (left), patent_id
(right)

open g_applicant_not_disambiguated:

   tag applicant if not missing(raw_applicant_organization)

   tag applicant if not missing(raw_applicant_name_last) and
   applicant_type=="applicant" and
   (applicant_authority=="assignee" or "obligated-assignee")

   keep tagged applicants

left merge tagged applicants on doc_id (left), patent_id
(right) if missing assignee data
```

Number of patents in a given CPC subclass

As a final use case, we describe how to tabulate the number of patents in each Cooperative Patent Classification (CPC) subclass. Beginning with the AIPD 2023, keep patents, remove reissue patents or use the de-duplicated data described above, and identify all patents determined to be AI at the 86 percent threshold (see the use cases above for more information on these steps). Finally, merge the set of AI patents with the PatentsView current CPC table (**g_cpc_current**)

[45] See MPEP 605.

using **doc_id** from the AIPD 2023 and **patent_id** from **g_cpc_current**. The revised pseudo-code from Giczy et al. (2022) is provided below.

Pseudo-code:

```
keep if flag_patent == 1

drop if doc_id begins with "RE" (or use dataset from above
where CON, DIV, CIP reissue patents are included))

keep if predict86_any_ai == 1

      (or keep if any ai_score_[component] >= threshold)

left merge g_cpc_current.tsv on doc_id (left), patent_id
(right)

by cpc_subclass: count number of observations
```

The code above includes all CPC subclasses regardless of being CPC First, CPC Inventive, or CPC Additional (see MPEP 905 for additional information on the CPC). To use only CPC First, keep only observations where **cpc_type** == "inventional" and **cpc_sequence** == 0. To use all inventive CPCs, keep if **cpc_type** == "inventional".