March 2022

Perspective
EXPERT INSIGHTS ON A TIMELY POLICY ISSUE

SHAWN MCKAY, GAVIN S. HARTNETT, BRUCE HELD

Airline Security Through Artificial Intelligence

How the Transportation Security Administration Can Use Machine Learning to Improve the Electronic Baggage Screening Program

Artificial intelligence (AI) and machine-learning (ML) technologies have been integrated into the structure of Americans' lives for decades. From agriculture to medicine, shopping, and social networking—to name just a few areas—AI and ML are invisibly and almost imperceptibly changing the nature and quality of individual and societal functioning. And although leveraging AI and ML capabilities started slowly, it has been growing in proportion to the expanding understanding of the science and the exponential growth of the necessary information infrastructure. The growth of AI and ML capabilities means that AI and ML will gain application into an ever–rapidly increasing number of tasks to improve performance and efficiency. This Perspective uses the Transportation Security Administration's (TSA's) Electronic Baggage

HSOAC
HOMELAND SECURITY OPERATIONAL ANALYSIS CENTER
An FFRDC operated by the RAND Corporation under contract with DHS

Screening Program (EBSP) as a case study to highlight how to think about the integration of AI and ML into existing programs.[1] In the case of EBSP, as will be discussed, AI and ML are not a single technology program meant to achieve some set of requirements on a specific schedule. Rather, the approaches discussed are meant to demonstrate how AI and ML can be iteratively integrated to positively and significantly improve a program over time to improve safety, cost, and travel convenience to the traveling public. The general approaches discussed in this Perspective should have relevance for many other federal programs.

Abbreviations

3D	three dimensional
AI	artificial intelligence
AUC	area under the curve
CT	computed tomography
DHS	U.S. Department of Homeland Security
DL	deep learning
EBSP	Electronic Baggage Screening Program
EDS	explosive detection system
ML	machine learning
ROC	receiver operating characteristic curve
S&T	Science and Technology Directorate
TSA	Transportation Security Administration
TSO	transportation security officer
VVT&E	validation, verification, testing, and evaluation
XRD	X-ray diffraction
Z_{eff}	effective atomic number

Machine Learning: Achieving Artificial Intelligence Through Data

The field of AI aims to "design, build, and experiment with computational systems that perform tasks commonly viewed as intelligent" (Poole, Mackworth, and Goebel, 1997, p. 2). ML is a subcomponent of AI in which intelligence is not explicitly programmed in the algorithms but is automatically learned using data (T. Mitchell, 1997). The basic idea behind ML is to allow a machine (software) to "program itself" through the extraction of statistical relationships from data sets. More concretely put, the basic operation of an ML algorithm is to determine the statistical relationships between input and output variables by processing a validated set of examples called *training data*. ML algorithms generally perform better when more training data are available, and the current technology requires extremely large data sets for realistic applications.

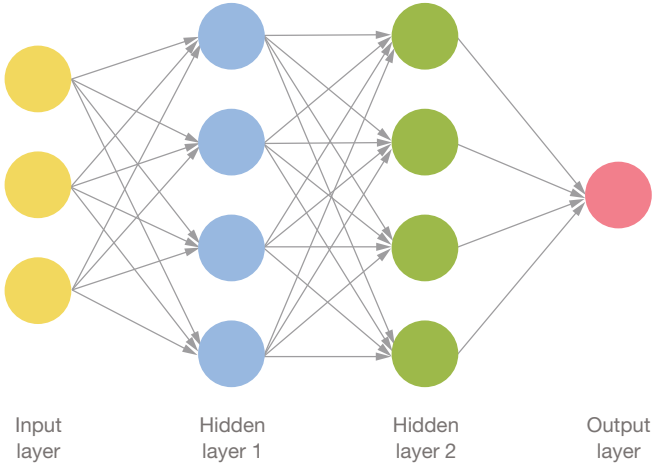
ML as a field has existed for decades, but it has only recently enjoyed such significant advances and rapid growth (to the point at which AI and ML are often conflated as equivalent concepts).[2] Two oft-cited reasons for this growth are the increase in the availability of large data sets and the vast improvements in the computational resources available to train ML models.

Clearly, an AI capability that relies on extracting information from data should improve as the amount and quality of data increase, and recent years have seen an explosion in the quality and

amount of data across all domains and modalities (e.g., images, videos, text, audio). Less obvious is the fact that an approach to AI that extracts information from data requires flexible statistical models to encode the learned information from these data sets. Most of the recent success of ML in solving complex tasks in such areas as computer vision, natural-language processing, and game playing has involved the use of artificial neural networks (see Figure 1). Artificial neural networks are modular in that they consist of "layers" that can be stacked onto one another. As more layers are added, the model becomes "deeper"; the more specialized term *deep learning* (DL) is used to refer to ML approaches that employ such networks. The many layers of a deep neural network allow for more-complex relationships to be learned from the data than are possible with other approaches. However, the addition of more layers generally comes at a cost—much greater computational resources and more-detailed data for training than simpler models.

FIGURE 1

A Simple Artificial Neural Network



SOURCE: Abu and Mohamad, 2020, p. 165 (CC BY 4.0).

Recently, both data and computational resources have become more widely available, increasing the number of areas in which ML can lead to improvements over the status quo. Also, although coding a neural network–based approach was previously a very difficult software engineering problem, the widespread success of DL has led to a very active open-source software ecosystem.[3] As a result of these advances, ML can outperform humans on a large and growing list of tasks, which is attributable to the demonstrated ability of deep neural networks to vastly outperform competing approaches (including humans) in computer vision competitions, in which the goal is to recognize the types and locations of objects of interest within a digital image (LeCun, Bengio, and Hinton, 2015).

As noted earlier, the basic operation of an ML algorithm is to determine a statistical relationship between a validated set of example inputs representing problem statements and a set of corresponding outputs representing problem solutions. The ML algorithm produces a program that captures and can "execute" the statistical relationship derived from the training data—that is, given a set of inputs similar to those in the training set, it will compute an appropriate set of outputs in line with the derived

statistical relationship. That program is called the *application*.

In summary, given a set of training data, an ML algorithm will produce an application to solve a problem represented by the training data. Technology that relies on extracting information from data, like ML does, should improve as the amount of data increases, but modern approaches to ML also require flexible statistical models to encode the learned information from these data sets, as well as significant computational resources.

The Potential of Machine Learning in Airport Baggage Screening

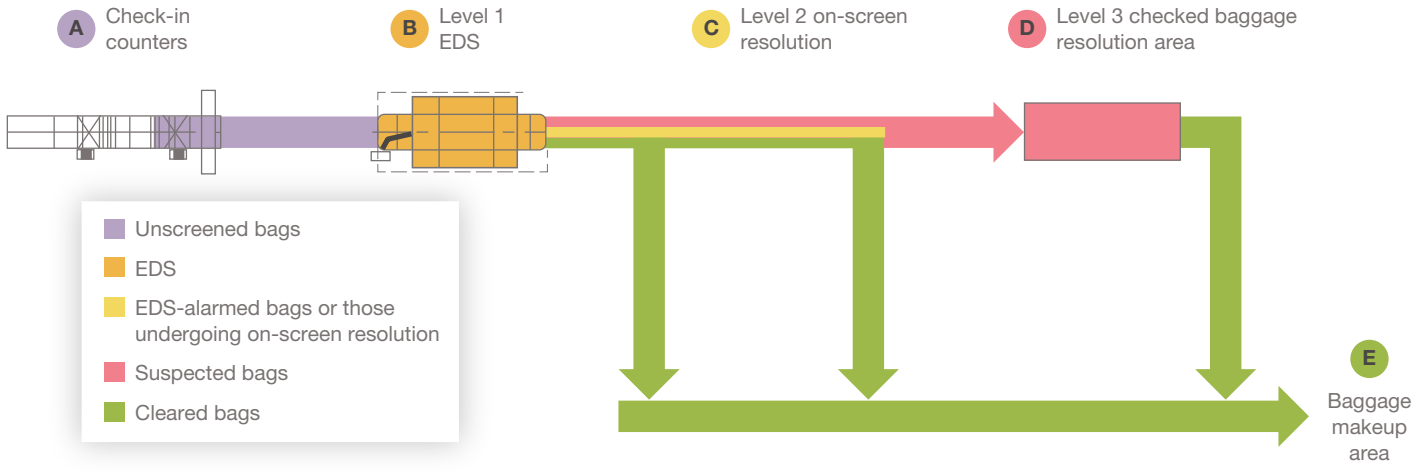
The Aviation and Transportation Security Act (Pub. L. 107-71, 2001) requires that TSA screen 100 percent of checked baggage at airports. To date, there have been no successful attacks through checked baggage since the Aviation and Transportation Security Act was passed, but threat actors are constantly looking for new ways to attack commercial airliners, and checked baggage remains a potential vector. To continue deterring and defeating new threats, TSA will continue to improve EBSP, and, given that much of the TSA baggage screening process involves the analysis of digital imagery, it is possible that computer vision methods could be used to improve the safety of air travel and the efficiency of the screening process.

The current TSA checked baggage screening process consists of three levels (see Figure 2). Level 1 screening is fully automated. During this step, explosive detection systems (EDSs) use X-rays and computed tomography (CT) technology to identify potential threat material. Two data outputs are produced at level 1 screening: a binary threat indicator (either no threat material or potential threat material) and a three-dimensional (3D) CT image of the baggage. To ensure high confidence that baggage containing threat material will be identified as such, much benign baggage is flagged as containing potentially threatening material. Flagged baggage is sent to level 2 screening, which consists of an inspection of the 3D CT image by a trained transportation security officer (TSO). The TSO looks for indicators either that the baggage is benign (i.e., is not a threat to the aircraft onto which it will be loaded) or that it could be a threat. If the piece of baggage cannot be cleared by level 2 screening, it proceeds to level 3. This final screening step can include examination of the 3D CT image, testing for explosive residue on and in the baggage, and, if needed, a physical inspection of the baggage's contents.

A simple-to-define yet difficult-to-solve task—namely, determining whether a bag is safe—is repeated hundreds of thousands of times each day, and the process involves analyzing complex image data (CT scans) in order to produce a binary outcome. This process is naturally suited for a variety of ML approaches because they typically excel at problems that have large amounts of labeled data.

FIGURE 2

Generic In-Line Checked Baggage Inspection System



SOURCE: TSA, 2020, p. 2-3.

In this Perspective, we review recent advances in ML and identify several algorithms and approaches that could be applied (either as is or with a reasonable level of modification) to the baggage screening process. Our goal is to facilitate the adoption of existing ML technologies that could improve the safety of air travel. Of course, there is a large divide between what is merely plausible and what is achievable, and much work will be needed to implement working solutions that pass TSA's stringent requirements. We conclude with a discussion of these challenges and the next steps that could be taken.

Threat-Detection Methods: Anomaly Detection and Binary Classification

ML has been successfully applied in a variety of areas. In this Perspective, we focus mainly on two closely related problems: anomaly detection and binary classification. Anomaly detection identifies instances that are nonstandard or not as expected. A classic example of anomaly detection is the problem of detecting credit card fraud: The overwhelming majority of transactions are not fraudulent, but the tremendous diversity of fraudulent transactions is ever-changing as criminals modify their techniques and their purchasing patterns evolve. The problem of identifying dangerous checked baggage can also

be recognized as an anomaly-detection problem; the majority of bags are benign, and the space of all possible dangerous baggage configurations is unknown.

Binary classification identifies instances as one thing or another. For example, an item might be classified as animate or inanimate. Anomaly-detection problems can be shaped to resemble binary classification algorithms, in which the two labels are *not anomaly* and *anomaly*. However, there can be many drawbacks to treating the problem this way:

- First, the anomalies are, by definition, quite rare. Special care must be taken to ensure that the model does not simply learn to always predict that something will not be an anomaly because, by doing so, the model could achieve a very low error rate simply by virtue of the fact that anomalies are rare.
- Second, treating all anomalies as being members of the same class assumes that they are all similar. This might be a valid assumption with some problems, but often, anomalies are truly anomalous in that they are distinct both from usual instances and from each other. In this case, an additional problem is that the previously observed anomalies in the training data set are not representative of new anomalies that will be encountered. For these reasons, it is preferable to not treat the problem as one of binary classification but instead treat the baggage screening task as one of anomaly detection.[4] However, before examining anomaly-detection ML possibilities for EBSP, we first look at a binary classification method using data that could be collected from current EDS technology.

Machine Learning Might Have Some Near-Term Applications for Binary Classifications

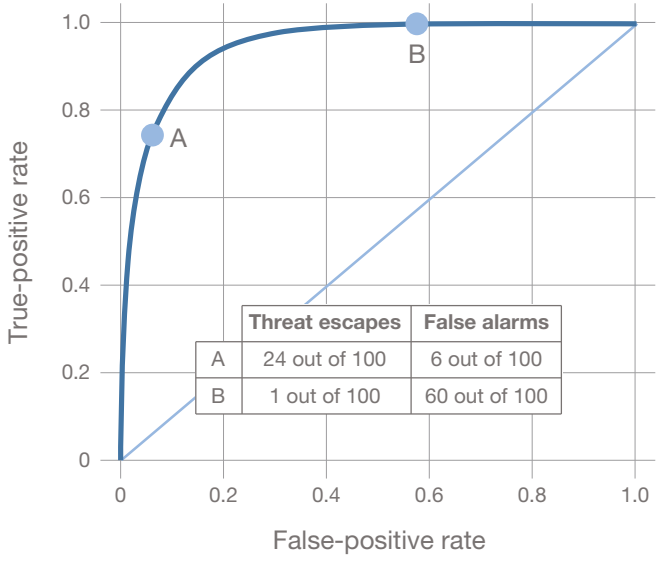
Some developments already suggest the possibility of leveraging ML for baggage screening by using X-ray data that could be available from existing level 1 EDS scanning capabilities. One paper described a method for using such a labeled data set to "teach" an ML-enabled algorithm to detect a chemical threat in a checked bag using X-ray scanning (Morris, Chien, and Goodman, 2018). Using a common dual-view EDS,[5] the authors of that paper created a training data set with a total of 4,760 scans, of which 1,960 were single-chemical threats. These researchers followed a classification ML approach by using a labeled data set, and they addressed the problem associated with detecting rare threats by using a balanced data set consisting of roughly 60-percent nonthreat and 40-percent threat material. For their model, the authors employed a convolutional neural network, a particular type of neural network architecture that has driven much of the recent progress in image recognition.[6]

The model performance was characterized by the receiver operating characteristic curve (ROC), a common metric for performance evaluation of

classification models.[7] Figure 3 depicts a notional ROC.[8] The true-positive rate is the probability that the algorithm will find a threat when one exists, and the false-positive rate is the probability that a threat will be flagged when a threat does not exist. An ideal algorithm would find all actual threats and would flag nothing that was not a threat, but this is rarely achievable in practice, and trade-offs are typically necessary. For example, the detection threshold can be lowered to reduce the number of missed threats (err on the side of caution), but this comes at the expense of predicting nonexistent threats (false positives). Although the ROC accurately depicts this fundamental trade-off, it is still convenient to have a single number to evaluate a model's performance. This can be achieved by calculating the AUC. In this application, the AUC was 0.95, which is considered excellent by general academic standards (Mandrekar, 2010), but not for TSA's purposes.[9] For the checked baggage application, an even higher AUC will be needed to yield both low false-positive rates and a very high probability that actual threats will not be misidentified as benign. To achieve this level of performance, additional development is necessary, and it is not clear whether positive results could be achieved before other approaches—such as the integration of additional scanning technologies—mature.

FIGURE 3

Model Accuracy Characterized by Receiver Operating Characteristic Curve



	Threat escapes	False alarms
A	24 out of 100	6 out of 100
B	1 out of 100	60 out of 100

SOURCE: Adapted from Morris, Chien, and Goodman, 2018.

NOTE: Points A and B correspond to two detection thresholds and illustrate the trade-off calculus that imperfect detection algorithms require. The diagonal line corresponds to a sequence of models based on simple random guessing; worse-than-random models lie to the right of this line, and better-than-random models lie to the left of this line. The notational ROC displayed here is entirely to the left of the dashed line, indicating that every model along the curve is better than random guessing.

The Morris, Chien, and Goodman example showcases how balanced, labeled data sets can be created to train an ML algorithm. This might be relevant for TSA in that its operations already create large amounts of data on a daily basis. Today's EDS scanners, for example, determine density and effective atomic number (Z_{eff}) for each voxel[10] in a piece of baggage, and this information is used to characterize the material in the voxel. But a simple algorithm can be used for this characterization, so ML technology is not necessary. But if an additional

data type were made available, ML technology could be trained to reduce false-positive detections with a labeled data set using three different material indicators. X-ray attenuation is an example of a potential additional datum, but collecting the data would require some modification of the EDS machines to assess it. An additional modification would be required to provide a means of outputting all the collected data for analysis. In addition, using the collected data to create a large, balanced data set for ML training would be challenging because of the near-total lack of identified actual threats in checked baggage. Creating the "threat" side of the balanced data set would require the generation of simulated threats that can be safely scanned in EDS machines and tagged as threats. Moreover, an acceptable ROC (or AUC) would need to be defined, and this would affect the required size of the training data set. Because we expect these requirements to be very high, the threat data set would need to be large.

As just discussed, the ability to leverage ML to get more out of existing EDSs shows some potential now, but additional development, including the creation of a usable data set, will be necessary to achieve the required level of performance, suggesting that a longer-term perspective might be required or more promising. Moreover, the advantages of ML will increase as additional and different scanning technologies are introduced. These would provide richer and more-diverse training data sets for ML tools to use (Mandrekar, 2010).

These longer-term potential applications of ML in EBSP are discussed next.

Deep Learning Methods for Analyzing X-Ray and Computed Tomography Scan Data

Recent years have seen significant progress in using deep neural networks[11] for image recognition. For example, AI now outperforms radiology technicians in identifying edges of brain tumors (Bondell, 2020), reading human handwriting (Captricity, 2018), diagnosing pneumonia from chest X-rays (Kundu et al., 2021), and diagnosing common fungal infections ("AI vs. Doctors," 2017). Here, we review in detail two recent methods for using deep neural networks to process X-ray and CT data. Both methods treat the problem of identifying threats as one of binary classification. However, these approaches could likely be modified to handle the related problem of anomaly detection, which is more appropriate for a baggage screening application.[12]

Many successful ML applications identify anomalies in CT images. Parallels can be drawn from the field of medical imaging, which has also recently begun to investigate whether ML techniques can improve the state of the art. The motivation for doing so is that, as populations age in many countries and as more and more medical scanning techniques become available, the requirement for radiological imaging analysis has grown, increasing the needed number of radiologists who must manually

read the images or imposing greater demands on radiologists already in the workforce (Brady, 2017).[13] As a result of the added requirements for radiologists, a considerable amount of work has been undertaken to develop ML algorithms that can use CT images to identify diseases and help reduce general radiological analysis error rates—which range from 3 to 5 percent with human radiologists—and the overall radiologist workload (Brady, 2017).[14]

One recent paper describes how ML methods could be applied to 3D CT chest scans to identify several types of diseases (Draelos et al., 2021). This is because 3D CT scans are an incredibly rich data format. In the context of chest scans, for example, each image might include different organ types and several of hundreds of different abnormalities. DL excels at problems like this, with most approaches requiring a high-quality labeled training data set. In this case, the labels would correspond to the disease type, but, unfortunately, such a data set did not exist. To surmount this obstacle, the researchers developed a method to automatically create the required training data set by extracting the labels from previously created radiologist reports. This allowed the ML algorithm to find the underlying relationships in the data that could be used to distinguish different diseases from benign scans. Although the models created by the researchers had decent performance overall, performance did vary by disease class.[15]

Similarly, as air traffic demand increases, the demands on existing TSOs doing level 2 review of CT images will increase, or more TSOs will need to be hired and trained. As described earlier, the current level 2 screening practice by TSA relies on analysis of CT scans by a TSO. The demonstrated utility of ML for CT scan analysis in the medical context suggests that similar methods could be useful for baggage screening as well.

Outlook: Building Toward an Artificial Intelligence–Assisted Baggage Screening Process

The work using ML to improve interpretation of medical images suggests possibilities for EBSP. This section describes two. We start in the medium term (the next five to ten years) with an anomaly-detection example that could use existing EDS image data to streamline level 2 screening.

The Medium Term (Five to Ten Years)

There is an opportunity for TSA to achieve cost efficiencies in level 2 screening by using ML to reduce the number of CT scans that TSOs need to review manually. The earlier examples demonstrate that deep neural networks can be used to analyze X-ray or CT imagery, provided that enough data of sufficient quality are available. There is no obvious existing labeled data source for TSA screening, which poses an immediate practical problem. This problem was also faced by the researchers investigating the use of DL to identify chest abnormalities

from CT scans. In that case, the researchers made clever use of existing medical reports to extract labels from assessments made by medical professionals. A similar avenue is available in the baggage screening problem because of the heavy reliance on human judgment in level 2 and 3 screening and the generation of a substantial amount of data created in level 1 screening.

The additional screening by a human TSO in level 2 presents an opportunity to augment these data and allow ML-based methods to improve the existing screening process. For example, a common item with a chemical characteristic that might alarm the EDS, such as a bottle of spirits, will often be recognized by the TSO, who will also note the lack of other items that would suggest the presence of a disguised bomb. In such circumstances, the TSO is not labeling the baggage or its contents, just assessing it as normal (i.e., not an anomaly). This situation is likely to happen thousands of times a day for a variety of common objects found in passenger luggage. The TSO will send baggage to level 3 only when doubts about the benign nature of the baggage cannot be adequately assessed. Because suspicion can be a point anomaly (the object looks like a bomb) or a context anomaly (9-V batteries are not typically used in laptops), a TSO's assessment can also be used to augment the label provided by the level 1 EDS.

The key advantage of this strategy is that the data are already available—they just need to be stored in the appropriate format. Additionally, different levels of implementation can provide confidence and flexibility. For example, after some initial period of training, statistics on the model's performance relative to the TSOs' can be collected, and thresholds on the confidence of the AI determinations can be established. So, for example, a threshold for an actual implementation decision might be something like, "When the AI and TSO assessments of 'send to level 3' agree 95 percent of the time for those bags for which the AI has indicated high confidence in the assessment, those bags can be sent to level 3 without intermediate assessment by a TSO." A similar threshold could be set for those bags destined to be sent to the plane. The implication here is that, once these thresholds are met and the ML screening is implemented, the TSO workload will be reduced. Moreover, because the ML model can be expected to improve over time as more data are collected, the percentage of baggage that does not need human TSO assessment should also increase over time.

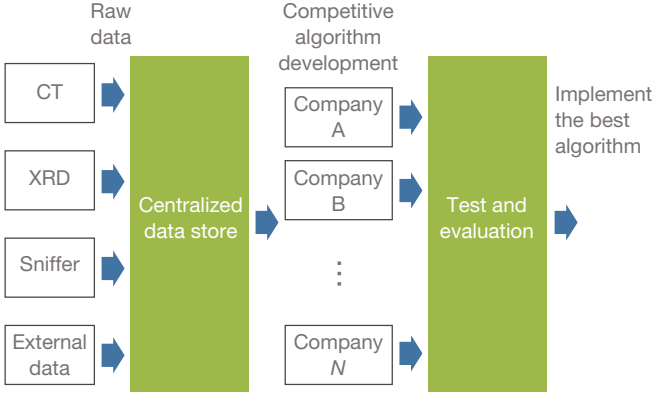
The Long Term (Beyond Ten Years)

The potential for ML integration becomes much greater if additional scanning and data collection methods are introduced into the checked baggage screening process. This represents a more involved overhaul of the current process and is therefore more appropriate for longer time frames than those in the medium-term proposal.

As new data-producing technologies are introduced into the scanning process, ML will likely be needed to take full advantage of the information that will become available.[16] For example, neutron scans are capable of penetrating metallic containers that cause EDS machines to alarm because they cannot assess all the contents of a piece of baggage. But they also provide additional information for characterizing the materials in checked baggage. As another example, X-ray diffraction (XRD) technology complements existing CT scans by utilizing the scatter of the X-rays to differently characterize the materials through which the X-rays are passing. If either or both of these scanning technologies are integrated into baggage screening, the assessment requirements are likely to be too onerous for human cognition in the required time frames. Additionally, future automated chemical-sensing technologies could provide more information to help distinguish potential threats, but the nature of trace-detection results could require sophisticated, real-time analysis to unwind all the chemical signatures likely on checked baggage. Nontechnical data, such as intelligence and passenger profile information, could also provide information to tune risk thresholds appropriately, and this information could be integrated into an overall assessment of the threat that a piece of baggage poses.

A future ML system based on analyzing many disparate types of data will, therefore, need to involve some form of data or sensor fusion, which is the process of integrating the different data sources into a more usable form for the algorithm. The development of such a system will require both changes to existing information technology infrastructure and progress in ML model design. A new database structure will be required to store all the different types of data. This database should be accessible both in situ at the point of operation for use in threat detection and remotely to assist in algorithm development, off-line data analysis, and maintenance. Standards in data formatting will be needed to ensure that the data can be integrated into a data set suitable for ML development. These data standards will need to be built into acquisition requirements and be flexible for future adjustments to facilitate needed iterations throughout the life cycle of the scanning technology. Once the data are fused in an acceptable format, ML algorithm development can begin. Figure 4 illustrates one possible

FIGURE 4
Machine-Learning Sensor Fusion Concept



high-level design for a future data fusion and ML development architecture.

Data fusion and competitive ML development are the two main components needed to ensure efficient and effective development of threat-detection ML algorithms. These two components require both new sensor technology acquisition requirements and a new acquisition endeavor in ML algorithm development. The new acquisition requirements for detection technology (e.g., CT, XRD) are needed to ensure that sensor data can be used for ML development. It is unlikely that a single manufacturer will produce all the different types of chemical and scanning tools of the future. Hence, to fuse data from multiple sources and leverage ML capabilities to interpret the data, sensor acquisition requirements should ensure that data are accessible in real time, through common interfaces, and in common formats. But this will require a change in TSA acquisition philosophy. Currently, scanning tools (such as the EDS) acquired by TSA come with vendor-developed, proprietary algorithms for explosive sensing. In other words, the EDS vendors develop the algorithm that produces the threat alarm, and TSA ensures only that the threat alarm performance meets its operational requirements. Because it is unlikely that a single manufacturer will produce all the different types of chemical and scanning tools of the future, it will be important for a process to be developed by which data from multiple detection technologies can be integrated and then leveraged by advanced ML algorithms (L. Mitchell, 2017).[17] We suggest that, to facilitate this data integration, TSA adopt, develop, and require open architectures for its checked baggage scanners so that the raw data of each sensor are accessible.[18] In addition, other data sources that can likely be repurposed for use in checked baggage threat detection are represented as the "external data" box in Figure 4. One example is passenger profile information.

Second, we recommend that TSA facilitate the development of high-performing ML algorithms through a competitive source selection process. As discussed earlier, traditional acquisition techniques depend on the sensor manufacturer to develop the algorithms. This approach becomes highly ineffective because the combined sensor data are likely to provide richer, and hence more-complicated, information to distinguish a threat. Data fusion and competitive ML development decouple this dependence on the sensor manufacturer for algorithm development and open algorithm development to broader competition. In several documented examples, competitive ML algorithm development has yielded not only better results but also faster development iterations (Papers with Code, undated). A test and evaluation organization will need to be involved in a source selection process that identifies the best-performing ML algorithm among the competing entities.

Challenges

Thus far, we have taken a rather optimistic perspective and argued that the existence of successful ML methods for addressing problems analogous to those in the TSA baggage screening process could be adapted to improve the performance and efficiency of the existing baggage screening process. However, AI is not magic, and the development of large-scale ML platforms will take both time and money. Moreover, the use of ML can introduce new issues and failure modes that must be carefully mapped out and understood for safety-critical applications, such as airport screenings. In this section, we discuss some key challenges and novel issues raised by ML. And it is worth noting that these challenges will apply to a broader range of applications for which ML is being considered, not just in baggage screening.

Class Imbalance

The defining characteristic of ML algorithms is that they produce a model by extracting information from data. As a result, problems with rich data sets tend to be amenable to ML solutions. The problem of screening checked baggage for airline security is a bit paradoxical in that there is both an abundance and a scarcity of data because of the extreme class imbalance. The overwhelming majority of screened bags are benign and do not contain banned substances. If the problem is modeled as a binary classification problem, in which each bag is assigned one of two labels (such as *no threat material* and *potential threat material*), the training data will be dominated by the first type of label. Class imbalance can also occur in many other ways of framing the problem. For example, if the task is to identify every object within the X-ray scan of the baggage, the model will have to choose among many thousands of different classes. Some of these will be quite common (for example, clothes, shampoo bottle, or boots), while others will be extremely rare. Crucially, there will be comparatively few real-world positive examples of banned substances, and it is probably not too difficult for a would-be terrorist or smuggler to devise a never-before-seen contraption or concealment method.

Although class imbalance is a difficult problem, many methods and approaches have been developed to handle it. For example, one common approach is to use resampling techniques to either oversample the rare classes (e.g., bombs) or undersample the common classes (e.g., clothes, shampoo bottles). Another approach is to achieve a similar effect through reweighting such that more importance is placed on the rare class during training. In the current case, the rare class happens to be the one about which we care more, so so-called cost-sensitive methods might be appropriate (Elkan, 2001).

Another possible solution to the class imbalance problem is to model the problem as one of anomaly detection, as mentioned earlier. Anomaly detection is similar to binary classification in that the basic problem is to classify data as belonging to

one of two classes (anomaly, not anomaly), but it is designed to handle cases in which there are very few known anomalies, or perhaps each anomaly is different in a hard-to-categorize way. For example, an anomaly-detection approach to baggage screening would involve assessing whether a given bag looked different from all the other benign bags rather than comparing it to previously flagged bags.

Privacy

A key ingredient in the development of ML models is training data. The richer and more plentiful the data, the better the ML model will perform. However, there are also privacy considerations associated with these data, and, as more and more data sources are incorporated into the model, these considerations will grow more complex and consequential. For example, in order to incorporate the X-ray or CT scan ML models discussed earlier, the existing data created during the baggage screening process will have to be saved so that the data can be used as training data for the models. The data could be tagged with metadata, which might include information that could identify the traveler. Care will have to be taken to avoid abuse of this information—by, for example, allowing TSA employees to learn about the contents of a given person's baggage. This same basic privacy concern is present with the current system, but it will become exacerbated as information is stored in large databases accessible by many people.

Cybersecurity

AI and ML development and deployment to airports will require greater network connectivity than exists today, and this will introduce new cyberthreats to manage. For example, the data needed to train the ML models will need to be extracted from level 1 and level 2 processes across airports and centrally stored. Once trained, the ML models will need to be periodically updated and pushed to airport locations, during which process they could be potentially intercepted or altered by malicious actors. Additionally, the reliance on ML systems exposes a new kind of vulnerability called *adversarial examples*, which are inputs to ML systems that have been specifically altered in order to fool the ML system while often appearing unaltered to a human inspector (Gilmer et al., 2018). These are likely just a few of the cyber risks that leveraging ML capabilities will pose. Unfortunately, the baggage screening problem is explicitly adversarial: The whole point of the baggage screening is to prevent adversaries from sneaking banned substances and materials onto planes. Therefore, any would-be ML-based method for improving the baggage screening process must address the risk of these attacks.

While acknowledging and addressing the additional risk of adversarial attacks that accompany reliance on ML methods, developers should not overestimate the extent of that risk. Many attacks require an adversary to have access to a defender's (in this case, TSA's) computer systems, in which case there are probably far worse things the attacker

could do than simply perturb pixel values. Adversarial attacks have also been shown to be somewhat brittle in some real-world application areas, such as computer vision tasks involved in developing self-driving cars (Lu et al., 2017), although many defenses and mitigating practices are available to help deal with the risk of adversarial attacks (Hartnett, Lohn, and Sedlack, 2019).

Machine-Learning Test and Evaluation

Currently, there are no widely accepted and systematic methods for doing validation, verification, testing, and evaluation (VVT&E) on the type of black-box applications produced by statistical ML algorithms. This is an area that will require further development because VVT&E is so connected to confidence in the capabilities of the technology and is necessary for ML model assessment, comparison, and selection.

This capability gap introduces some issues that will need to be addressed:

- First, U.S. Department of Homeland Security (DHS) acquisition processes will need to evolve to accommodate new and, potentially, evolving testing procedures.
- Another source of risk is using such systems in any kind of safety-critical application. This can be partially addressed by proceeding incrementally, carefully monitoring the algorithm's decisions, and erring on the side of caution in designing and testing the system's sensitivity thresholds.

Incremental and Iterative Development of Artificial Intelligence

In thinking about a future AI-assisted baggage screening process, we note that progress in AI has historically been incremental—and we see no reason that this trend would not continue. The reason we list this as a challenge is that initial efforts in ML development might produce unsatisfactory results; however, there are efficient ways to improve ML performance iteratively. As research produces better models and algorithms, and as data and computational resources become more plentiful, AI models can be expected to improve over time. This observation is relevant to TSA because it motivates an iterative product development, acquisition, and maintenance strategy that best takes advantage of the likely cycle of improvement in model performance. It also motivates an initial investment in the computational infrastructure and data storage capabilities required for ML-based solutions because these will continue to be used by future iterations of ML technologies.

To illustrate this point, we briefly highlight the accomplishments produced from the ImageNet initiative, which showcases the iterative nature of ML algorithm development, the power of open competition, and the utility of a standard database for algorithm development. ImageNet is an open-source

database containing more than 14 million labeled images (Figure 5).

In the past decade, researchers in both academia and industry have competed to use the ImageNet database to develop leading ML algorithms for computer vision tasks (Devopedia, 2021). Figure 6 shows how ML algorithm accuracy in this area has greatly improved in the past decade. Accuracy is measured by the percentage of time the model's top guess matches the image label, known as *top 1 accuracy*. At the beginning of the ImageNet Challenge, the top 1 accuracy of the best models was around 50 percent, and it has since grown to slightly over 90 percent.

The ImageNet database and challenge initiative illustrates many principles with ML development that are relevant for TSA. ML model development requires iterations, sometimes spanning years. Establishing a standard, accessible database provides a means of streamlining subsequent iterations. Also, incorporating a competitive aspect into the ML

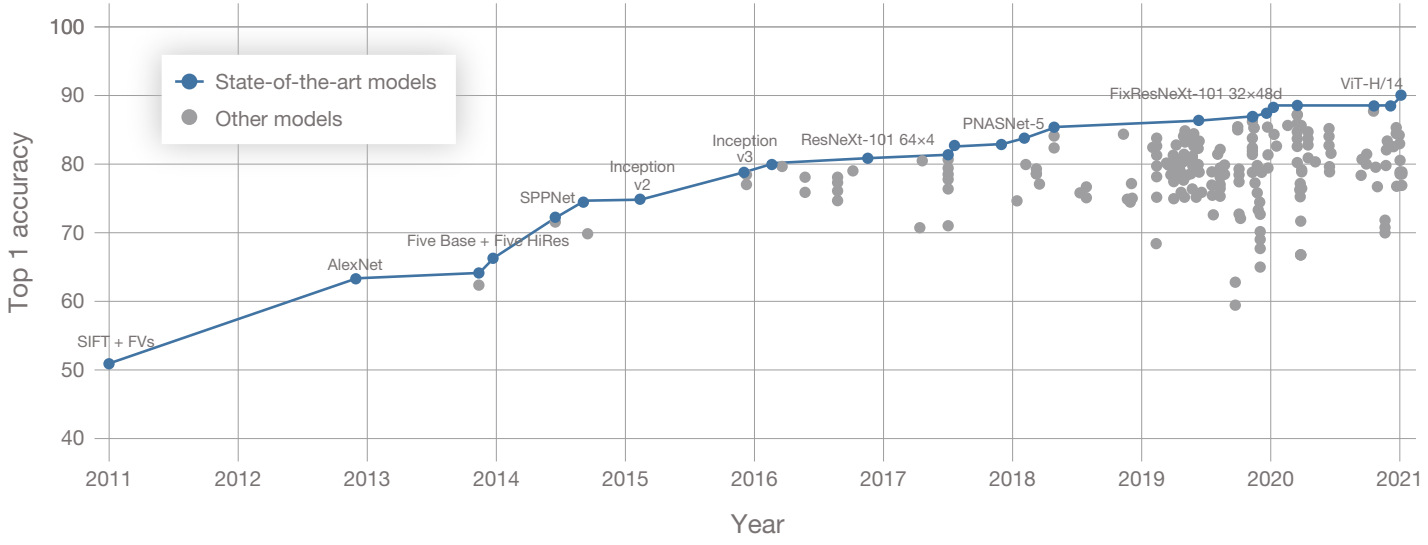
FIGURE 5

Examples of ImageNet Images

SOURCE: © 2009 IEEE. Reprinted, with permission, from Deng et al., 2009, p. 249.

FIGURE 6

Machine-Learning Algorithm Performance Improvement with the ImageNet Database



SOURCE: Papers with Code, undated (CC-BY-SA 4.0).

NOTE: The points on the line are ML programs that have performed best at the computer vision task over time.

algorithm development process provides additional advantages:

- First, there is a host of different ML algorithms and ways to parameterize them, making it unlikely that a single group of developers will identify a top-performing model (Burkov, 2019).
- Second, the more eyes are on the data, the more likely it is that critical issues will be uncovered and addressed.

Along these lines, several such issues have also been uncovered from ImageNet participants. For example, it has been shown that algorithms tend to have an overreliance on color, texture, and background cues (Greene, 2019). The identification of these issues points to where further research and development are needed.

Conclusion

The past decade has seen dramatic improvements in AI's ability to solve complex problems across a wide variety of domains. Most of the progress has come from ML, in which the general goal is to develop a

model tailored for solving a given task or problem by learning from large amounts of data, rather than relying on more-traditional AI systems in which the intelligence is more directly hand-coded into the software. In particular, the large amounts of data that are now readily available, combined with improvements in computational resources, have made it possible to train deep neural networks to successfully solve many difficult problems—in some cases, better than humans can. The obvious promise is the potential to address many challenging tasks more effectively and efficiently than can be done through human perception and cognition.

In this Perspective, we have suggested that there is an opportunity for these same methods to be used to improve TSA's checked baggage screening process. By reviewing recent ML successes in related problems, such as medical imaging and anomaly detection, we have argued for the plausibility that existing models and algorithms could be extended or adapted to address various technical problems associated with screening baggage. The appeal of incorporating ML into this process leads to two suggestions:

- First, these methods could result in an improvement in performance over current practices.
- Second, they could do so in a way that is ultimately more efficient and effective, potentially at a cost savings.

TSA can take steps to incorporate ML in both the medium and long terms. In the medium term, existing ML technologies for analyzing X-ray and CT imagery could be straightforwardly incorporated into the existing baggage screening process, although the engineering of such a system would certainly require substantial effort. In the longer term, the screening process would benefit from richer scanning and data collection practices and data fusion. This would enable more-powerful ML models, but it would also require a more substantial overhaul of the existing screening process and current TSA technology acquisition practices.

Finally, although we are generally optimistic that ML could improve airline security, we do not want to overstate the case. ML methods have limitations, and their performance is typically constrained by the quality and amount of the data used to train them, which presents challenges in TSA's operating environment and would require investments in and development of new data systems. In the case of checked baggage screening, these data will likely be extremely imbalanced because the overwhelming majority of bags are benign, privacy considerations must be taken into account, and, given the safety-critical nature of the problem, VVT&E must be developed and maintained to ensure acceptable performance and ongoing confidence in the capability. Fortunately, ML technologies are well suited to iteration and incremental improvement.

Given the potential for ML approaches to improve baggage screening, we suggest near-term investments by TSA and DHS, the department that houses it, to develop that potential. Moreover, the

iterative introduction of ML technologies into the baggage scanning process could also help set the stage for a wider set of investments to leverage ML technologies to enhance existing capabilities and derive value from new capabilities as they are introduced across any number of programs. For example, DHS programs that are manpower intensive and that require rapid assessment of complex information, border surveillance for example, could leverage ML in a manner analogous to the earlier suggestions for improving assessments of three-dimensional X-ray images of checked baggage. Or, programs that already leverage multiple data sources, such as customs inspections or disaster preparedness, could benefit in the near to medium term from ML algorithms that can process large amounts of disparate data to find informative patterns that would likely escape human assessments and help develop improved planning outcomes. Furthermore, given the iterative nature of ML development, investments now in these kinds of programs could result in materially better outcomes as additional data sources come online. And importantly, given the iterative and incremental development potential of ML, the sooner developmental efforts start, the more valuable the future results are likely to be.

Notes

[1] We were part of a Homeland Security Operational Analysis Center research team conducting technology development analysis for the EBSP that also included Kelly Klima, Thao Liz Nguyen, Ian Mitch, Amado Cordova, Ryan Consaul, Jarrett Catlin, and Emily A. Gunn. This Perspective builds on that research, and much of the content is drawn from the non–publicly available documentation of that work.

[2] As noted earlier, early forms of AI and ML have been developing for decades.

[3] Multiple highly optimized and well-maintained DL libraries, such as TensorFlow or PyTorch, have greatly facilitated research and development progress in ML.

[4] An ML algorithm can be trained to detect any of three types of anomalies:

- **Point anomalies** are the simplest type in that they are far outside the entirety of the data set. A picture of an asteroid would be a point anomaly in an image database of North American animals.
- **Context anomalies** are detected when additional information characterizes the value as abnormal. For example, consider time series. Triple-digit Fahrenheit temperatures are normal for Arizona but not in December. Adding the temporal aspect to the anomaly detection helps distinguish values that might be otherwise overlooked.
- **Collective or group anomalies** occur when a group of entities shows abnormal behavior together but the behavior individually is not abnormal. A car turning right or left out of a parking lot is normal; however, cars consistently turning right out of the same parking lot (absent a change in traffic signals) is a collective anomaly.

Anomaly-detection ML approaches have been successfully developed for all three of these problem types and for different data modalities (e.g., image, numeric) (Chalapathy and Chawla, 2019).

[5] A Rapiscan dual-view system was the scanning technology used in the study.

[6] This approach does not conduct continuous, real-world training. Rather, it requires that the balanced training data set be updated as new threats or ways of disguising the threats are discovered or realized. The ML algorithm would be retrained using an updated but still balanced training data set on a periodic or as-needed basis.

[7] A much simpler metric is only the accuracy (i.e., the fraction of instances that are correctly classified), although, in many classification problems, accuracy alone is not able to fully characterize the performance of the system. For example, consider a classification problem in which 99.99 percent of objects are class 0 (no threat) and only 0.01 percent are class 1 (threat). A classifier that always predicts class 0, regardless of the input data, will do quite well when judged by accuracy—it will have an average accuracy of 99.99 percent. Obviously, this classifier is simply taking advantage of the base rate (or prior) of each class and has not actually learned to distinguish between the two classes. This motivates the need for the ROC.

[8] Morris, Chien, and Goodman did not publish their ROC, only the area under the curve (AUC). Later in this Perspective, we depict a symmetrical notional ROC to illustrate how tradeoffs must be made between threat escapes and false alarms.

[9] Morris, Chien, and Goodman also published their database of scans, providing other researchers the opportunity to improve on their research. To date, the development of and access to databases has been a key element in maturing ML algorithm development.

[10] A voxel is a 3D pixel. The dimensions of the voxels are specific to each type of EDS machine.

[11] DL neural network models consist of multiple layers of processing that are stacked onto one another. As more layers are added, the model becomes deeper, which allows more-complex relationships to be learned from the data.

[12] Although the authors of these papers considered only the binary classification problem, the successful demonstration of these approaches requires that the neural networks learn useful representations (or features) of the X-ray and CT data. These features are likely useful for a variety of ML tasks, including anomaly detection.

[13] Other factors, such as a lack of radiologist experience, cost pressures that reduce the time that can be spent on each image, and fatigue due to long hours, are also issues (Brady, 2017).

[14] The number of publications on ML applications using DL, for example, has exponentially grown in the past two decades.

[15] The researchers developed a convolutional neural network–based model that achieved a ROC of 0.77 over 83 disease classes. Eighteen of the disease classes together had a collective ROC of greater than 0.9, while six had a collective ROC of less than 0.6. Although the successful examples are promising, these initial results show that ML applications can be very effective in some areas of classification but struggle in others.

[16] There is some potential to expand data collection techniques using existing scanning machines. Right now, these machines use a combination of the Z_{eff} and density to assess baggage contents with fairly simple algorithms. However, if additional data could be extracted from the EDS machines, such as through X-ray attenuation, a trained AI system might be able to eliminate some of the uncertainty between threatening and innocuous materials. However, more research is needed to determine whether adding attenuation data to Z_{eff} and density would provide enough additional discrimination information to make such an effort worthwhile. Moreover, contractual and design hurdles would need to be overcome to make the data available from existing EDS machines.

[17] Vendors of scanning tools consider the algorithms that their scanners use to be a proprietary competitive advantage and are unlikely to allow anyone access to them or to the data produced by the scanning technology and used by the algorithm, because control of the algorithm and data helps ensure that those vendors maintain an ongoing business relationship with TSA and helps keep potential competitors from entering the market.

[18] Open architectures invite information security vulnerabilities that must be protected and would complicate their development and fielding.

References

Abu, Sanusi Darma, and Fatma Susilawati Mohamad, "Approaches of Deep Learning in Persuading the Contemporary Society for the Adoption of New Trend of AI Systems: A Review," *International Journal of Scientific and Technology Research*, Vol. 9, No. 12, December 2020, pp. 163–177.

"AI Versus Doctors," *IEEE Spectrum*, Vol. 54, No. 10, October 2017, p. 13.

Bondell, Sara, "Can Artificial Intelligence Outperform Humans?" *Endeavor*, December 2, 2020. As of June 24, 2021: https://moffitt.org/endeavor/archive/can-artificial-intelligence-outperform-humans/

Brady, Adrian P., "Error and Discrepancy in Radiology: Inevitable or Avoidable?" *Insights Imaging*, Vol. 8, No. 1, February 2017, pp. 171–182.

Burkov, Andriy, *The Hundred-Page Machine Learning Book*, Canada, 2019.

Captricity, "Artificial Intelligence Surpasses Human Ability to Read Handwriting," press release, Insurance-Canada.ca, March 27, 2018. As of June 24, 2021: https://www.insurance-canada.ca/2018/03/27/captricity-ai-read-handwriting/

Chalapathy, Raghavendra, and Sanjay Chawla, "Deep Learning for Anomaly Detection: A Survey," arXiv preprint, version 2, last revised January 23, 2019, arXiv:1901.03407. As of September 8, 2021: https://arxiv.org/abs/1901.03407

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

Devopedia, "ImageNet," webpage, last updated April 7, 2021. As of September 8, 2021: https://devopedia.org/imagenet

Draelos, Rachel Lea, David Dov, Maciej A. Mazurowski, Joseph Y. Lo, Ricardo Henao, Geoffrey D. Rubin, and Lawrence Carin, "Machine-Learning–Based Multiple Abnormality Prediction with Large-Scale Chest Computed Tomography Volumes," *Medical Image Analysis*, Vol. 67, January 2021, art. 101857.

Elkan, Charles, "The Foundations of Cost-Sensitive Learning," *IJCAI'01: Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Vol. 2, August 2001, pp. 973–978.

Gilmer, Justin, Ryan P. Adams, Ian Goodfellow, David Andersen, and George E. Dahl, "Motivating the Rules of the Game for Adversarial Example Research," arXiv preprint, last revised July 20, 2018, arXiv:1807.06732. As of September 8, 2021: https://arxiv.org/abs/1807.06732

Greene, Tristan, "AI Fails to Recognize These Nature Images 98% of the Time," *Next Web*, July 18, 2019. As of June 24, 2021: https://thenextweb.com/news/ai-fails-to-recognize-these-nature-images-98-of-the-time

Hartnett, Gavin S., Andrew J. Lohn, and Alexander P. Sedlack, "Adversarial Examples for Cost-Sensitive Classifiers," arXiv preprint, submitted October 4, 2019, arXiv:1910.02095. As of September 8, 2021: https://arxiv.org/abs/1910.02095

Kundu, Rohit, Ritacheta Das, Zong Woo Geem, Gi-Tae Han, and Ram Sarkar, "Pneumonia Detection in Chest X-Ray Images Using an Ensemble of Deep Learning Models," *PLoS ONE*, Vol. 16, No. 9, September 7, 2021, article e0256630. As of September 8, 2021: https://doi.org/10.1371/journal.pone.0256630

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature*, Vol. 521, May 28, 2015, pp. 436–444.

Lu, Jiajun, Hussein Sibai, Evan Fabry, and David Forsyth, "No Need to Worry About Adversarial Examples in Object Detection in Autonomous Vehicles," arXiv preprint, submitted July 12, 2017, arXiv:1707.03501. As of September 8, 2021: https://arxiv.org/abs/1707.03501

Mandrekar, Jayawant N., "Receiver Operating Characteristic Curve in Diagnostic Test Assessment," *Journal of Thoracic Oncology*, Vol. 5, No. 9, September 2010, pp. 1315–1316.

Mitchell, Lloyd, "Is Open Architecture Open Season for Hacking?" *Federal News Network*, December 29, 2017. As of August 5, 2021: https://federalnewsnetwork.com/all-news/2017/12/is-open-architecture-open-season-for-hacking/

Mitchell, Tom M., *Machine Learning*, New York: McGraw-Hill, 1997.

Morris, Trevor, Tiffany Chien, and Eric Goodman, "Convolutional Neural Networks for Automatic Threat Detection in Security X-Ray Images," *2018 17th IEEE International Conference on Machine Learning and Applications*, 2018, pp. 285–292.

Papers with Code, "Image Classification on ImageNet," undated. As of August 2, 2021: https://paperswithcode.com/sota/image-classification-on-imagenet

Poole, David, Alan Mackworth, and Randy Goebel, "Computational Intelligence and Knowledge," *Computational Intelligence: A Logical Approach*, New York: Oxford University Press, 1997, pp. 1–22. As of June 24, 2021: https://www.cs.ubc.ca/~poole/ci/ch1.pdf

Public Law 107-71, Aviation and Transportation Security Act, November 19, 2001. As of September 8, 2021: https://www.govinfo.gov/app/details/PLAW-107publ71

Transportation Security Administration, *Planning Guidelines and Design Standards for Checked Baggage Inspect Systems*, version 7.0, Washington, D.C., August 21, 2020, updated October 8, 2020. As of September 8, 2021: https://sam.gov/opp/5223ca7b095942749bde423905f507d7/view

TSA—*See* Transportation Security Administration.

Acknowledgments

We are grateful for the support that was provided by two people in particular from the office of our sponsoring office, the U.S. Department of Homeland Security (DHS) Science and Technology Directorate (S&T): David Brandenburg and, before him, DaShunda McDonley. Their ongoing support and guidance, along with the advice and assistance of Steven Sundbeck and Anthony Visk from S&T's EBSP program office, made this project possible and kept us on track. Craig Mosford, from TSA, provided us with a tremendous amount of information (both technical and procedural) about electronic baggage screening. This information was critical for our understanding of the issues associated with baggage screening. Finally, we thank the many people in TSA, academia and other federally funded research and development centers, and other government agencies who consented to and participated in technical discussions with us. Their candor and patience helped us understand the complex requirements and issues that are associated with checked baggage screening and the developmental activities—ongoing or completed—to address these issues.

We thank our peer reviewers, Rand Waltzman and William Shelton, for their constructive reviews. The Homeland Security Operational Analysis Center Acquisition and Development Program management team, including Emma Westerman, Paige Rudin, Bergen Goesch, and Natalie Richards, also have our gratitude. Lisa Bernard, our editor, greatly improved the final Perspective and deserves our appreciation. Finally, we also thank the team that collaborated on the larger project, including Kelly Klima, Amado Cordova, Thao Liz Nguyen, Ian Mitch, Ryan Consaul, and Jarrett Catlin. That work provided the basis for this Perspective.

About This Perspective

The Aviation and Transportation Security Act (Pub. L. 107-71, 2001) requires that the Transportation Security Administration (TSA) screen 100 percent of checked baggage at airports. To fulfill this mandate, TSA procures, installs, and maintains screening systems for checked baggage at airports through its Electronic Baggage Screening Program (EBSP). The Homeland Security Operational Analysis Center was asked to produce a 20-year EBSP technology roadmap to inform the future of checked baggage screening.

One possibility for improved baggage screening is to incorporate recent advances in artificial intelligence (AI) and machine learning (ML) in the screening process. There is a large and growing list of cognitive tasks for which ML can outperform humans, and many of these tasks are similar to those involved in TSA's screening process for checked baggage. This Perspective describes the current state, challenges, and outlook for AI and ML, using the EBSP example as a case study. It builds on research and analysis conducted more broadly on technology development for EBSP, and much of this Perspective is drawn from the non–publicly available documentation of that work. The findings will likely be of interest to those planning to incorporate AI or ML into legacy technology systems. This research was sponsored by the U.S. Department of Homeland Security (DHS) Science and Technology Directorate (S&T) and conducted within the Acquisition and Development Program of the Homeland Security Operational Analysis Center federally funded research and development center.

About the Authors

Shawn McKay is a senior engineer at the RAND Corporation. His current research has focused on AI applications in decisionmaking. He has a Ph.D. in engineering.

Gavin Hartnett is a physical scientist at the RAND Corporation and a professor at the Frederick S. Pardee RAND Graduate School. A theoretical physicist by training, he conducts policy-oriented research into a wide range of topics, including artificial intelligence, machine learning, blockchains, and quantum computing. He has a Ph.D. in physics.

Bruce Held is a senior policy researcher at the RAND Corporation. Since coming to RAND, he has focused his research efforts on defense acquisition, the defense industrial base, and defense research and development policy. He has an M.S. in engineering and a J.D.

HSOAC
HOMELAND SECURITY OPERATIONAL ANALYSIS CENTER

An FFRDC operated by the RAND Corporation under contract with DHS

The Homeland Security Act of 2002 authorizes the Secretary of the Department of Homeland Security (DHS), acting through the Under Secretary for Science and Technology, to establish one or more federally funded research and development centers (FFRDCs) to provide independent analysis of homeland security issues. The RAND Corporation operates the Homeland Security Operational Analysis Center (HSOAC) as an FFRDC for DHS under contract HSHQDC-16-D-00007.

The HSOAC FFRDC provides the government with independent and objective analyses and advice in core areas important to the Department in support of policy development, decisionmaking, alternative approaches, and new ideas on issues of significance. The HSOAC FFRDC also works with and supports other federal, state, local, tribal, and public- and private-sector organizations that make up the homeland security enterprise. The HSOAC FFRDC's research is undertaken by mutual consent with DHS and is organized as a set of discrete tasks.

The information presented in this publication does not necessarily reflect official DHS opinion or policy.

This research was published in 2022. For more information on this publication, visit www.rand.org/t/PEA731-1.