11 Scoring Simulations with Artificial Intelligence

Carter Gibson, Nick Koenig, Joshua Andrews, and Michael Geden

CONTENTS

The world is currently living through what some have called the fourth industrial revolution (Schwab, 2017). In this framework, the first three revolutions related to water and steam power, electric power, and electronics and information technology to automate production, respectively. This fourth industrial revolution is characterized by technologies that blur the lines among the physical, digital, and biological spheres and include such advances as the Internet of Things, 3D printing, nanotechnology, quantum computing, and, most importantly for this chapter, artificial intelligence (AI). Use of AI has increased in the last decade and is a large driver of innovation (Rust & Huang, 2014). In obvious and less obvious ways, AI is already impacting many areas of daily life (Poola, 2017). AI gets attention for high-profile uses, such as in autonomous vehicles or how it's proving more accurate than expert

DOI: 10.1201/9781003401360-11

radiologists (e.g., Hosny et al., 2018; Schwarting, Alonso-Mora, & Rus, 2018). AI is also being used in smaller ways to subtly improve areas of modern life, such as in unlocking a phone with facial recognition, giving grammar advice for writing, filtering emails as spam, giving users personalized ads when browsing the web, or helping banks identify fraud (e.g., Ryman-Tubb, Krause, & Garn, 2018). Already ubiquitous, AI, and the fourth industrial revolution more broadly, promises to impact almost every field and job, including simulations and training.

Crucial to understanding how AI is changing the field of simulations and training is first examining where simulations started. Simulations are built to mimic or reproduce a specific context. The typical goal is to train or measure in an environment at lower risk than learning on the job. For example, it's not best practice for a pilot to learn how to perform a difficult maneuver or a surgeon to try out a new technique in the high-stakes context of their actual work. Perhaps an organization simply wants to train a leader to be a better communicator or give higher-quality performance appraisals. Simulations allow for structured, safe, and deliberate practice in a lower-stakes environment to develop skills that will transfer to the higher-stakes circumstances in the workplace.

Simulations exist across a wide continuum, from highly realistic and technical (e.g., a flight simulator that accurately reproduces all of the controls in a plane) to more conceptually representative, like a paper-and-pencil activity. At a high level, the concept of fidelity can be demonstrated by how closely a simulation can recreate the appearance and potential dynamics of the simulated scenario. A realistic flight simulator would be high fidelity, whereas the paper-and-pencil task would be lower fidelity. Though some have criticized the concept of fidelity as poorly defined (e.g., Norman et al., 2018), the term is still useful for framing thinking about simulations in the field. High-fidelity simulations may create colossal amounts of data. In the context of a flight simulator, the computer could record what inputs are made, how quickly they're made, how much pressure is applied, where the individual in the simulation is looking, and vital signs of the participant. With many of these issues, a novice may not be able to comprehend the importance of so many measures, but experts are able to take all of these inputs and provide specific feedback, advice, or general conclusions that can improve the performance of the participant in the exercise. On the other end of the fidelity spectrum, a training exercise could have a leader going through a developmental assessment center, where they work through an in-basket task writing emails, solving problems, or organizing their calendar (Motowidlo, Dunnette, & Carter, 1990). Again, a large amount of data is being created and historically has relied on expert human judgment to determine the quality of performance across the range of constructs being measured.

How can simulation performance be accurately scored across varying levels of fidelity? High-fidelity simulations may have a large number of variables with varying degrees of importance, but performance in low-fidelity simulations may still be difficult to measure and quantify empirically. For example, how do you quantify the outcomes of the simulation designed to score a proficiency exam or determine who among a group of applicants should be hired for a position? While much research has gone into measurement, less work exists to combine these sources of data to predict

important criteria (Sydell et al., 2013). We introduce the concept of fidelity to show that AI can be useful across a wide range of types of simulations, regardless of the type of data that is generated.

The issue of what to do with all of the data generated by simulations and how to score them in reliable and valid ways are where AI has the potential to significantly impact how simulations are used. And while AI is changing the field, these changes were not unanticipated. That is, scholars have pointed to this future long before technology had the models and processing power to bring them about. Scholars pointed to two major predictive improvements in the area of scoring simulations: (1) combining information across item types and assessment experiences, and (2) leveraging the power of increasingly large sample sizes (and data sources) (Sydell et al., 2013). AI is following through with these promises and has significant implications for training and simulations, specifically because of its ability to automate scoring of data sources that were previously impossible to score by machines (or even, sometimes, by humans) and through more accurate models of scoring.

Of course, like many new technologies, AI won't destroy and replace what came before it, but rather provides a new tool. This chapter discusses what AI can do, what it can't, and offers suggestions for users to start incorporating it into their own work. We believe AI will change the way simulations are used to select and train individuals by allowing simulations to be more easily scaled, automating previously manual scoring approaches, and helping experts design better and more predictive weighting schemes to create more optimal scoring models of complex behaviors.

ARTIFICIAL INTELLIGENCE AND REPRODUCING EXPERT RATINGS

Automation within the field of human factors has been an area of interest for years. Since their invention, computers have become exponentially more powerful and cheaper following a pattern called Moore's Law (Moore, 1965). Several books within the field of human factors have been dedicated to the subject (e.g., Parasuraman & Mouloua, 1996; Mouloua & Hancock, 2020), and hardware and software are consistently making more complex automation possible. Mosier and Manzey (2020) discuss automated decision support systems (DSSs) and the value these systems provide in reducing user bias across a variety of industries. But what if it was possible to use expert human judgment to train a system and remove the experts from the loop?

In many disciplines, expert human judgment is leveraged for decision-making in complex tasks. From doctors reviewing patient MRIs to assessors evaluating assessment center candidates, expert judgment has been shown to outperform novice judgment (Salkowski & Russ, 2018; Schleicher et al., 1999). Wickens et al. (2016) described decision-making as including the following key features: uncertainty, time, and expertise. These features are certainly present in many high-stakes environments where decisions must be made. The complexity of human decision-making is ripe ground for AI as advances in hardware and software make replicating complex human decisions more feasible.

While AI from the 1950s to the 1980s often involved explicitly programming symbolic representations of logic and decision-making into the computer, coined

"Good Old-Fashioned Artificial Intelligence" (Haugeland, 1985, pg. 112), more recent conceptualizations of AI have leveraged the idea of machine learning and the understanding that the software can program itself if given enough data. The availability of more processing hardware at lower costs has allowed for more and more model parameters and the introduction of deep learning. Model parameters, which are defined as variables internal to the model that are estimated using data, are extremely important to all of machine learning. An example of a model parameter in a linear regression is a beta weight, which is estimated by optimizing for best fit. In the field of natural language processing, parameters in the order of magnitude of several thousand via bag-of-words were considered large just a decade or two ago. Now, we have architectures like BERT with 345 million parameters (Devlin et al., 2019) and, more recently, GPT-3 and its 175 billion parameters (Brown et al., 2020). These algorithms consist of an input layer, where data is fed into the model, several hidden layers, and an output layer where a specific prediction is made. This is considered deep learning because the neural network has several hidden layers.

The increase in the size of the parameter space leads to more and more complex representations of the data via the layers and internal neurons' ability to extract very specific subsets of information from the input data. These complex algorithms make it possible to replicate human judgment on natural language-based tasks, such as evaluating the quality of an essay, a written job simulation, or even a job candidate's interview response. This approach can be leveraged for automating simulations and systems where trained professionals currently need to evaluate natural language-based responses and make decisions. The need to automate the understanding of human language exists across several domains. In the medical field, this technology has been used to take clinical notes and predict hospital readmission (Huang, Altosaar, & Ranganath, 2020) and patient diagnoses from electronic health records (Franz, Shrestha, & Paudel, 2020). In the field of human resources, natural language processing has been used to automate job analysis (Mracek et al., 2021) and the scoring of work simulations (Tonidandel et al., 2020). In the following pages, we will outline a process for developing an algorithm that can replicate trained subject matter experts when it comes to evaluating work simulations and interview responses based on the written or spoken English language.

Assessment centers, in-person or virtual, often have several writing exercises in the form of in-baskets that require participants to respond to an email from a peer, boss, or customer. These unstructured text responses can then be evaluated for competencies relevant to success in the role. Extracting scores from these responses requires no small investment in resources. The responses can be lengthy and accurate ratings require review by trained evaluators, knowledge workers who receive substantial compensation for their expertise. In addition, the process itself is extremely repetitive, which can lead to a vigilance decrement (Thomson et al., 2015) extremely common in such work. Transformer-based NLP algorithms are most effective for use with such long-form responses. Simple chat simulations or short answer responses would likely not benefit from the added complexity these algorithms provide.

TRADITIONAL APPROACH TO SCORING OPEN-ENDED CONTENT: RATER TRAINING

The traditional method for scoring open-ended content is an expensive process in terms of both time and money. Almost by definition, the SMEs qualified to provide ratings on a complex subject are going to be both busy and expensive. Using them to rate and evaluate a large sample of any product will be challenging, whether in an academic or applied context. In the context of an organization, maintaining a stable group of trained judges can be challenging as people leave the organization or their original role. Furthermore, if ratings of a particular product are needed in a timely fashion, it may be difficult to get quick work from a judge. Plus, several of the steps require the judges to coordinate and discuss a shared frame of reference.

AI isn't going to remove humans, or in this case experts, on a given topic, but rather change their role. Using AI won't be as simple as just applying an algorithm and solving a problem; the tools described in this chapter still require significant effort to ensure appropriate data is being fed into the system. If the end goal is a program that can rate a work product, such as a writing sample, a large pool of writing samples as well as expert ratings of these samples would be needed. The way these expert ratings are collected, even for the purposes of building an AI model, is going to look a lot like it has traditionally. A large pool of literature dating back several decades exists on how to train raters most effectively (e.g., Bernardin & Buckley, 1981), and all of these proven steps still need to be followed. At its core, what data scientists are trying to do is translate a large amount of text or other data into a useful and reliable score in a standardized way. Once this training is completed, a set of judges with a shared mental model of the constructs being assessed will have been created. Deep learning can then be used to recreate these expert ratings, and ultimately be able to rate new writing samples independent of human judges in a reliable and valid way.

While many approaches could be used, frame-of-reference training is perhaps one of the most popular (Roch et al., 2012). Conceptually, the goal of this training is to get all judges onto the same metric to minimize differences due to judges with unique ideas about what is important when rating task performance. The first step is to create rating benchmarks and standards for all variables to measure. For example, in a writing sample, it's possible to rate overall quality and also more specific constructs such as grammar, vocabulary, and style. To create these benchmarks, the collection of writing samples would be reviewed to find examples of various anchors on the scale, such as at 1, 3, and 5 on a 5-point scale. This process would be repeated for each of the variables the judges will rate. Once the benchmarks have been built, all the judges meet to work through a small sample of cases. For each of these cases, they would provide ratings for each variable and then discuss them until they came to a consensus (i.e., shared-mental model) on what a "5" looks like, what a "3" looks like, etc. Once the judges appear to be rating in a consistently similar way, the judges would proceed to review the entire sample and rate all cases. For the purposes of AI, several hundred rated samples would be needed, typically with at least three judges to ensure confidence that the "true" rating for each writing sample

has been obtained. There may be a need to have periodic meetings and calibrations to account for things like rater drift (Harik et al., 2009). Once the judges have rated several hundred cases, an important check is to review inter-rater reliability, or the degree to which raters are consistently giving the same ratings on each variable of interest for each sample of writing. See Hall and Brannick (2008) for a good review of the various considerations when choosing a metric for inter-rater reliability, and Gibson and Mumford (2013) for an example of this rater-training process used in practice.

Of course, many judgment calls will need to be made about specifics in the process. We prefer to be conservative about ratings to ensure the highest quality data to train the model. For example, in some contexts, once judges have established sufficient agreement, only one judge may be needed to rate new data. This is more likely to be viable in cases with extremely high inter-rater agreement or when rating more concrete variables (e.g., a construct that has been very specifically operationalized). In other cases, raters may be allowed to have different ratings on a specific instance as long as, on average, they're in agreement (e.g., one rater gives a "2" and another gives a "5" on the same product). Given the relative newness of using these algorithms, we have opted to be more conservative, such as expecting ratings to have evaluations within one point of each other on a 1-to-5 rating scale. Thus, if the first rater evaluates a work sample as a "4," the third rater would need to have an evaluation of 3, 4, or 5, for that same work sample; otherwise, they would need to meet to discuss and draw a shared conclusion.

Now that labels have been created, the next step is identifying the algorithm to be used. While bag-of-words (Zhang, Jin, & Zhou, 2010) and long short-term memory recurrent neural networks (Hochreiter & Schmidhuber, 1997) are reasonable methods to use, they both have shortcomings that are beyond the scope of this chapter. Note that many of the approaches described in this chapter are new and are still actively being developed, so rather than dive into a technical guide of a given method, it is more useful to review broadly the considerations when choosing an analytic technique. More recently, Vaswani et al. (2017) introduced the transformer architecture, and Devlin et al. (2019) and Liu et al. (2019) expanded upon the transformer architecture to create the current state-of-the-art model: the bidirectional encoder representations architecture (BERT). This architecture has the capacity to take in embeddings as representations of the words and adjust those representations depending on the words coming both before and after it. These word embeddings hold information about the word or token's relationship to other words. This makes it possible for the language model to understand that in most cases, the words "customer" and "client" are very similar. While this architecture was originally very successful at predicting masked words (i.e., what a hidden word was most likely to be given surrounding words) and similarly creating state-of-the-art language translations (i.e., Google Translate), the researchers quickly realized it could also excel in downstream tasks, like question answering, predicting sentence sentiment, and more. These downstream tasks are tasks that the model wasn't explicitly trained for, but could be trained to do with new data and sufficient computational power. Because of its ability to produce accurate results on a number of complex natural language

processing tasks, this architecture is ideal for the downstream task of replicating human evaluations on complex human behaviors within written work simulations.

The Architecture

BERT is freely and readily available via Hugging Face (huggingface.co). There are many variants to choose from, but we recommend RoBERTa's base model. There is also the need for hyperparameter tuning of the model. Hyperparameters are parameters outside of the algorithm itself that control how the algorithm performs. In neural networks, these can be things like learning rate, percentage of nodes/neurons that are dropped (dropout), optimization function, batch size, and more. We found success trying differing learning rates and dropout rates while using the largest batch size that could fit into memory.

The Data

An important point about machine learning and deep learning in general is that the algorithms are extremely powerful when it comes to learning from the data they were trained on. For this reason, users need some form of a holdout set to ensure predictions will generalize on responses outside of the specific responses the algorithm was trained on. When hyperparameter tuning, users will want to use a cross-validation strategy to prevent inadvertently overfitting to the holdout set. Hyperparameters, unlike model parameters, are parameters that need to be set outside of the model itself, but can have an impact on the quality of prediction. For this reason, it may make sense to test a variety of combinations to identify the best set for your given data. One common hyperparameter for the transformer architecture is learning rate, which is simply the size of the update step used for moving along the gradient. Another common hyperparameter is neuron dropout. This is the proportion of neurons within the hidden layers that are randomly set to zero. This is an extremely effective regularization technique for deep learning. First, test different hyperparameters on a *k*-fold cross-validation sample set. A *k*-fold cross-validation set involves slicing your data into *k*-folds; common *k*'s used would be 5 or 10. For example, given a dataset with a sample size of 1,000 and a 5-fold cross-validation, 800 responses and labels would be used to train the algorithm, 200 responses would be predicted, and then the process would be repeated with each set of 200 being used as the holdout set. When hyperparameters are found that provide satisfactory results, train the final model on a training set consisting of roughly 80% of the data and evaluate the model's performance on a holdout set of 20% of the data.

We also recommend data stratification, which consists of identifying differences in the data and ensuring those differences are consistent across the folds. The outcome/label is that it is important to ensure it is stratified across folds. Other things to consider may be the length of the responses, the population the sample was taken from, and any other parameters that may differ within the data. This stratification ensures that the model is consistently being trained and evaluated on very similar data.

An example we have used this on was virtual assessment center in-basket email responses used as part of a multi-method assessment for job selection. Candidates responded to a fictitious email from a colleague inquiring about a problem they were facing and asking for guidance on how to proceed. The candidate was asked to respond to the email with recommendations for handling the specific situation. This response was evaluated on a number of competencies, from effective communication to an ability to drive results. One thousand responses were labeled by two trained subject matter experts on each of the competencies operationalized using predetermined behaviorally anchored rating scales, and the two raters had to come to a consensus on the competency rating. The responses and labels were then stratified across label distributions into both 5-folds and a final unrelated 80/20 split for final algorithm training.

Several different dropout rates ranging from 0.05 to 0.15 and several different learning rates ranging from 1-e3 to 1-e6 were tested. The means and standard deviations of the correlations along with the means and standard deviations of the mean squared errors (MSEs) on the holdout folds were compared to one another and a final set of hyperparameters was chosen. MSE was chosen as the optimization function because our specific purpose involved a regression-based output. The best hyperparameters were then used to train a final model on the original 80/20 stratified split.

Output

The final result is a deterministic algorithm that can be used to make predictions on new work samples within milliseconds of when the candidates produce them. The algorithm is deterministic in the sense that, given identical inputs, it will always produce an identical output. This differs significantly from the stochastic nature of the deep learning training process, where given the way the model is trained each time you retrain it, you will end up with different parameter weights and thus different predictions. A current paper by some of the authors (Thompson et al., forthcoming) found correlations between SMEs and the algorithm predictions that averaged above 0.84 on seven separate competency/work simulations. We also found that, on average, predictions on the competencies were within one point of the consensus (on a 1–5 scale) SME evaluations 91% of the time and within 0.5 of the consensus SME evaluations 66% of the time, providing evidence that the algorithms are consistently and accurately replicating the SMEs' evaluation of the job candidate on these job-relevant competencies. In an effort to examine how good or bad these hit rates were, the algorithm was compared to the pre-consensus SMEs. This evaluation found that 75% of the time the SMEs were within one point of each other on a rating before consensus. This research suggests that not only is the algorithm evaluating the responses very similarly to the consensus rating provided by the two SMEs, but it is also producing more consistent ratings than any one SME.

Other Considerations

As was mentioned earlier, there are other options for replicating work simulation evaluations on natural language data. Our research found that the transformer

architecture outperforms the bag-of-words and LSTM architectures using 25% of the sample size. With a training set of 250 responses, we found that the evaluations on the holdout of 250 responses were more accurate than when using 750 responses to train the bag-of-words model. This is almost certainly a result of the transfer learning these transformers provide. As briefly discussed earlier, the transformer architecture comes with data built into it via its language model. It naturally has a vocabulary and representations of that vocabulary from the original purpose of the architecture, which was to predict masked words and upcoming sentences. This encoded information allows the transformer architecture to make robust and generalizable predictions on several downstream tasks with fractions of the data of more naïve architectures/implementations.

While the above may not provide a step-by-step review of how to use the described tools, it should provide a conceptual overview of the process by which work simulations can be evaluated using SMEs and leveraging the transformer architecture to produce highly accurate predictions on never-before-seen responses to the same work simulation, effectively removing the human rater from the loop and automating what was once considered an extremely complex task that required human expert judgment and decision-making.

SCORING ACTIONS IN SIMULATED ENVIRONMENTS

Beyond unstructured text (i.e., text that does not have a predefined format), simulations capture detailed data about an individual's actions within the environment, including their motions, interactions with objects, and contextual information (e.g., time stamp, NPC). Event and motion data could have transformative potential and utility for scoring performance and providing feedback in virtual simulations. Actions within simulations are often logged by the software in a way that would be impossible for human raters to comprehend. Simulations produce a rich source of data on individuals, providing a considerable opportunity for training AI/machine learning models to identify new metrics and features relevant to success. Nonetheless, this source of data also comes with additional complexity and considerations that makes it difficult to structure and analyze.

Traditional Approaches to Scoring Simulations

Scoring and rating simulations still rely heavily on SMEs to generate scoring metrics and provide trainee feedback. The current gold standard for scoring such simulations is to leverage SMEs to develop a mapping from an individual's actions to the quality of their performance within the simulation, typically by providing ratings along the relevant dimensions of interest for the virtual simulation (Boyle et al., 2018; Oquendo et al., 2018). This is often performed during validation of a training simulation or when the simulation is used for evaluation purposes. Trained raters then use assessment tools (e.g., rating rubrics, checklists) to score and provide feedback to trainees.

Nonetheless, scoring approaches that use SMEs have a focused perspective that often uses only a small amount of the available simulation data. SMEs are especially

valuable because they can identify and measure complex and abstract behaviors and constructs within simulated environments, but often at a high cost due to the required expertise and training. The reliance on SMEs for scoring and rating simulations is not cost-effective in providing scalable feedback and high-fidelity training to enable widespread adoption. AI/machine learning techniques are particularly good at identifying patterns and therefore could help supplement traditional scoring approaches with a scalable alternative. Expert ratings for unstructured, non-text data would be invaluable for training an AI to identify mistakes and potential errors in real time, which in turn would enhance feedback and skill acquisition.

An alternative approach for more scalable scoring is to produce simple metrics based on SME domain knowledge that can be easily calculated in an automated fashion. This manual feature engineering typically results in a small number of easily interpretable metrics. For example, the Fundamentals of Robotic Surgery (FRS) has offered standardization for scoring using a set of metrics, such as time-to-completion and deviation in cutting from a prespecified region. It is also commonly employed in game-based assessments using evidence-centered design (ECD), which specifies relationships between actions in a simulation with concepts the student is trying to learn (Mislevy, Steinberg, & Almond, 2003). This approach, while scalable and simple to deploy, suffers from multiple drawbacks. First, these methods are often poor indicators of actual performance on these tasks (Mills et al., 2017). Additionally, the extraction of SME knowledge requires the use of time-consuming methods, such as cognitive task analyses, to derive the simple scoring system. Finally, manually engineered metrics are often highly specified in a context, making generalization to new scenarios challenging. For example, we would likely observe large differences between ideal metrics for scoring heart surgery compared to bone surgery.

Ideally, simulations should provide a scalable method for the administration and scoring of a scenario relative to the size of the trainee body. The point of simulations is to provide easy and safe practice for skills training, where the quality of learning during simulations is dependent upon the quality of feedback provided to the learner. Research has demonstrated that machine learning-aided skills evaluation is a scalable and automated means for measuring and collecting data on multi-dimensional evaluation constructs (Vedula, Ishii, & Hager, 2017). However, as we note, the applicable approaches are dependent upon the collection and structuring of training data, which often includes a time-based component.

Data Representations for Modeling Simulations

Virtual simulations generate detailed logs recording the actions made by the user and events that occurred within the virtual environment. These logs provide a rich source of information about the user; however, they are typically stored in formats unsuited for direct application in machine learning models. The logs contain information unrelated to modeling goals, are sampled at excessively high frequencies (e.g., location at each millisecond), and are stored in a representation that is suboptimal for direct use in modeling. The first step of the machine learning pipeline is to preprocess this data into a format amenable to analysis. Simulators sample at a

high frequency, as it is trivial to record the data and critical to capture every relevant action taken by the user, causing data files to quickly become extremely large (e.g., gigabytes per person). The high-frequency sampling rate causes many data points to be redundant, since the time between samples is so low that few actions have been performed. Down-sampling the data while ensuring critical actions are still recorded can dramatically speed up modeling time at little to no cost to predictive performance. It can also be necessary to adjust the sampling rate of data when merging across multiple sources, such as eye-trackers, videos, and tool motion (Vedula, Ishii, & Hager, 2017). The multi-modal data streams may be collected at different sampling rates through separate software, requiring adjusting the sampling rate of the disparate data streams to align time stamps for merging. Next, the log data is transformed to remove noise and make it easier for the model to learn the desired relationships. Transformations can occur across two axes: static and temporal. Static transformations translate representations on the observation level (i.e., each time point) from the goal of efficient storage for the simulator software to being more closely aligned with the objective of the model. For example, in game-based learning environments, it can be helpful to encode the user's goals, accomplishments, actions (e.g., talking to a non-playable character or NPC), and the entities with which they are performing the actions (e.g., the name of the NPC) (Geden et al., 2020; Min et al., 2017). It can also be useful for defining complex actions, such as the type of strokes made with a surgical tool in surgical simulations (Ahmidi et al., 2015). Additionally, temporal transformations remove trends and seasonal changes from the time series to remove undesirable artifacts from the data. A couple of common temporal transformations are taking the moving average, differencing, and detrending sessional components (Wei, 2006).

Manual feature engineering can also be used to improve the performance of a model by providing it with an SME's heuristic for interpreting the environment. While certain models are able to automatically generate features from raw data (e.g., deep neural networks), these approaches require large amounts of data to learn these complex relationships. SMEs can provide a curated set of features relevant to the task, allowing for the model to focus on the mapping from the heuristics to the outcome variable without having to learn the intermediate representation (Vedula, Ishii, & Hager, 2017; Uzuner, 2009; Kuhn & Johnson, 2019; Krajewski et al., 2009; Garla & Brandt, 2012). While potentially effective, manual feature engineering is a time-consuming and domain-specific process requiring SMEs to encode their knowledge, illustrating the continuing importance of SMEs in the development of new models even as AI tools are sought to replace them.

Machine Learning Methods for Scoring Simulations

A wide breadth of machine learning models have been successfully applied for scoring simulations across diverse domains such as education, medical, and transportation settings (Anh, Nataraja, & Chauhan, 2020; Henderson et al., 2020; Beninger et al., 2021). The diversity of machine learning models is partly fueled by the No Free Lunch theorem (Wolpert & Macready, 1997), which states that there is no single

"best" model to use across all circumstances, requiring the researcher to explore multiple methods and tailor their solution to the structure that is unique to their problem. This makes it impossible to provide a simple prescription for the selection of a model, as it may depend upon a number of factors, including the volume of data available, the form that the data is represented in, the type and number of criterion variables (i.e., classification, regression), and the unique structure of the task. For example, the transformer method described in the previous section has been extremely successful when handling text data; however, it is not applicable to tasks without a sequential component (e.g., credit loans) or non-language tasks with a small sample size (i.e., cannot apply pre-trained models).

An important aspect of modeling simulator data is that criterion will rarely be available for each time point collected within the simulation. Instead, criterion data will be intermittently gathered during a window of time, such as the task performed for the last 10 minutes of the simulation. This creates a discrepancy in structure between the three-dimensional feature data (event × time-window × feature) and the two-dimensional criterion data (event × criterion). The relationship between multiple time points of the feature data and a single criterion impacts both how the researcher should structure their data and what models they can use. Broadly, there are two approaches available: traditional machine learning methods can be used if the feature data is compressed along the time axis to create a summarized representation aligned with the criterion, or time series methods can be used that natively handle the problem.

Static Methods Using Summarized Representations

The compression of the time series predictors from a three-dimensional structure (event × time series × predictors) to a two-dimensional summarized representation (event × time series) is typically accomplished using either simple summary statistics or manually crafted features. The predictive performance of the machine learning model is entirely dependent upon the quality of the summarized features, making it critical that the researcher thinks carefully about which statistics relate to the criterion of interest. For illustrative purposes, we will walk through this approach using three commonly employed supervised learning models: support vector machines (SVMs) (Cortes & Vapnik, 1995), random forests (Breiman, 2001), and deep learning models (Rumelhart, Hinton, & Williams, 1985). All methods can be easily found in many programming languages (e.g., R, Python, MATLAB).

SVMs are a robust, non-probabilistic, linear classifier that finds the decision boundary that maximizes the distance between classes. SVMs bear a strong resemblance to logistic regression, as both create predictions based on a linear combination of features; however, this is with one notable difference: the objective being optimized. Logistic regression minimizes the negative log-likelihood of the data, providing a probabilistic interpretation of the likelihood of each sample belonging to a particular class. SVMs use the hinge loss with a regularization term to maximize the distance between classes, encouraging the model to not only differentiate between classes but also to do so confidently. Due to the regularized hinge

loss, SVMs provide a robust and scalable method that produces sparse solutions (e.g., coefficients encouraged to be 0). Zepf et al. (2019) used an SVM to predict drowsy driving in a simulated driving environment based on features automatically extracted from an EEG using principal frequency bands. Mirichi et al. (2019) used an SVM to create an interpretable model for predicting expertise in a VR simulation of a subpial tumor resection.

Random forests are an ensemble method constructed from a multitude of decision trees trained on random subsets of the data (Brieman, 2001). Decision trees are directed acyclic graphs that use simple binary rules (e.g., $X < 15$) to create predictions. Decision trees are able to map nonlinear structures; however, they have a tendency to overfit to the data and are very sensitive to outliers. Random forests address this limitation by combining many diverse decision trees to create a more stable, robust model. Beninger et al. (2021) used random forests, neural networks, and SVMs to predict inattention during driving in a simulated driving environment. They first preprocessed the data by lowering the sampling rate, sampling a 1-minute window of feature data before each event, normalizing the features within the window, and calculating summary statistics to flatten the data (e.g., minimum, maximum, median). In their evaluations, random forests outperformed the linear SVM and neural network. McDonald and colleagues (2014) used random forests to predict drowsy driving in a simulated driving environment based on features extracted from steering wheel motions.

Neural networks are directed acyclic graphs composed of layers with multiple nodes and are a universal function approximator (Rumelhart, Hinton, & Williams, 1985). Neural networks' extremely flexible structure has led to their widespread success and adoption, particularly on complex tasks with large amounts of data, such as text and image processing (Sun et al., 2017). The simplest neural network can be constructed from an input layer and an output layer with a single node and a linear activation function, which is the same structure as a linear regression. The most complex neural networks are composed of billions of parameters and thousands of layers (Wang et al., 2017; Devlin et al., 2019). Anh, Nataraja, and Chauhan (2020) demonstrated that deep neural networks were able to accurately assess surgical skill in suturing, knot tying, and needle passing. Richstone et al. (2010) used neural networks to predict expertise based on eye movements in simulated surgical environments.

Time Series Methods

Time series methods directly model the three-dimensional predictor data, sidestepping the need for creating a two-dimensional summarized representation. These methods typically require stronger assumptions about the structure of the time series data or the use of complex and flexible frameworks. Multivariate autoregression is a probabilistic model that predicts the criterion based on a linear combination of multiple previous time points of the predictors. Autoregressive methods require the researcher to specify the temporal dependence of the model (i.e., model order); this is usually found during exploratory data analysis by identifying trends and seasonal

relationships in the data. Multivariate autoregression is an interpretable model which supports statistical inference; however, it does not natively support nonlinear relationships between features and criteria. Loukas and Georiou (2011) use multivariate autoregression to predict laparoscopic skills (i.e., knot tying and needle driving) during surgical training based on hand motions.

Another commonly employed method is recurrent neural networks (RNN). RNNs model temporal data by recursively calling a node based on feature data at the current time-step and intermediate data from the previous time-step of the RNN node (Rumelhart, Hinton, & Williams, 1985). RNNs make no assumptions about the data and can model nonlinear relationships; however, they are uninterpretable, black-box models that do not support inferential reasoning. RNNs can be difficult to train due to their recursive structure and struggle with long-term temporal dependencies, which led to the development of numerous variants. One of the most successful and well-known variants is long short-term memory (LSTM) networks inspired by the mechanics of human memory (Hochreiter & Schmidhuber, 1997). LSTMs modify RNNs by adding in the ability for the model to "forget" information, addressing the training stability issues of RNNs while allowing them to better model long-time dependencies. Hong and Wang (2020) used LSTMs to predict drowsy driving based on an individual's facial and steering features. Nguyen et al. (2019) use a modified form of LSTMs to predict surgical skill levels based on their hand motions within a surgical simulation.

Applications

With a growing body of literature concerning the collection and structuring of data, as well as the potential utility of various models, it is important to focus research attention on the pragmatic application of AI/machine learning techniques. To date, high-fidelity simulation (e.g., virtual reality) technology has seen wide adoption in military and medical applications, where environmental realism is worth the high price of technology. However, the last decade has seen a major expansion of virtual reality technology, including a boom in the gaming industry. Like all other technologies, improvements in hardware, a shared programming knowledgebase, and capital investments are making virtual reality technology less expensive and more accessible. Organizations are already using virtual reality for process planning and factory layout planning (Mujber, Szecsi, & Hashmi, 2004; Gong et al., 2019). Organizations can further use simulated environments to allow candidates to virtually tour a work facility with the ability to simulate daily activities and train new employees on the use of heavy equipment like forklifts and transport container cranes without risk of injury (Yuen, Choi, & Yang, 2010; Bruzzone & Longo, 2013; Choi, Ahn, & Seo, 2020).

Meaningful translation of technical advances in computer science to meet the specific contextual needs of human factors research could have a transformative influence on the field. There is increasing need for continued research concerning appropriate model and feature selection in addition to a nuanced understanding of when a model becomes confident enough to provide meaningful and stable feedback

to a user or trainee. Additionally, traditional human factors topics concerning feedback communication should see revitalized attention in the new context of computer-simulated environments. The following sections highlight some important areas for consideration when adopting AI/machine learning techniques for applied settings.

Trainee Feedback

The major benefit of using simulations and simulated environments is the opportunity to provide trainee feedback in a structured, high-fidelity, and safe environment. Feedback is a long-standing topic of research in human factors, providing an abundance of relevant literature while leaving substantial room for new insights. Traditional topics, such as the design and evaluation of warning signals (Wogalter, Conzola, & Smith-Jackson, 2002), have seen revitalized attention given these new applications and contexts. The following sections highlight a few of the upcoming and important topics that are well-suited for examination through a human factors lens.

In general, AI/machine learning techniques have been criticized for lack of decision-making interpretability. Interpretability of model output has particular importance in providing feedback from training simulations (Mirichi et al., 2019). As we have discussed in this chapter, machine learning approaches are well-adapted for identifying meaningful patterns in data; however, deciphering the decision-making process of machine learning models remains difficult. For training feedback to be meaningful, trainees must understand why they received certain scores and how they can improve. Typical black-box approaches are not always well-suited for providing this level of feedback. Nonetheless, researchers have proposed means by which machine learning models can be used to provide meaningful feedback during and after simulations.

Early Prediction

Real-time detection has already demonstrated budding utility for static diagnosis and training simulations, but it also has applications beyond training, such as in computer-enhanced surgical assistance (Thai et al., 2020). In all instances, it is important to consider when and how a system should begin providing feedback to a user or trainee; accordingly, there are two distinct avenues of consideration when deploying this technology. The first is a technical perspective on when a model becomes confident enough to make a stable prediction of a future event or state (i.e., early prediction and model uncertainty). The second is the psychological consideration of how feedback should be delivered to a user or trainee in real time.

Early prediction specifically concerns confidence in prediction when working with temporal data. It concerns the question of when a model's prediction confidence is high enough (or uncertainty low enough) to make a stable prediction. For example, in medical research, early prediction would concern a model's ability to accurately detect an abnormality or disease at early stages of diagnosis or identify symptoms of early onset (e.g., Hsu & Holtz, 2019). Early prediction methods have

shown tremendous utility for enhancing diagnostics across diverse circumstances, including predicting circulatory failure (Hyland et al., 2020), sepsis shock (Lin et al., 2018), and diabetes (Alam et al., 2019).

Real-Time Feedback

Advancements in early prediction pave the road for real-time feedback. Recent findings show tremendous potential for using motion data to provide meaningful feedback to trainees. For example, researchers have determined corollaries for motion inference in games-based settings and provided evidence that motion information could be used for early indications of events (Hart, Vaziri-Pashkam, & Mahadevan, 2020). A natural expansion of this research is the application of advanced machine learning/deep learning models that could infer dangerous motion and provide early warning indications. Researchers have used an adaptation of random forest and LSTM neural network models to create a real-time feedback tool for a temporal bone surgery simulator by identifying characteristics in drilling strokes that improved surgery performance (Ma et al., 2017a, 2017b). More abstractly, researchers have explored machine-learning approaches to predict early warning signs of critical transitions within dynamic systems (Lade & Gross, 2012; Füllsack, Kapeller, Plakolb, and Jäger, 2020). The concept of identifying predictors of critical transitions using machine learning models could have widespread applications for monitoring dynamic systems, such as predicting communication breakdowns in military squadrons or oncoming disequilibrium in a patient during operation.

Early prediction coupled with meaningful feedback would be useful across a variety of training settings, including team communication and education. However, for real-time feedback to be meaningful, it must be effectively received and processed by a user or trainee. Computerized systems should facilitate integration of established best practices, such as the use of personalized warnings (Wogalter, Racicot, Kalsher, & Simpson, 1994), meaningful use of alarms (Edworthy & Hellier, 2006), integration of multisensory warning signals (Ho, Reed, & Spence, 2007; Baldwin et al., 2012), and ensuring that warnings are maximally informative (Fagerlönn & Alm, 2010). Meaningful application of AI to training simulations will require an interdisciplinary perspective that can translate powerful analytics into products with pragmatic value.

Adaptive Simulations

AI's potential to produce real-time prediction and feedback will also enable advancements in adaptive simulations. The goal of AI-enhanced adaptive simulations would be to further increase the fidelity and learning opportunities within simulated training environments. Perhaps the most straightforward application is to adapt the difficulty of a simulation to create additional challenges when trainees are performing well. In addition to adaptive difficulty, simulations could include adaptive scenarios for dangerous events such as hydroplaning, losing control of an object when using machinery like cranes and forklifts, and emergency medical scenarios during

surgery. Allowing simulations to adaptively introduce these events, especially in connection to real-time data about the environmental state, would increase simulation fidelity and better prepare trainees for low-frequency but high-risk events.

A recent review of adaptive simulations found that the most common simulation adaptation was adjustment to difficulty, such as adjustment to speed or resistance in rehabilitation exercises (Zahabi & Razak, 2020). The study also found that most simulations with adaptive content did not provide adaptive real-time feedback. In our discussion, we express the need to provide adaptation of controlled elements in the environment and the usefulness of adaptive feedback. Achieving these goals will require that AI gain sufficient knowledge about current states, desired states, and future states and be able to process this information quickly and efficiently. Codifying the knowledge of when an action should have been performed, and then adaptively providing feedback to redirect toward the desired state, is not a task that can be easily programmed from a flat representation of actions, especially when moving beyond simple motion data. Furthermore, as it pertains to the application of such technologies outside of training simulations, the implementation of AI image (video) information is much more difficult in applied settings than in simulations (Vedula et al., 2017).

CONCLUSION

The implications of AI for society are immense, and such techniques are already being used to change how we think about and score simulations. This chapter has discussed domains where AI can help with two of the greatest challenges in scoring simulations: NLP models for handling unstructured text data and various other techniques for dealing with unstructured data such as event and motion data. While we framed the techniques around real-world examples of their use, it should be noted that we have not tried to be exhaustive in terms of the possible analytic techniques or provided enough information so a reader could jump to using the above analytics directly. Each of these techniques likely deserves an entire chapter devoted to real-world usage, but such detail would be far beyond the scope of this chapter. Our goal was to outline both the implications of these technologies for how we score simulations and exposure to the various analytics techniques that could be, and often already are being, used in the field to score simulations. The truth is that the best techniques at a given time are rapidly changing, and any guide or cookbook for how to use a technique will become dated quickly. So rather than focus on the specifics, we wanted to outline consistent themes and limitations that should be considered whenever AI is being used and describe what may be possible.

A consistent theme across everything discussed in this chapter is that the importance of SMEs, and an understanding of the domain at hand, will continue to be of critical significance. While AI is capable of extraordinary things, getting the most from these tools requires collaboration between data scientists and domain-specific SMEs, as the model is only as effective as the quality of data and simulation being assessed. Broadly speaking, we see two primary areas where AI will forever change how we score simulations. One will be the automation of scoring unstructured text

data that previously would require highly trained human raters, and the other would be new sources of data too complex for humans to process. We argue both areas have promising futures in the realm of simulations, whether for training or evaluative purposes. Lastly, we believe AI will create a golden age in the use of simulations as we are able to create models that help simulations become more accurate, scalable, and comprehensive.

REFERENCES

Ahmidi, N., Poddar, P., Jones, J. D., Vedula, S. S., Ishii, L., Hager, G. D., & Ishii, M. (2015). Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. *International Journal of Computer Assisted Radiology and Surgery*, 10(6), 981–991.

Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., Hussain, A., Malik, M., Raza, M., Ibrar, S., & Abbas, Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, 16, 1–6.

Anh, N. X., Nataraja, R. M., & Chauhan, S. (2020). Towards near real-time assessment of surgical skills: A comparison of feature extraction techniques. *Computer Methods and Programs in Biomedicine*, 187, 105234.

Baldwin, C. L., Spence, C., Bliss, J., Brill, C., Wogalter, M., Mayhorn, C., & Ferris, T. (2012). Multimodal cueing: The relative benefits of the auditory, visual, and tactile channels in complex environments. *Proceedings of the Human Factors and Ergonomics Society*.

Beninger, J., Hamilton-Wright, A., Walker, H. E., & Trick, L. M. (2021). Machine learning techniques to identify mind-wandering and predict hazard response time in fully immersive driving simulation. *Soft Computing*, 25(2), 1239–1247.

Bernardin, H. J., & Buckely, M. R. (1981). Strategies in rater training. *Academy of Management Review*, 6, 205–212.

Boyle, W. A., Murray, D. J., Beyatte, M. B., Knittel, J. G., Kerby, P. W., Woodhouse, J., & Boulet, J. R. (2018). Simulation-based assessment of critical care 'front-line' providers. *Critical Care Medicine*, 46(6), e516.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models as few shot learners. Arxiv: https://arxiv.org/pdf/2005.14165.pdf

Bruzzone, A., & Longo, F. (2013). 3D simulation as training tool in container terminals: The TRAINPORTS simulator. *Journal of Manufacturing Systems*, 32, 85–98.

Choi, M., Ahn, S., & Seo, J. (2020). VR-based investigation of forklift operator situation awareness for preventing collision accidents. *Accident Analysis and Prevention*, 136, 1–9.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies*, 4171–4186. Minneapolis, MN.

Edworthy, J., & Hellier, E. (2006). Alarms and human behaviour: Implications for medical alarms. *British Journal of Anesthesia*, 97(1), 12–17.

Fagerlönn, J., & Alm, H. (2010). Auditory signs to support traffic awareness. *IET Intelligent Transport Systems*, 4(4), 262–269.

Füllsack, M., Kapeller, M., Plakolb, S., & Jäger, G. (2020). Training LSTM-neural networks on early warning signals of declining cooperation in simulated repeated public good games. *MethodsX*, 7, 100920.

Franz, L., Shrestha, Y. R., & Paudel, B. (2020). A deep learning pipeline for patient diagnosis prediction using electronic health records. *BioKDD 2020: 19th International Workshop on Data Mining in Bioinformatics*. San Diego, CA.

Garla, V. N., & Brandt, C. (2012). Ontology-guided feature engineering for clinical text classification. *Journal of Biomedical Informatics*, 45(5), 992–998.

Geden, M., Emerson, A., Rowe, J., Azevedo, R., & Lester, J. (2020). Predictive student modeling in educational games with multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1), 654–661.

Gibson, C., & Mumford, M. D. (2013). Evaluation, criticism, and creativity: Criticism content and effects on creative problem-solving. *Psychology of Aesthetics, Creativity, and the Arts*, 7, 314–331.

Gong, L., Berglund, J., Berglund, A., Johansson, B., & Borjesson, T. (2019). Development of virtual reality support to factory layout planning. *International Journal of Interactive Design and Manufacturing*, 13, 935–945.

Hall, S., & Brannick, M. T. (2008). Performance assessment in simulation. In D. A. Vincenzi, J. A. Wise, M. Mouloua, & P. A. Hancock (Eds.), *Human factors in simulation and training* (pp. 149–168). Boca Raton, FL: CRC Press.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, 46, 43–58.

Hart, Y., Vaziri-Pashkam, M., & Mahadevan, L. (2020). Early warning signals in motion inference. *PLoS Computational Biology*, 16(5), e1007821.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. Cambridge, MA. The MIT Press.

Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., … Lester, J. (2020). Enhancing student competency models for game-based learning with a hybrid stealth assessment framework. *International Educational Data Mining Society*, 13, 92–103.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation* 9(8), 1735–1780.

Ho, C., Reed, N., & Spence, C. (2007). Multisensory in-car warning signals for collision avoidance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(6), 1107–1114.

Hong, L., & Wang, X. (2020). Towards drowsiness driving detection based on multi-feature fusion and LSTM networks. *International Conference on Control, Automation, Robotics and Vision*,732–736.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18, 500–510.

Hsu, P., & Holtz, C. (2019). A comparison of machine learning tools for early prediction of sepsis from ICU data. *2019 Computing in Cardiology (CinC)*, 46, 1–4.

Huang, K., Altosaar, J., & Ranganath, R. (2020). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. Arxiv: https://arxiv.org/pdf/1904.05342.pdf

Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., Bock, C., Horn, M., Moor, M., Rieck, B., Zimmermann, M., Bodenham, D., Borgwardt, K., Rätsch, G., & Merz, T. M. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3), 364–373.

Krajewski, J., Sommer, D., Trutschel, U., Edwards, D., & Golz, M. (2009). Steering wheel behavior based estimation of fatigue. *Proceedings of the International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 118–124.

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Abingdon, UK: Taylor & Francis Group.

Lade, S. J., & Gross, T. (2012). Early warning signals for critical transitions: A generalized modeling approach. *PLoS Computational Biology*, 8(2), e1002360.

Loukas, C., & Georgiou, E. (2011). Multivariate autoregressive modeling of hand kinematics for laparoscopic skills assessment of surgical trainees. *IEEE Transactions on Biomedical Engineering*, 58(11), 3289–3297.

Lin, C., Zhang, Y., Ivy, J., Capan, M., Arnold, R., Huddleston, J. M., & Chi, M. (2018). Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM. *Proceedings of the International Conference on Healthcare Informatics*, 219–228.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. Arxiv: https://arxiv.org/pdf/1907.11692.pdf

Ma, X., Wijewickrema, S., Zhou, Y., Zhou, S., O'Leary, S., & Bailey, J. (2017). Providing effective real-time feedback in simulation-based surgical training. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 566–574.

Ma, X., Wijewickrema, S., Zhou, Y., Zhou, S., Mhammedi, Z., O'Leary, S., & Bailey, J. (2017b). Adversarial generation of real-time feedback with neural networks for simulation-based training. *Proceedings of the International Joint Conference on Artificial Intelligence*, 3763–3769.

McDonald, A. D., Lee, J. D., Schwarz, C., & Brown, T. L. (2014). Steering in a random forest: Ensemble learning for detecting drowsiness-related lane departures. *Human Factors*, 56(5), 986–998.

Mills, J. T., Hougen, H. Y., Bitner, D., Krupski, T. L., & Schenkman, N. S. (2017). Does robotic surgical simulator performance correlate with surgical skill? *Journal of Surgical Education*, 74(6), 1052–1056.

Min, W., Mott, B., Rowe, J., Taylor, R., Wiebe, E., Boyer, K., & Lester, J. (2017). Multimodal goal recognition in open-world digital games. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 13(1), 80–86.

Mirichi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., & Del Maestro, R. (2019). The virtual operative assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS One*, 15(2), 1–15.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics Magazine*, 38(8), 114–117.

Mosier, K. L., & Manzey, D. (2020). Humans and automated decision aids: A match made in heaven? In M. Mouloua & P. A. Hancock (Eds.), *Human performance in automated and autonomous systems: Current theory and methods* (pp. 19–42). Boca Raton: CRC Press.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, 75, 640–647.

Mouloua, M., & Hancock, P. (2020). *Human performance in automated and autonomous systems: Current theory and methods*. Boca Raton: CRC Press.

Mracek, D. L., Peterson, N., Barsa, A., & Koenig, N. (2021). DEEP*O*NET: A neural network approach to leveraging detailed text descriptions of the world of work. In K. Nei (Chair), *Demonstrating natural language processing for improving job analysis*. Symposium conducted at the meeting of the Society for Industrial/Organizational Psychology, New Orleans, LA.

Mujber, T., Szecsi, T., & Hashmi, M. (2004). Virtual reality applications in manufacturing process simulation. *Journal of Materials Processing Technology*, 155–156, 1834–1838.

Nguyen, X. A., Ljuhar, D., Pacilli, M., Nataraja, R. M., & Chauhan, S. (2019). Surgical skill levels: Classification and analysis using deep neural network model and motion signals. *Computer Methods and Programs in Biomedicine*, 177, 1–8.

Norman, G. R., Grierson, L. E. M., Sherbino, J., Hamstra, S. J., Schmidt, H. G., & Mamede, S. (2018). Expertise in medicine and surgery. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *Cambridge handbooks in psychology: The Cambridge handbook of expertise and expert performance* (pp. 331–355). Cambridge: Cambridge University Press.

Oquendo, Y. A., Riddle, E. W., Hiller, D., Blinman, T. A., & Kuchenbecker, K. J. (2018). Automatically rating trainee skill at a pediatric laparoscopic suturing task. *Surgical Endoscopy*, 32(4), 1840–1857.

Parasuraman, R., & Mouloua, M. (1996). *Automation and human performance: Theory and applications*. New York: CRC Press.

Poola, I. (2017). How artificial intelligence is impacting real life every day. *International Journal of Advance Research, Ideas and Innovations in Technology*, 2., 96–100.

Richstone, L., Schwartz, M. J., Seideman, C., Cadeddu, J., Marshall, S., & Kavoussi, L. R. (2010). Eye metrics as an objective assessment of surgical skill. *Annals of Surgery*, 252(1), 177–182.

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, 85, 370–395.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*. California Univ San Diego La Jolla Inst for Cognitive Science. Cambridge, MA: Bradford Books/MIT Press.

Rust, R. T., & Huang, M. (2014). The service revolution and transformation of marketing science. *Marketing Science*, 33, 206–221.

Ryman-Tubb, N. F., Krause, P., & Garn, W. (2018). How artificial intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. *Engineering Applications of Artificial Intelligence*, 76, 130–157.

Salkowski, L. R., & Russ, R. (2018). Cognitive processing differences in experts and novices when correlating anatomy and cross-sectional imaging. *Journal of Medical Imaging*, 5(3), 031411.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (1999). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Schwab, K. (2017). *The fourth industrial revolution*. New York: World Economic Forum.

Schwarting, W., Alonso-Mora, J., & Rus, D. (2018). Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1, 187–210.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *Proceedings of the IEEE International Conference on Computer Vision*, 843–852.

Sydell, E., Ferrell, J., Carpenter, J., Frost, C., & Brodbeck, C. C. (2013). Simulation scoring. In M. Fetzer & K. Tuzinski (Eds.), *Simulations for personnel selection* (pp. 83–107). New York, NY: Springer.

Thai, M., Phan, P., Hoang, T., Wong, S., Lovell, N., & Do, T. (2020). Advanced intelligent systems for surgical robotics. *Advanced Intelligent Systems*, 2, 1–33.

Thomson, D. R., Besner, D., & Smilek, D. (2015). A resource-control account of sustained attention: Evidence from mind-wandering and vigilance paradigms. *Perspectives on Psychological Science*, 10(1), 82–96.

Thompson, I., Koenig, N., Mracek, D. L., & Tonidandel, S. (forthcoming). Integrating deep learning and measurement science: Automating the subject matter expertise used to evaluate candidate work samples. *Journal of Applied Psychology*.

Tonidandel, S., Thompson, I. B., Mracek, D. L., & Koenig, N. (2020). Automating subject matter expertise used to evaluate candidate work samples. In E. Campion & M. Campion, (Chairs). *The Construct Validity of Computer-Assisted Text Analysis (CATA)*. Symposium conducted at the annual conference of the Society for Industrial and Organizational Psychology, Austin, TX.

Uzuner, Ö. (2009). Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4), 561–570.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomex, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing*, 6000–6010.

Vedula, S. S., Ishii, M., & Hager, G. D. (2017). Objective assessment of surgical technical skill and competency in the operating room. *Annual Review of Biomedical Engineering*, 19, 301–325.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., … Tang, X. (2017). Residual attention network for image classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Wei, W. W. (2006). Time series analysis. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology: Vol. 2*. Oxford, UK: Oxford University Press.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2016). *Engineering psychology and human performance* (4th ed). New York: Routledge.

Wogalter, M., Conzola, V., & Smith-Jackson, T. (2002). Research-based guidelines for warning design and evaluation. *Applied Ergonomics*, 33(3), 219–230.

Wogalter, M., Racicot, B., Kalsher, M., & Simpson, S., (1994). Personalization of warning signs: The role of perceived relevance on behavioral compliance. *International Journal of Industrial Ergonomics*, 14(3), 233–242.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82.

Yuen, K., Choi, S., & Yang, X. (2010). A full-immersive CAVE-based VR simulation system of forklift truck operations for safety training. *Computer-Aided Design & Applications*, 7(2), 235–245.

Zahabi, M., & Razak, A. M. A. (2020). Adaptive virtual reality-based training: A systematic literature review and framework. *Virtual Reality*, 24, 725–752.

Zepf, S., Stracke, T., Schmitt, A., van de Camp, F., & Beyerer, J. (2019, December). Towards real-time detection and mitigation of driver frustration using SVM. *Proceedings of the International Conference on Machine Learning and Applications*, 202–209.

Zhang, Y., Jin, R., & Zhou, Z. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1, 43–52.