CHAPTER 5

Evaluating Learning for Intelligence

"Intelligence is the ability to adapt to change"

—Stephen Hawking

As a field, machine learning is still in its infancy. Advanced machine learning has only been explored over the last 25 years, which has fueled data science as a profession. As a result, the data science industry is still in a phase of wonderment at the endless potential of AI and machine learning. With this comes both excitement and confusion—and an industry that is gathering knowledge, experience, and first-time problems.

Typically, the most laborious tasks within a machine learning project are identifying the appropriate model and engineering features, which make a substantial difference to the output of the model. In fact, the features chosen can often have more impact on the quality of a model compared to the model choice itself. Therefore, it is important to evaluate the learning algorithm that will determine the model's intelligence to predict the output of an unknown sample. This is usually done using various metrics, which are discussed further.

A. Panesar, *Machine Learning and AI for Healthcare*,
https://doi.org/10.1007/978-1-4842-3799-1_5

Model Development and Workflow

To successfully deploy a machine learning model, there are several stages of development and evaluation that take place, as illustrated in Figure 5-1.

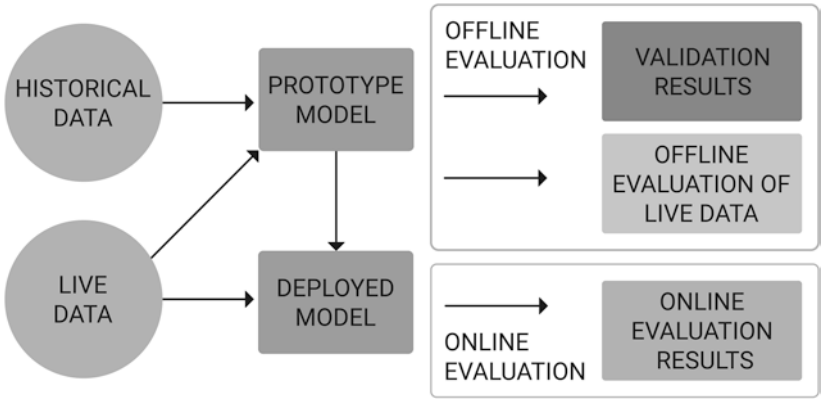

Figure 5-1. *Model development and work flow*

The first stage is the prototype phase. During this phase, a prototype is created through testing various models on historical data to determine the best model. Hyperparameter tuning, as discussed later in this chapter, is a requirement of model training. Once the best prototype model is chosen, the model is tested and validated. Validating a model requires splitting datasets into training, testing, and validation sets as discussed in Chapter 3. Consider the fact that there is no such thing as a random dataset and instead the randomness applies to the splitting of the dataset. Be aware of biases that may appear in the data. Once the model has been successfully validated, it is deployed to production. The model is then usually evaluated by one (or several) performance metrics.

There are two ways of evaluating a machine learning model: offline evaluation and online (or live) evaluation.

Why Are There Two Approaches to Evaluating a Model?

A deployed machine learning model consumes data from two sources: historical data (or the data that is used as the experience to be learned from) and live data. Many machine learning models assume stationary distribution data—that the data distribution is constant over time. However, this is atypical of real life, as distributions of data often change over time—known as a distribution shift. For instance, consider a system that predicts the side effects of medications to patients based on their health profile. Medication side effects may change based on population factors such as ethnicity, disease profile, territory, medication popularity, and new medications. The distribution of relevant side effects based on patient data can vary quickly over time, and hence it is essential for a model to detect a shift in distribution and accordingly evolve the model. The method in which this is typically assessed is through the performance of the model based on live data, evaluated through the validation metric used in the testing and validation of the model on historical data.

Model performance that is similar to or within a threshold of permissibility when evaluated on live data is deemed as a model that continues to fit the data. Degradation of model performance indicates that the model does not fit the data and requires retraining.

Offline evaluation measures the model based on metrics learned and evaluated from the historical, stationary, distributed dataset. Metrics such as accuracy and precision-recall are typically used within the offline training stage. Offline evaluation techniques include the hold-back method and n-fold cross-validation.

Online evaluation refers to the evaluation of metrics once the model is deployed. The key takeaway is that these metrics may differ from the metrics used to evaluate performance when the model is deployed live. For instance, a model that is learning on new pharmacological treatments may seek to be as precise as possible in training and validation; but when placed online, it may need to consider business goals such as budget or treatment value when deployed. Online evaluation, particularly in the digital age, can support multivariate testing to understand best-performing models.

Feedback loops are key to ensuring systems are performing as intended and help to understand the model in the context of use better. This can be performed by a human agent or automated through a contextually intelligent agent or users of the model.

It is important that the evaluation of a machine learning model is based on a statistically independent dataset and not on the dataset it is trained on. This is because the evaluation of the training dataset is optimistic about the model's true performance as it adapts to the dataset. By evaluating the model with previously unseen data, there is a better estimate of the generalization error. New data can be hard to find; hence it is important to be able to have new, unseen data from the current dataset. Methods such as n-fold cross-validation discussed in Chapter 3 are useful techniques for this purpose. Often the data used is more important than the algorithm choice; and the better the features used, the greater the performance of the model.

The evaluation metrics discussed can be found in the metrics package for R and scikit-learn for Python.

Evaluation Metrics

There are a plethora of evaluation metrics for machine learning problems. Metrics exist for the variety of machine learning tasks—classification, regression, clustering, association rule mining, NLP, and so on.

Classification

Classification problems seek to give a label or classification to an input. There are several methods by which to measure performance, including accuracy, precision-recall, confusion matrices, log-loss (logarithmic loss), and AUC (area under the curve).

Accuracy

Accuracy is the simplest technique used in identifying whether a model is making correct predictions. It is calculated as a percentage of correct prediction over the total predictions made.

Accuracy = number of correct predictions/number of total predictions

Confusion matrix

Accuracy is a general metric that does not consider the division between classes. Therefore, it does not consider misclassification or the associated penalty with misclassification. For instance, a medical misdiagnosis that is a false positive (e.g., take a patient diagnosed with breast cancer when they do not have it) has substantially different consequences compared to a false negative, whereby a patient is told that they do not have breast cancer when in fact they do. A confusion matrix breaks down the correct and incorrect classifications made by the model and attributes them to the appropriate label.

- True positive: Where the actual class is yes, and the value of the predicted class is also yes.
- False positive: Actual class is no, and predicted class is yes

- True negative: The value of the actual class is no, and the value of the predicted class is no
- False negative: When the actual class value is yes, but predicted class is no

Take an example whereby a model predicts whether a patient has breast cancer or not based on 50 example inputs from the test dataset with an equal distribution between positive and negative labeled examples. The confusion matrix would be as in Table 5-1.

Table 5-1. *Confusion matrix*

	Prediction: Positive	Prediction: Negative
Labeled positive	20	5
Labeled negative	15	10

From the confusion matrix, it is determined that the positive class has greater accuracy than the negative class. The accuracy of the positive classification is 20/25 = 80%. The negative class has an accuracy of 10/25 = 40%. Both metrics differ from the overall accuracy of the model, which would be determined as (20 + 10)/50 = 60%. It is apparent how a confusion matrix adds more detail to the overall accuracy of a machine learning model.

As a result, accuracy can be rewritten as the following:

Accuracy = (correctly predicted observation)/(total observation) = (TP + TN)/(TP + TN + FP + FN)

Per-class accuracy

Per-class accuracy is an extension of accuracy that takes into account the accuracy of each class. As a result, the preceding example has a per-class accuracy of (80% + 40%)/2 = 60%. Per-class accuracy is useful

in distorted problems where there are a larger number of examples within one particular class compared to another. The class with greater examples dominates the calculation, and therefore accuracy alone may not suffice for the nature of your model; thus it is useful to evaluate per-class accuracy also.

Logarithmic loss

Logarithmic loss (or log-loss for short) is used for problems where a continuous probability is predicted rather than a class label. Log-loss provides a probabilistic measure of the confidence of the accuracy and considers the entropy between the distribution of true labels and predictions.

For a binary classification problem, the logarithmic loss would be calculated as follows:

$$\text{Log-loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1-y_i)\log(1-p_i)$$

Where P_i is the probability of the *i*th data point belonging to a class and y_i the true label (either 0 or 1).

Area Under the Curve (AUC)

The AUC plots the rate of true positives to the rate of false positives. The AUC enables the visualization of the sensitivity and specificity of the classifier. It highlights how many correct positive classifications can be gained allowing for false positives.

The curve is known as the receiver operating characteristic curve, or ROC as shown in Figure 5-2. A high AUC or greater space underneath the curve is good, and a smaller area under the curve (or less space under the curve) is undesirable. In Figure 5-2, test A has better AUC as compared to test B, as the AUC for test A is larger than for test B. The ROC visualizes the trade-off between specificity and sensitivity of the model.

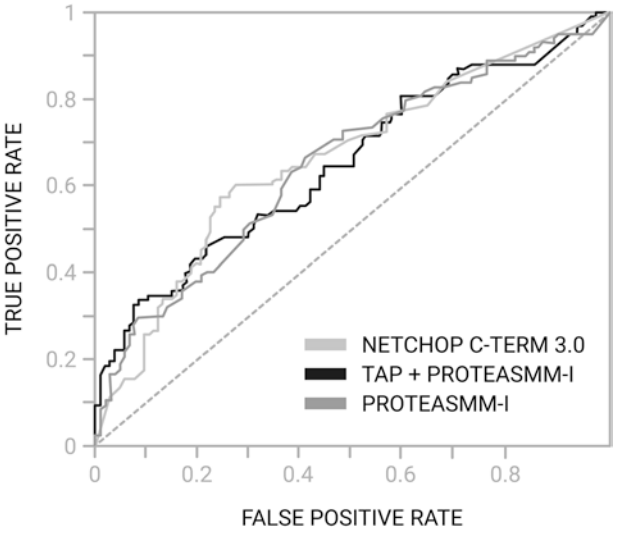

Figure 5-2. *ROC curve*

Precision, recall, specificity, and F-measure

Precision and recall are two metrics used together to evaluate model performance. Precision evaluates how many items are truly relevant compared to the total number of items correctly classified. Recall evaluates how many items are predicted to be relevant by the model from the items that are relevant.

- Precision: (correctly predicted Positive)/(total predicted Positive) = TP/TP + FP
- Recall: (correctly predicted Positive)/(total correct Positive observation) = TP/TP + FN

Specificity refers to how well the model performs at returning incorrect classifications and is calculated as in Figure 5-3.

- Specificity: (correctly predicted Negative)/(total Negative observation) = TN/TN + FP

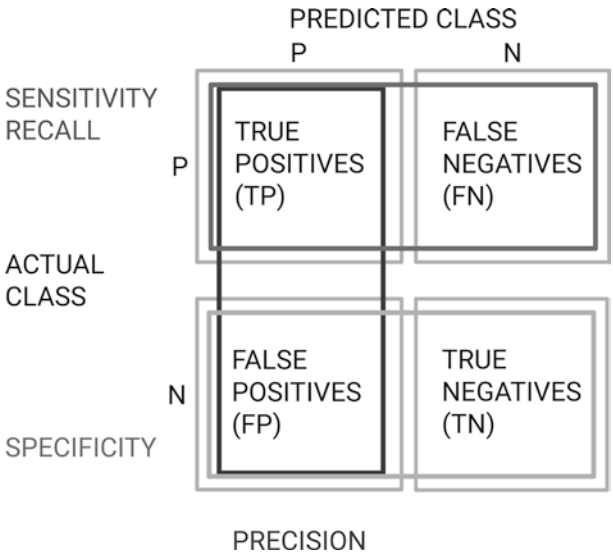

Figure 5-3. *Specificity classification diagram*

F-measure goes beyond the arithmetic mean and calculates the harmonic mean of precision and recall:

$$F = \frac{1}{\frac{1}{2}\left(\frac{1}{p}+\frac{1}{r}\right)} = \frac{2pr}{p+r}$$

Where p denotes precision and r denotes recall.

Regression

Regression machine learning models output continuous variables, and root-mean-squared error (RMSE) is the most commonly used evaluation metric for these problems.

RMSE

RMSE calculates the square root of the sum of the average distance between predicted and actual values. This can also be understood as the average Euclidean distance between the true value and predicted value vectors. A criticism of RMSE is that it is sensitive to outliers.

$$RMSE = \sqrt{\frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{2}},$$

where y_i denotes the actual value and $\hat{y}_i$ denotes predicted value.

Percentiles of errors

Percentiles (or quantiles) of error are more robust as a result of being less sensitive to outliers. Real-world data is likely to contain outliers, and thus it is often useful to look at the median absolute percentage error (MAPE) rather than the mean.

$$MAPE = median\left(\left|\left|\left(y_i - \hat{y}_i\right) / \left(y_i\right)\right|\right|\right),$$

where y_i denotes the actual value and $\hat{y}_i$ denotes predicted value.

The MAPE is less affected by outliers by using the median of the dataset. A threshold or percentage difference for predictions can be set for a given problem to give an understanding of the precision of the regression estimate. The threshold depends on the nature of the problem.

Skewed Datasets, Anomalies, and Rare Data

An experienced data scientist treats all data with suspicion. Data can be inconsistent; and as a result, skewed datasets, imbalanced class examples, and outliers can all significantly affect the performance of a model. Having more examples within one class compared to another can lead to an underperforming model. Furthermore, outliers or data anomalies can further skew performance evaluation metrics. The effect of large outliers can be mitigated using percentiles of error. In practice, good data cleansing, removal of outliers, and normalization of variables can reduce the sensitivity to outliers.

Parameters and Hyperparameters

Hyperparameters and parameters are often used interchangeably, yet there is a difference between the two. Machine learning models can be understood as mathematical models that represent the relationship between aspects of data.

Model parameters are properties of the training dataset that are learned and adjusted during training by the machine learning model. Model parameters differ for each model, dataset properties, and the task at hand. For instance, in the case of an NLP predictor that output the sophistication of a corpus of text, parameters such as word frequency, sentence length, and noun or verb distribution per sentence would be considered model parameters.

Model hyperparameters are parameters to the model building process that are not learned during training. Hyperparameters can make a substantial difference to the performance of a machine learning model. Hyperparameters define the model architecture and effect the capacity of the model, influencing model flexibility.

Hyperparameters can also be provided to loss optimization algorithms during the training process. Optimal setting of hyperparameters can have a significant effect on predictions and help prevent a model from overfitting. Optimal hyperparameters often differ between datasets and models.

In the case of a neural network, for example, hyperparameters would include the number and size of hidden layers, weighting, learning rate, and so forth. Decision trees hyperparameters would include the desired depth and number of leaves in the tree. Hyperparameters with a support vector machine would include a misclassification penalty term.

Tuning Hyperparameters

Hyperparameter tuning or optimization is the task of selecting a set of optimal hyperparameters for a machine learning model. Optimized hyperparameters values maximize a model's predictive accuracy. Hyperparameters are optimized through running training a model, assessing the aggregate accuracy, and appropriately adjusting the hyperparameters. Through trialing a variety of hyperparameter values, the best hyperparameters for the problem are determined, which improves overall model accuracy.

Hyperparameter Tuning Algorithms

Hyperparameter tuning is like training a machine learning model. The task at hand is one of optimization. Model parameters can be expressed as a loss function, whereas hyperparameters cannot be expressed as such, as it depends entirely on the model training process. There are several approaches to hyperparameter tuning, with the most common being grid search and random search.

Grid Search

The grid search is a simple, effective, yet resource expensive hyperparameter optimization technique that evaluates a grid of hyperparameters. The method evaluates each hyperparameter and determines the winner. For example, if the hyperparameter were the number of leaves in a decision tree, which could be anywhere from n = 2 to 100, grid search would evaluate each value of n (i.e., points on the grid) to determine the most effective hyperparameter.

It is often a case of guessing where to start with hyperparameters, including minimum and maximum values. The approach is typical of trial and error, whereby if the optimal value lies toward either maximum or minimum, the grid would be expanded in the appropriate direction in an attempt to further optimize the model's hyperparameters.

Random Search

Random search is a variant of grid search that evaluates a random sample of grid points. Computationally, this is far less expensive than a standard grid search. Although at first glance it would appear that this is not as useful in finding optimal hyperparameters, Bergstra et al. demonstrated that in a surprising number of instances, a random search performed roughly as well as grid search.[65] The simplicity and better-than-expected performance of a random search means that it is often chosen over grid search. Both grid search and random search are parallelizable.

More intelligent hyperparameter tuning algorithms are available that are computationally expensive as the result of evaluating which samples to try next. These algorithms often have hyperparameters of their own. Bayesian optimization, random forest smart tuning, and derivative-free optimization are three examples of such algorithms.

Multivariate Testing

Multivariate testing is an extremely useful method of determining which model is best for the particular problem at hand. Multivariate testing is known as statistical hypothesis testing and determines the difference between a null hypothesis and alternative hypothesis. The null hypothesis is defined as the new model not affecting the average value of the performance metric; whereas the alternate hypothesis is that the new model does change the average value of the performance metric.

Multivariate testing compares similar models to understand which is performing best or compares a new model against an older, legacy model. The respective performance metrics are compared, and a decision is made on which model to proceed with.

The process of testing is as follows:

1. Split the population into randomized control and experimentation groups.
2. Record the behavior of the populations on the proposed hypotheses.
3. Compute the performance metrics and associated *p*-values.
4. Decide on which model to proceed with.

Although the process seems relatively simple, there are a few key aspects for consideration.

Which Metric Should I Use for Evaluation?

Choosing the appropriate metric to evaluate your model depends on the use case. Consider the impact of false positives, false negatives, and the consequences of such predictions. Furthermore, if a model is attempting

to predict an event that only happens 0.001% of the time, an accuracy of 99.999% can be reported but not confirmed. Build the model to cater to the appropriate metrics.

One approach is to repeat the experiment, thus performing repeat evaluations. Although not a fail-safe, this reduces the change of illusionary results. If there is indeed change between the null and alternate hypothesis, the difference will be confirmed.

Correlation Does Not Equal Causation

The phrase correlation does not equal causation is used to stress that a correlation between two variables does not suggest that one causes the other. Correlation refers to the size and direction of a relationship between two or more variables.

Causation, also known as cause and effect, emphasizes that the occurrence of one event is related to the presence of another event. It may be tempting to assume that one variable causes the other; however, in models with several features, there may be hidden factors that cause both variables to move in tandem.

For instance, smoking tobacco is a cause that increases the risk of developing a variety of cancers. However, it may be correlated with alcoholism, but it does not cause alcoholism.

What Amount of Change Counts as Real Change?

Defining the amount of change required before the null hypothesis is rejected once again depends on the use case. Specify a value at the beginning of the project that would be satisfactory and adhere to it.

Types of Tests, Statistical Power, and Effect Size

There are two main types of tests—one-tailed and two-tailed tests. One-tailed tests evaluate whether the new model is better than the original. However, it does not specify whether the model is worse than the baseline. One-tailed tests are thus inherently biased. With two-tailed tests, the model is tested for the possibility of change in two directions—positive and negative.

Statistical power refers to the probability that the difference detected during the testing reflects a real-world difference.

Effect size determines the difference between two groups through evaluating the standardized mean difference between two sets. Effect size is calculated as the following:

Effect size = ((mean of experiment group) – (mean of control group))/ standard deviation

Checking the Distribution of Your Metric

Many multivariate tests use the *t*-test to analyze the statistical difference between means. The *t* value evaluates the size of the difference relative to the variation in your sample data. However, the *t*-test makes assumptions that are not necessarily satisfied by all metrics. For instance, the *t*-test assumes both sets have a normal, or Gaussian, distribution.

If the distribution does not appear to be Gaussian, select a nonparametric test that does not make assumptions about a Gaussian distribution, such as the Wilcoxon–Mann–Whitney test.

Determining the Appropriate *p* Value

Statistically speaking, the *p* value is a calculation used in hypothesis testing that represents the strength of the evidence. The *p* value measures the statistical significance, or probability, that a difference would arise

by chance given there was no real difference between two populations. It provides the evidence against the null hypothesis and is a useful metric for stakeholders to draw conclusions from.

A *p* value lies between 0 and 1, and is interpreted as follows:[66]

- a *p* value of ≤ 0.05 indicates strong evidence against the null hypothesis, thus rejecting the null hypothesis
- a *p* value of > 0.05 indicates weak evidence against the null hypothesis, hence maintaining the null hypothesis
- a *p* value near 0.05 is considered marginal and could swing either way

The smaller the *p* value, the smaller the probability that the results are down to chance.

How Many Observations Are Required?

The quantity of observations required is determined by the statistical power demanded by the project. Ideally, this should be determined at the beginning of the project.

How Long to Run a Multivariate Test?

The duration of time required for your multivariate testing is ideally the amount of time required to capture enough observations to meet the defined statistical power. It is often useful to run tests over time to capture a representative, variable sample.

When determining the duration of your testing phase, consider the novelty effect, which describes how user reactions in the short term are not representative of the long-term reactions. For instance, whenever Facebook updates their news feed layout or design, there is an uproar.

However, this soon subsides once the novelty effect has worn off. Therefore, it is useful to run your experiment for long enough to overcome this bias. Running multivariate tests for long periods of time are typically not a problem in model optimization.

Data Variance

The control and experimentation sets could be biased as the result of not being split at random. This may result in biases in the sample data. If this is the case, other tests can be used, such as Welch's t-test, which does not assume equal variance.

Spotting Distribution Drift

It is key to measure ongoing performance of your machine learning model once deployed. Data drifts and system development require the model to be confirmed against the baseline. Typically, this involves monitoring the offline performance, or validation metric, against data from the live, deployed model. If there is a sizeable change in the validation metric, this highlights the need to revise the model through training on new data. This can be done manually or automated to ensure consistent reporting and confidence in the model.

Keep a Note of Model Changes

Keep a log of all changes to your machine learning model with notes on changes. Not only does this serve as a change log for stakeholders, it provides a physical record of how the system has changed over time. The use of versioning software within a development enviornment (test/staging to live deployment) will enable software changes to automatically be noted. Versioning software provides a form of technical governance and can be used to deploy software with extensive rollback and backup facilities.

CHAPTER 6

Ethics of Intelligence

> *"People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world."*
>
> —Pedro Domingos

From supermarket checkouts to airport check-ins and digital healthcare to Internet banking, the use of data and AI for decision-making is ubiquitous. There has been an astronomic growth in data availability over the last two decades, fueled by, first, connectivity, and now the Internet of Things. Traditional data science teams focus on the use of data for the creation, implementation, validation, and evaluation of machine learning models that can be for predictive analytics.

In the past decade, there has been a wealth of data-driven AI and technological advance in healthcare, including the following:

- Noninvasive blood glucose levels: There have been many advances in noninvasive blood glucose management. Google developed a contact lens that determines the user's blood glucose levels in 2016. French company PKvitality developed a small wristwatch that determines the level of glucose in the blood, and there have been advances in determining blood glucose levels through using far-infrared signals.[67]

A. Panesar, *Machine Learning and AI for Healthcare*,
https://doi.org/10.1007/978-1-4842-3799-1_6

- Artificial pancreas: As a combination of systems, the artificial pancreas uses two devices. The first device measures blood glucose using a sensor and communicates with a secondary device (or patch), which administers insulin to the patient. The insulin delivery system can adjust the insulin dosage according to the blood glucose levels.[68]
- Bioprinting of skin constructs: 3-D printing has been used to reproduce blood vessels and skin cells to facilitate wound healing for burn patients.[69]
- Digital peer support: Peer support communities such as Diabetes.co.uk enable peer-to-peer support have demonstrated in research studies the ability to improve qualitative and quantitative health outcomes.[70]
- Foot ulcer detection: Concerns such as foot ulceration, bruising, and wider diabetic foot concerns can be increasingly detected through machine learning to expedite ulceration detection, prevent amputations, determine effective treatments, and improve healing times.
- Open source data sharing:[71] Data repositories and public APIs have enabled data sharing between systems. Historically conservative firms are now embracing open source analytical, AI, and data management software. Within many organizations, employees are actively discouraged, or given autonomy on whether to use proprietary tools. Cost and performance are drivers toward open source data, mainly as open data sources have become more robust

> and accepted. The adoption is also driven by the latest generation of data scientists and university graduates aware of the safety and capabilities of open source data.

Novel solutions to problems developed through machine learning are themselves leading to questions of morality and ethics. Currently, governance is moving at the pace of the industry itself. There are many scenarios within AI for which there are no precedents, regulations, or laws. As a result, it is paramount to consider the ethical and moral implications of creating intelligent systems.

The World Health Organization (WHO) reports that the prevalence of chronic disease will rise to 57% of the global population by 2050.[72] Unfortunately, the WHO also reports that there is a global, growing shortage of healthcare workers, which by 2035 will rise to 12.9 million.[73] Lack of professionals to provide healthcare services paints a stark image of the future with grave consequences for humanity. Healthcare professional shortages are being offset through AI, digital interventions, IoT, and other digital technologies that cannot only replace manual and cognitive working tasks but can also improve the reach, precision, and availability of healthcare. At the same time, advancements in the detection and diagnosis of diseases, genomics, pharmacology, stem cell and organ therapy, digital healthcare, and robotic surgery are expected to minimize the cost of treating illness and disease. As AI penetrates humanity's day-to-day activities, philosophical, moral, ethical, and legal questions are raised. This is amplified in healthcare, where clinical decisions can mean the difference between life and death. Even if AI can aid diagnosis of conditions or predict future mortality risk, will humans ever prefer an AI's advice over their doctor? As humankind becomes accustomed to living side-by-side to intelligent systems, there are a host of hurdles to overcome.

What Is Ethics?

Ethics or moral philosophy refers to the moral codes of conduct (or set of moral principles) that shape the decisions people make and their conduct. Morality refers to the principles that distinguish between good/right or bad/wrong behavior. Ethics in the workplace, for example, is often conveyed through professional codes of conduct for which employees must abide.

What Is Data Science Ethics?

Data science ethics is a branch of ethics that is concerned with privacy, decision-making, and data sharing.

Data science ethics comprises three main strands:

- Ethics of data

  This area of data science ethics focuses on the generation, collection, use, ownership, security, and transfer of data.

- Ethics of intelligence

  This area of data science ethics covers the output or outcomes from predictive analytics that data is used to develop.

- Ethics of practices[74]

  The ethics of practices was proposed by Floridi and Taddeo, referring to the morality of innovation and systems to guide emerging concerns.

Data Ethics

More smartphones exist in the world than people—and phones, tablets, and digital devices alongside apps, wearables, and sensors are creating millions of data points a day. There are over 7.2 billion phones in use,

112 million wearable devices sold annually, and over 100,000 healthcare apps available to download on your mobile phone.[75] IBM reports more than 2.5 quintillion bytes (2.5×10^{18}) of data are created daily.[76] Data is everywhere. Moreover, it's valuable.

The topic of data ethics has been thrust into the public spotlight through high profile fiascos such as the Facebook Cambridge Analytica scandal. Facebook, one of the world's largest and most trusted data collection organizations, had user data harvested through a quiz hosted on its platform. Behavioral and demographic data of the 1.5 million completers of the quiz was sold to Cambridge Analytica. The data is largely considered to have been used to target and influence the outcome of the United States 2017 elections.[77] What's more concerning is that this breach of security was reported over 2 years after the initial data leak.

We are in a time where fake news can travel quicker than the truth. Society is at a critical point in its evolution, where the use, acceptance, and reliance on data must be addressed collectively to develop conversations and guiding principles on how to handle data ethically.

The ethical and moral implications of data use are vast, and best demonstrated through an example. During this chapter, we will refer to the following hypothetical scenario:

SCENARIO A

John, type 2 diabetic, aged 30, has a severe hypoglycemic episode and is rushed to the hospital.

John has been taken to the hospital unconscious for treatment. John, a truck driver, was prescribed insulin to treat his type 2 diabetes by his doctor's advice, which is most likely the cause of his hypoglycemia. Lots of data is generated in the process—both in the hospital, by healthcare professionals, and also on John's Apple Watch—his heart rate, heart rate variability, activity details, and blood oxygen saturation to look out for signs of a diabetic coma.

John's blood sample is also taken, and his genome identified. Let's suppose all this data is used, and it's useful.

John's Apple Watch was used to monitor his heart on the way to the emergency room, through which it was suspected that he has an irregular heartbeat, later confirmed with the hospital's medical equipment.

Upon waking from his hypoglycemic episode, John is pleased to learn that genetic testing does not always give bad news, as his risks of contracting prostate cancer is reduced because he carries low-risk variants of the several genes known in 2018 to contribute to these illnesses. However, John is told of his increased risks of developing Alzheimer's disease, colon cancer, and stroke from his genetic analysis.

Informed Consent

Informed consent refers to the user (or patient) being aware of what their data will be used for. Informed consent refers to an individual being legally able to give consent. Typically this requires an individual to be over 18, of sound mind, and able to exercise choice. Consent should ideally be voluntary.

Scenario A demonstrates how useful data can be given a particular context (or use case) and demonstrates the many intricacies of informed consent.

Freedom of Choice

Freedom of choice refers to the autonomy to decide whether your data is shared and with whom. This refers to the active decision to share your data with any third party. For example, should John, who has type 2 diabetes, now be required to demonstrate that his blood glucose levels are under

control and within the recommended range before being allowed to drive his truck again? A person's choice as to whether to share their data could result in a future where people are exempt from opportunities until otherwise demonstrated by their data. In an ideal world, each person should have a choice as to whether to share their data. In practice, this is neither realistic nor plausible.

There are ethical implications for John not consenting, or wanting to consent, to the healthcare team using his Apple Watch data—namely, that John's Apple Watch data belongs to John. Should the emergency response team have used John's heart rate only to ensure it was beating or was it ethical to diagnose John's atrial fibrillation subsequently? With informed consent, John would have a choice, which is a fundamental pillar of data ethics.

This is a choice that patients previously did not need to make. Before the datafication of modern life, there were fewer devices to capture such data and far less sophisticated means of predicting future events. The advances of AI and machine learning in healthcare mean that today there are two groups of people: those that seek out health information to help them plan and manage future scenarios, and those that are happy to live in a state of not knowing. As the datafication of everything continues, those that are happy to live in a state of ignorant bliss are finding less opportunity to do so.

Should a Person's Data Consent Ever Be Overturned?

In an ideal world, an individual's decision to share their data should be respected. However, as an absolute concept, this is neither realistic nor plausible. For instance, let's assume John was to decline consent to use his data. On the assumption that the emergency response team's duty is to safeguard the health of its patients, John's data helped to monitor his vital

signs, which contributed to his survival. Arguably the emergency response team would be acting unethically if they were not to use the full variety of data available to them at that given moment. Perhaps even John himself would overrule his consent if it meant increasing his chances of survival.

Precedent demonstrates freedom of choice and consent is ultimately a utopian concept. In October 2016, a man accused of the rape and murder of a 19-year-old medical student in Germany had the health data from his smartphone used against him at trial.[78] The suspect, who was identified by a hair discovered at the crime scene, refused to give police the PIN code to his smartphone. Police enlisted the help of a cyberforensics firm in Munich who broke into the device. Data on the suspect's iPhone recorded his steps and elevation, which police analyzed. Police suggested the suspect's elevation (stair climb) data could correlate to him dragging his victim down a riverbank and climbing back up. As well as locating the suspect's movements, the phone also suggested periods of strenuous activity, which included two peaks the onboard smartphone app put down to climbing stairs. Police investigators mimicked how they believed the suspect disposed of the body and demonstrated the same two peaks of stair climbing detected on the suspect's iPhone.

Public Understanding

The Facebook Cambridge Analytica fiasco went on to highlight just how unaware the public are on the topic of data privacy. The US Senate's interrogation of Facebook CEO Mark Zuckerburg illustrated just how ignorant and unaware the public was on the topic of technology. Senators were puzzled as to how Facebook made money as a free platform, referred to sending e-mails through WhatsApp, and queried as to whether encrypted messages could ever be used to provide targeted advertising.[79] The misunderstandings and ignorance demonstrated by critical stakeholders on the topic of data governance were astounding.

Increased public awareness and understanding are required to educate the public on the use of data and to empower people to decide how, where, and by whom their data can be used.

Who Owns the Data?

People generate thousands of data points daily. Nearly every transaction and behavior creates a data trail left through devices such as smartphones, televisions, smartwatches, mobile apps, health devices, contactless cards, cars, and even fridges. However, who owns the data? The topic of data ownership is an example of a complex first-time problem that has facilitated the development of international policy and governance. Historically, user data is owned by companies rather than the individual.

EHRs enable machine learning-based predictive analytics that will eventually enable providers to provide enhanced levels of care. Although it seems unlikely a patient would not want to not share their health data for improved morbidity and mortality, one would reasonably assume it would become more difficult to withdraw such privileges, much akin to closed-circuit television (CCTV) where consent is not often noted or obvious, and is widely assumed to have utilitarian benefits over the long term. Many platforms prevent users from accessing their full range of services if data sharing is disabled. This is driven by the organizational need to store data in a robust and central repository for safety, governance, and improvement.

Take the example of a patient who uses a connected blood glucose device and mobile app to track and record their blood glucose. Blood glucose data is delivered from the user's blood glucose meter to her mobile application. Although the data would be presented on the user's phone, the data is held in the mobile application provider's database, subject to the terms and conditions of use. Many of today's web sites, mobile apps, connected devices, and health services state that data can and will be used in an anonymized and aggregated format, or in some instances in an identifiable format, by the organization providing the service and also typically with selected partners.

Data collection, usage, and sharing have become a fundamental part of supporting the data-driven quality improvement processes.

Types of shared data include the following:

- Anonymized data

  Anonymized data is data that has identifiable features removed. Identifiable features are features that enable someone to recognize the person that the data has come from. For example, an oncology ward's spreadsheet of patients would be anonymized if the name, date of birth, and patient number were removed.

- Identifiable data

  Identifiable data refers to data that can be used to identify an individual. For instance, an oncology ward's spreadsheet of patients would be identifiable if it reported the name, date of birth, and patient number.

- Aggregate data

  Aggregate data refers to data that has been combined and where a total is reported. Following the oncology example, if the ward's spreadsheet covered ten patients, aggregate data reporting could include the male-to-female split or age brackets of patients. Data is cumulatively reported for the population within the dataset.

- Individualized data

  Individualized data is the opposite of aggregated data. Instead of data being combined, data is reported for each person within the dataset. Individualized data does not have to be identifiable.

Concerns over data privacy have spurred a global response. In May 2018, GDPR (or the General Data Protection Regulation) came into force across Europe. GDPR legislation governs organizations in how they use and share user data.[80] Because of GDPR, organizations have been forced to collect the opt-in consent of their memberships and declare to users how they would use user data and with whom they should share it. The GDPR legislation places data control firmly in the user's hands. Users of data-driven systems are now able to exercise their rights in seeing the data held on them, with whom their data is shared and why, the right to be forgotten, and the right to be deleted. GDPR also defined labels for those collecting and processing data:

- Data controller: The data controller refers to the person or organization that controls, stores, and makes use of the data.
- Data processor: The data processor refers to the person or organization who processes data on behalf of a data controller. Based on this definition, agents such as calculators could be considered data processors.

For those working with data, it is useful to understand the distinction between data processor and controller, and the responsibilities of each party.

GDPR demonstrates how vital data safety has become. GDPR is tightening data access, security, and management in an ambitious attempt to protect the EU's 500 million citizens. It has single-handedly repositioned control with users. The maximum fine for the worst offenders against GDPR regulations is set at either €20 million (£17.6m) or 4% of global revenues for larger organizations.[81] Unfortunately, regulations affect all businesses and organizations, and a flurry of reconsent requests flooded the Internet. There is much confusion as to how GDPR laws are applied, as demonstrated when Mark Zuckerberg was summoned to the European

Parliament in 2018. Just like in the United States, key figures were ignorant and unaware as to how data and connectivity work in the twenty-first century.

The topic of a user's right to be forgotten within machine learning has posited several compelling ethical questions. For instance, if John's data were to be used in a machine learning algorithm to predict the likelihood of severe hypoglycemia, and if John were to request his data to be deleted, it would be close to impossible to decouple John's data from the machine learning model that is learned. Should John's consent be overridden if there is a utilitarian benefit?

Further still, if John's data were to be for some reason valuable in developing an algorithm to diagnose and predict disease, it could be claimed that it is unethical for John to refrain from sharing his data or requesting his data to be removed from such an algorithm. If John were to be the only patient in the world to have a genetic defect, or if his data was for some reason useful to progress medical understanding, it is evident as to how and why John's data would be useful to humankind.

What Can the Data Be Used For?

Data is already used to fuel a variety of decisions. Employers have used psychometric and parametric testing for decades to understand potential candidates. Nowadays, employers also look at supporting sources of data. Employers often scour the social media profiles of prospective employees for concerns of reputational risk. In the same vain, history is plagued with people who have lost their jobs as the result of (deemed inappropriate) social media use.

What an individual's data can be used for raises ethical and moral apprehensions. The concepts of identity and free will are challenged by the notion that an individual's data could be used to make decisions with direct ramifications that may be unbeknownst to that person.

Car insurance is an industry that has evolved to optimize the use of data. Historically, car insurance premiums have been grounded in accident claim data, which is generalized for segments of the population. The advent of the black box enables more precise premiums based on a variety of demographic and behavioral factors—such as the age of the driver, times that the car is used, the speed of driving, and frequency of erratic acceleration among others. This data can also be used to reduce the cost of car insurance premiums should the driver be sensible and enables the increase in car insurance premiums should the driver not follow the stipulations set out by the insurer. Others see this constant real-time data analysis (and subsequent feedback) as "big brother."

Similarly, life and health insurance are beginning to follow in the footsteps of car insurance. Life and health insurers are increasingly providing incentivized products that reward positive behaviors and chastise negative behaviors. Life insurer Vitality, for instance, offers a product whereby it provides an Apple Watch to its members to encourage positive and healthy behaviors. Similarly, British organization Diabetes Digital Media provide evidence-based digital health interventions to insurers and bill payers worldwide to incentivize wellness.[82]

Individuals should be aware that data could also be used for cynical, more sinister purposes. For instance, patient data could be used to decline an individual from a treatment, operation, or opportunity. The public has been incredibly trusting of data use, and robust data governance is required to prevent its future misuse. GDPR is regarded as the first step toward improved understanding and data governance. GDPR is considered by some to be as revolutionary to data science as the advent of the Internet itself.

Privacy: Who Can See My Data?

Conversations of data ownership naturally lead to who can see your data. The key is in ensuring only approved services and organizations have access to your data. For instance, it is unlikely you would want a life insurance company to be given your medical data without your consent, particularly if there was data that could influence your coverage.

Data sharing between applications is commonplace, with APIs enabling accessibility and faster connectivity between independent services. For instance, users can import nutrition data from apps like MyFitnessPal into their diabetes management or fitness apps. These services often replicate user data among a variety of independent architectures, which leads to concerns over managing approved data access. Applications that enable data integration must also provide the facility to decouple patient data. Systems must be able to verify imported data for data governance, auditing, and patient safety.

The debacle surrounding Facebook and Cambridge Analytica went on to demonstrate that even if one trusts an aggregator of data, it is possible that third parties are engaging with that very data without your knowledge. Facebook requested Cambridge Analytica to delete its Facebook user data. Even though Cambridge Analytica agreed to the request and said they had deleted the user data, they were later demonstrated not to have deleted the data.[83] This act of deception leads to questions of accountability, the ramifications of data leaks, and who is accountable in situations of third-party data leaks.

Moreover, even if data is anonymized, it may not guarantee privacy. Netflix recently published 10 million movie rankings from 500,000 members as part of a challenge to improve the Netflix recommendation system. Although data was anonymized to remove personal details of members, researchers at the University of Texas were able to de-anonymize some of the Netflix data through comparing similar rankings and time

stamps with public information in IMDb (Internet Movie Database). There are natural security problems with anonymous data—mainly that it does not mean that you are anonymous.

When it comes to sharing data, it is useful to distinguish between people and patients. The general public's attitude toward data use in the realm of healthcare is far more welcoming than its attitude toward nonmedical data use. A survey of 5,000 people by Diabetes.co.uk reported that 56% of people are hesitant to share their data without good reason. When a group of patients with type 2 diabetes was asked a similar question about whether their data could be collected and used for further research, 83% of patients opted to share their data.[84] Patients appear to understand or are at the very least hopeful that, through sharing of health data, they are progressing medical understanding and treatment. People, however, are far more skeptical of beneficial data use. Confidentiality is a core tenet of data ethics.

How Will Data Affect the Future?

The sharing of patient data and aggregation of big datasets is being used to enhance diagnosis, treatment, and care. As the types and quality of data improve, healthcare's precision will become more precise in the areas discussed next.

Prioritizing Treatments

Big data medical sets can enable predictive analytics to determine optimal treatment pathways for various segments of the population. There is the potential to prioritize only those who are likely to see efficacy in the treatment, which begs the question as to how best to support those that do not.

Determining New Treatments and Management Pathways

The analysis of real-world experience data, clinical studies, randomized clinical trial (RCTs), and pharmacological data is facilitating treatment and management discovery. Digital health interventions have been demonstrated to reverse type 2 diabetes, reduce the number of epilepsy seizures. Data has the potential to develop new pharmacological interventions and disrupt traditional treatment paradigms.

More real-world evidence

The use of real-world evidence (RWE) at scale from patient communities, digital education programs, and health tracking apps is being increasingly demonstrated to improve the self-management landscape. A concern with real-world evidence has typically been lack of academic robustness. However, RWE is being increasingly used in research to determine population usage and benefits. Real-world evidence is a vital part of AI's ethical journey, and there must be trust between patient and provider.

Enhancements in Pharmacology

The precision audiences required for RCTs and academic studies can be identified quicker and easier through digital platforms. This has benefits, such as quicker project recruitment times, greater potential, and multisource comparison. Real-world data is being used to develop better and more effective drugs.

Optimizing Pathways Through Connectivity—Is There a Limit?

Machine learning's primary use in the short term involves data analysis and predictive analytics.

Security

There are significant privacy concerns that come with a truly unified system. What if such a system were to be compromised? Do people, or patients, want a truly unified system, or is the possibility to use it for malintent greater than the potential good? The consequences of vulnerabilities—whether security, operational, or technical—is amplified in unified systems.

For years, large companies have talked about ways to combine and analyze the giant repositories of data they separately gather about people: location data from smartphones, financial data from banks, relationship data from social networking apps, and search data from browsers—to build a complete picture of a person's behavior.

Facebook reportedly offered to match the data that hospitals had about individual patients with social information gleaned from the social networking site, such as how many friends the person has, or whether they seem to engage with others on the site. The company paused the project after privacy concerns were raised by the Cambridge Analytica scandal.

Further still, there are security concerns around IoT devices, which now range from thermometers, cars, and washing machines to blood glucose meters, continuous glucose monitors, and insulin pumps. The requirement of network connectivity to operate as a smart device leaves all devices prone to vulnerability. Distributed denial of service attacks (DDoS) is used on IoT devices to break in and leave malicious code to develop botnets, leak data, or otherwise compromise the device. WeLiveSecurity

reported on 73,000 security cameras with default passwords in 2016.[85] A basic but key to learning from this is that it is always advisable to change credentials that use default passwords. DDoS attacks are nothing new, but the extent of vulnerability is only being uncovered through research and first-time problems, or security breaches. Research students from the University of Florida were able to compromise Google's Nest thermostat in under 15 seconds.[86]

DDoS attacks can be crippling for any organization. Develop mitigation procedures and ensure network infrastructures allow visibility of traffic entering and exiting the network. It is good practice to develop a DDoS defense plan, which should be kept and regularly updated and rehearsed.

Ethics of Artificial Intelligence and Machine Learning

A primary application of machine learning in healthcare involves patient diagnosis and treatment. AI models are deployed to help physicians diagnose patients, especially in cases involving relatively rare diseases or when outcomes are hard to predict.

Imagine a future where doctors know exactly how many times you've eaten fast food in the last month, and that it is connected to your medical record. That data is used to suggest what foods to eat, comparing your health record to hundreds or thousands of other people to identify who is like you. Moreover, imagine this could directly affect life insurance or health insurance, with possible rewards for positive behavior and avoiding bad foods. Imagine a system that was able to predict your likelihood of developing any particular disease in real time.

The ethics of machine learning refers specifically to the questions of morality surrounding the outputs of machine learning models that use data, which come with their ethical concerns.

Machine learning has already been used to develop intelligent systems that have been able to predict mortality risk and length of life from health biomarkers. AI has been used to analyze data from EHR to predict the risk of heart failures with a high degree of certainty.

Moreover, machine learning can be used to determine the most effective medication dosage learning on patient real-world and clinical data, reducing healthcare costs for the patients and providers. AI can not only be used in determining dosage but also in determining the best medication for the patient. As the genetic data becomes available, medications for conditions such as HIV and diabetes will accommodate for variations among races, ethnicities, and individual responses to particular drugs. Within the same data, medication interactions and side effects can be tracked. Where clinical trials and requirement for FDA approval look at a controlled environment, big real-world data provide us with real-time data such as medication interactions and the influence of demographics, medications, genetics, and other factors on outcomes in real time.

As the limitations of technology are tested, there are ethical and legal issues to overcome.

Machine Bias

Machine bias refers to the way machine learning models exhibit bias. Machine bias can be the result of several causes, such as the creator's bias on data used to train the model. Biases are often spotted as first-time problems with subtle and obvious ramifications. All machine learning algorithms rely on a statistical bias to make predictions about unseen data. However, machine bias reflects a prejudice from the developers of the data.

The capabilities of AI regarding speed and capacity of processing far exceed that of humans. Therefore, it cannot always be trusted to be fair and neutral. Google and its parent company, Alphabet, are leaders when it comes to AI, as seen with Google's Photos service, where AI is used to identify people, objects, and scenes. But it can still go wrong, such as when

the search engine showed insensitive results for comparative searches for white and black teenagers. Software used to predict future criminals have been demonstrated to show bias toward black people.[87]

Artificially intelligent systems are created by humans, who are biased and judgmental. If used correctly, and by those who positively want to affect humanity's progress, AI will catalyze positive change.

Data Bias

Bias refers to a deviation from the expected outcome. Biased data can lead to bad decisions. Bias is everywhere, including the data itself. Minimize the impact of biased data by preparing for it. Understand the various types of bias that can creep into your data and effect analysis and decisions. Develop a formal and documented procedure for best practice data governance.

Human Bias

For as long as humans are involved in decisions, a bias will always exist. Microsoft launched an AI chatbot named Tay in 2017. Tay interacted with other Twitter users and learned from its interactions with others. Once the Twittersphere seized hold of Tay, trolls steered the conversation from positive interactions to interactions with trolls and comedians. Within hours, Tay was tweeting sexist, racist, and suggestive posts. Sadly, Tay only lived for 24 hours. This experiment raises the question of whether AI can ever really be safe if it learns from human behavior.[88]

Intelligence Bias

Machine learning models are only as good as the data they are trained on, which often results in some form of bias. Human thinking AI systems have demonstrated bias amplification and caused many data scientists to discuss the ethical use of AI technology. Early thinking systems built on

population data showed significant signs of bias regarding sex, race, social standing, and other issues.

An infamous example of algorithmic bias can be found in criminal justice systems. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm was dissected in a court case about its use in Wisconsin. Disproportionate data on crimes committed by African Americans were fed into a crime prediction model, which then subsequently output bias toward people from the black community. There are many examples and definitions of biased algorithms.[89] Algorithms that assess home insurance risk, for instance, are biased against people who live in particular areas based on claim data. Data normalization is key. If data is not normalized for such sensitivities and systems not properly validated, humanity runs the risk of underrepresenting and skewing machine learning models for minorities and underrepresenting many groups of people.

Removing bias does not mean that the model will not be biased. Even if an absolute unbiased model were to be created, we have no guarantee that the AI won't learn the same bias that we did.

Bias Correction

Bias correction begins with the acknowledgment that bias exists. Researchers began discussions on machine learning ethics in 1985 when James Moor defined implicit and explicit ethical agents.[90] Implicit agents are ethical because of their inherent programming or purpose. Explicit agents are machines given principles or examples to learn from to make ethical decisions in uncertain or unknown circumstances.

Overcoming bias can involve post-processing regarding calibration of our model. Classifiers should be calibrated to have the same performance for all subgroups of sensitive features. Data resampling can help smoothen a skewed sample. But, for many reasons, collecting more data is not very easy and can cause budgetary or time problems.

The data science community must actively work to eliminate bias. Engineering must honestly question preconceptions toward processes, intelligent systems, or how bias may expose itself in data or predictions. This can be a challenging issue to tackle, and many organizations employ external bodies to challenge their practices.

Diversity in the workplace is also preventing bias from creeping into intelligence. If the researchers and developers creating our AI systems are themselves lacking diversity, then the problems that AI systems solve and training data used both become biased based on what these data scientists feed into AI training data. Diversity ensures a spectrum of thinking, ethics, and mind-sets. This promotes machine learning models that are less biased and more diverse.

Although algorithms may be written to best avoid biases, doing so is extraordinarily challenging. For instance, even the motives of the people programming AI systems may not match up with those of physicians and other caregivers, which could invite bias.

Is Bias a Bad Thing?

Bias raises a philosophical question on the premise that that machine learning assumes that biases are generally bad. Imagine a system that is interpreted by its evaluators to be biased, and so the model is retrained with new data. If the model were to output similarly biased results, the evaluation might wish to consider that this is an accurate reflection of the output and hence require a reconsideration of what bias is present. This is the beginning of a societal and philosophical conflict between two species.

Prediction Ethics

As advanced machine learning algorithms and models are developed, more accurate and reliable conclusions will be achieved in a short space of time. Technologies are being used currently to interpret a variety of

images, including those from ultrasound, magnetic resonance imaging (MRI), X-rays, and retina scans. Machine learning algorithms can already effectively identify potential regions of concern on images of the eye and develop possible hypotheses.

It is key to ensure trust in your goals, data, and organization through good governance and transparency. This is fundamental for AI going forward.

Explaining Predictions

As AI algorithms become smarter, they also become more complex. Remaining ignorant about the construction of machine learning systems or allowing them to be constructed as black boxes could lead to ethically problematic outcomes. If an agent were discovered to be predicting incorrectly, it would be a daunting task discovering the behavior that caused an event that is hidden away and virtually undiscoverable.

Interpretability of both data and machine learning models is a critical aspect of intelligent systems. This not only ensures the model integrity but also that it is attempting to solve the correct problem. Users of solutions that embed data science will always prefer experiences where they are understandable and explainable. Data scientists can also use interpretability metrics as the basis for validation and improvement.

Machine learning black boxes could harbor bias, unfairness, and discrimination through programmer and data choices to never be known. Neural networks are an example of a typically unexplainable algorithm. The backpropagation algorithm's computed values cannot be explained. As AI develops in its accuracy in resembling humans, there will be a greater requirement to ensure AI isn't picking up the bad habits of humans.

Protecting Against Mistakes

Intelligence comes from learning, whether you are a human or machine. And intelligent machines, just like humans, learn from mistakes. Data scientists will typically develop machine learning models with a training, testing, and validation phase to ensure systems are detecting the correct patterns within a defined tolerance. The validation phase of machine learning model development is unable to cover all possible permutations of parameters that may be received in the real world. These systems can be fooled in ways that humans wouldn't be. Governance and regular auditing is required to ensure that AI systems perform as intended and that people cannot influence the model to use it for their own ends.

The incorrect classification of predictions can lead to scenarios where the outcome is a false positive or false negative. The impact of both should be considered in the context of your domain. Take for instance an incorrect diagnosis of breast cancer. In a false positive scenario, a patient would be informed they had breast cancer when they did not. Being informed that this classification was incorrect will come to some relief to the patient. A false negative would result in a patient's disease progression and the eventual correct re-diagnosis. The mental and physical trauma as a result of a false negative prediction should be considered. Patients should always be informed of the level of accuracy of results.

Information governance for systems that will be used by patients (patient-facing), whether predictive or otherwise, should provide robust risk mitigation procedures for mistakes in outcomes. Emotional or psychological patient support should be considered for those experiencing trauma as the result of the incorrect diagnosis.

As well as acting as a catalyst for innovation, the speed at which agile digital technologies are rushed to the market is also AI's biggest pitfall. Technology company LG infamously unveiled an AI bot at CES 2018 (Consumer Electronics Show, an annual trade show organized by the Consumer Technology Association) that ignored the presenter's

instructions. The AI bot, which was marketed as providing innovative convenience, failed to respond to any of its master's comments, was either malfunctioning, or chose to ignore the commands.[91] Whether in healthcare or not, AI system performance metrics should be transparent and audited regularly. The cost, particularly in healthcare, is too high not to. AI systems in healthcare have been demonstrated to fail exorbitantly. In the UK, an NHS breast cancer screening system failed to appropriately invite women for screening, with observers claiming up to 270 females may have died as a result.[92] In this particular case, the organization responsible for running the system placed blame with a contractor. The ethical implications and public perception of such mistakes are enormous.

Validity

There is a requirement to ensure the validity of machine learning models over time to ensure the model can be generalized and generalizations are valid. Regular testing and validation of models are essential to maintaining integrity and precision of your machine learning model. A suboptimal predictive analytics model will provide unreliable results and damage integrity.

Preventing Algorithms from Becoming Immoral

Algorithms can, and do, already act immorally. To date, AI algorithms have mainly performed in unethical ways due to design. This has been demonstrated very well outside healthcare by organizations such as Uber and Volkswagen. Uber's Greyball algorithm attempted to predict which passengers were undercover police officers and used it to identify and deny transport.[93] Volkswagen's algorithm allowed vehicles to pass emission tests by reducing nitrogen oxide emissions during the test phase.[94] Both organizations were internationally condemned for their public

deception and lack of transparency. Precedents such as these from the world's largest companies demonstrate that internal and external auditing is required to validate the integrity and ethics of algorithms and indeed their organizations.

There is a genuine concern that AI may learn to act immorally not only from its creators but also from its experience.

The Ecole Polytechnique Fédérale of Lausanne's Laboratory of Intelligent Systems in Switzerland conducted a project that monitored robots designed to collaboratively search for positive resources and ignore dangerous items.[95] Robots were designed as genetic agents equipped with sensors and a light that was used to flag the identification of a positive resource, which was finite in number. Each agent's genome dictated its response to stimulus and was subject to hundreds of generations of mutations. Agent learning was reinforced by the agent receiving positive marks for identifying positive resources and negative marks for its proximity to poisonous items. The top 200 highest performing genomes were mated and mutated at random to produce the next generation of agents. In their first generation, agents switched their lights on when they discovered a positive resource. This enabled other agents to find the positive resource. Due to the limited number of positive resources, not all agents were able to benefit. Overcrowding meant that some agents would also be distanced from the positive resource they initially found. By the 500th generation, the majority of agents had evolved to keep their light switched off when they discovered a positive resource, and a third of agents evolved to act in a way that was the exact opposite of their programming. Some agents had grown to identify lying agents that had an aversion to the light. Agents that were initially designed to cooperate eventually ended up lying to each other due to scarcity.

These concerns are amplified when applied to incentivized clinical decision-support systems that could generate increased profits for their

architects or providers—a system that was recommending clinical tests, treatment, or devices in which they hold a stake, or by altering referral patterns.

In healthcare, this scenario is very concerning. All machine learning models regardless of their use must be governed and verifiable. A machine learning model that was incentivized to make decisions should also adhere to robust standards of transparency, morality, and scrutiny.

Unintended Consequences

The adoption of AI in medicine is becoming more commonplace; and as a result, first-time problems or unintended consequences will govern the media perception and direction of AI ethics. There are very few technologies that have been embedded into healthcare without unintended, or adverse, effects. The question quickly comes to how data scientists, on behalf of humanity, can mitigate these risks.

The ethics of AI and unintended consequences have been significantly directed by The Three Laws, published by science fiction writer Isaac Asimov in 1942.[96] Asimov's Laws, found in the "Handbook of Robotics, 56th Edition, 2058 A.D.," were designed as a safety feature to prevent robots from harming humans:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given by a human being unless it conflicts with the First Law.
3. A robot must protect its existence as long as such protection won't conflict with the First or Second Law.

In Asimov's stories, man-like machines behave in counterintuitive ways. These acts are the unintended consequence of how the agent applies Asimov's Laws to its environment. Only 70 years later, Asimov's science fiction fantasy is eerily turning into reality. In 2012, South Korea published a Robot Ethics Charter to prevent ills and malintent;[97] and the IEEE and British Standards Institute have both published best practice guides for ethical agent engineering.[98] Best practice guides to designing ethically sound agents are typically grounded in Asimov's Laws.

Fundamentally, intelligent agents are made of binary code and do not contain or obey Asimov's Laws. Their human architects are responsible for implementing the laws and reducing the risk of unintended consequences.

AI is unlikely to become evil and turn on humanity in the way depicted by Hollywood. Rather, a lack of context could lead to AI making unintended and disastrous actions. For instance, an intelligent agent that was tasked with eradicating HIV in the global population could eventually come to some conclusion that to achieve its objective; it should kill everyone on the planet. It is easier to frame AI intelligence in a negative setting. Without careful management, an AI agent's utility function could allow for potentially harmful scenarios. There is a limited foundation to suppose that an AI agent would have such an extreme adaptation, yet it is worth consideration. There are undoubtedly many positive unintended consequences to AI that could save humanity from itself.

How Does Humanity Stay in Control of a Complex and Intelligent System?

Over millions of years, humans have combined ingenuity and intelligence to create methods and tools to control other species. As a kind, humans have evolved to dominate, driven primarily by the capacity to learn from the mistakes of themselves or others. Through this, humans have come

to develop tools to master bigger, faster, stronger animals and techniques such as mental or physical training to achieve optimal performance for such tasks.

What will happen when AI is more intelligent than humans? Human-made AIs are already able to surpass human cognition in niche areas. AlphaGo Zero, the reinforcement learning agent developed by Alphabet, acted as its teacher to master the game of Go.[99] AlphaGo was able to learn itself, without the aid of humans or historical datasets over thousands of game interactions. Are humans doomed to be mastered by an evolving AI race? This concept is known as the singularity—the point at which AI intelligence surpasses that of humans. It cannot be expected that an AI whose intelligence surpasses humans can be switched off, either. A reasonably intelligent agent may anticipate this action and potentially defend itself. With AI that is more intelligent than humans, particularly if learning itself (with its hyperintelligence), humankind is no longer able to predict outcomes and must accordingly be prepared for the situation.

Intelligence

Intelligence refers to the outcomes of AI models and how they are used. Intelligence powers the ads you see, the apps you download, the content you see on the Internet, the cab you hire, and the price you pay for items such as loans and mortgages.

The ethics of intelligence posits questions such as whether an autonomous vehicle should protect its driver or others? Should a car prioritize protecting its passengers or instead protect other drivers and the public? What if an incident were to cause loss of life—what should be the approach then? This may become less of a problem once autonomous vehicles are prevalent and fewer humans are involved. Until then, if an autonomous vehicle on autopilot and without human steering does have

an accident, who should be liable? Court cases involving autonomous vehicle accidents have demonstrated that drivers are often using autopilot on impact.

Assessments have expressed that humans do not want a fair, utilitarian approach and want their vehicle to protect the driver. Regulation on topics such as these will be instrumental in the long-term adoption of autonomous vehicles.

A significant application of machine learning is NLP. Virtual assistants such as Alexa have revolutionized methods of agent engagement. As the use of AI becomes embedded with everyday life, humans must not be assumed to vigilant and aware, and organizations should take steps to assure the public. The vocal skills of Google's Duplex bot were demonstrated at its developer conference in 2018 where it was shown booking a hair appointment. The onstage unveiling involved the Duplex software conducting a conversation with a hair salon receptionist. The computer-generated voice used pauses, colloquialisms, and circumlocutions usually present in human speech. The voice was under control of Google's DeepMind WaveNet software that has been trained using lots of conversations, so it knows what humans sound like and can mimic them effectively. Although welcomed by some, many were concerned at the deliberate deception of a human by AI. This has amplified the public's voice that humans demand explicit confirmation when engaging with AI.[100]

Natural language can be separated into two components: style and content. Analysis of communications such as engagements, e-mails, or messages can easily identify sentiment and opinion. Negative terms, for example, could be identified to understand the state of someone's emotional health. The implications of this, taking place publicly or privately, are vast. Analysis of social media profiles and other unstructured sources of data could lead to employers or insurers making predictions about people based on engagement, communication, or sentiment data available for mining.

Health Intelligence

Health intelligence refers specifically to the intelligence developed through healthcare AI. There are a variety of areas that health intelligence is being used in industry, driven by improvements in patient health, costs, and resource allocation:

- Healthcare services—patients are increasingly engaging with predictive analytic services to diagnose disease. For instance, diabetic retinopathy is being detected through AI systems developed through a partnership between Alphabet and Great Marsden hospital. Similarly, British organization Diabetes Digital Media developed a foot ulceration detection algorithm for earlier and quicker referral to podiatry clinics.
- Pharmacology—new drugs are being discovered through the use of learning on real-world data and patient profiles. This is enabling pharma to develop new and more precise medications.
- Life and health insurance—patients with health conditions (or risks) such as type 2 diabetes and prediabetes, who would traditionally receive loaded premiums, can use digital health interventions to manage and improve their condition: and as a result of their sustained engagement, receive reduced premiums to incentivize wellness. The benefits to life and health insurers are vast. Not only are insurers able to engage with more of the population, through connecting insurance with improved wellness, but insurers also can save money through fewer and reduced claims, reduced paramedical and medication costs, improve their risk profile, and optimize risk algorithms and underwriting.

Health intelligence cannot be developed in an environment where the public's ethical AI requirements are ignored, such as with personal data. The ethics of AI should be integral to the development of systems capable of health intelligence, identifying the risks of agents before they are realized. Health intelligence ethics are integral to understanding the intention of data and use of outcomes in medicine.

Who Is Liable?

AI's application in healthcare is an exciting opportunity to improve patient care and cost savings within a short period. Accurate, timely decisions influence the diagnosis, treatment, and outcomes of patients with any number of illnesses. Healthcare professionals can analyze a finite number of images, tests, and samples; and clinical decisions are still prone to human error. With AI, clinical decisions can be made with varying confidence through the analysis of infinite samples in near real time. The medical and technological limitations of AI appear easier to overcome than the potential moral and legal issues. If a breast cancer diagnosis algorithm falsely predicts breast cancer or an algorithm fails to identify signs of diabetic retinopathy on an eye scan, who should be to blame?

There are three liability possibilities in the diagnosis of a patient's disease:

- A human doctor decides a patient diagnosis, with no assistance from an external agent. A human doctor may be very accurate when symptoms are evident and often spots less obvious concerns at first glance. In this case, the liability is always with the doctor.
- An intelligent agent predicts the patient diagnosis, with a 99% accuracy. Seldom will a patient be misdiagnosed, but errors such as death are neither the liability of the agent or human. If an AI was to make such a mistake, how would a patient raise their issue?

- A human doctor is assisted by an intelligent agent. In this case, it may prove harder to identify responsibility, as the prediction is shared. Even if the human's decision was final, it could be argued that the countless experience of the AI agent would be a key factor in influencing a decision.

Liability for an AI agent's behavior can be understood as belonging to the following:

- The organization that developed the AI agent.
- The human team that designed the AI agent is responsible for any unexpected functioning or inaccurate predictions.
- The AI agent is responsible for any unexpected behaviors itself.

AI agent development is often the result of numerous engineers and collaborators, and thus holding the developers to account is not only difficult to manage but may put off potential engineers from joining the industry. Holding the AI development organization to account sounds like the best route to ethical adherence.

How do we ensure AI systems will not overturn humans? What regulations govern our safety? Friedler and Diakopoulos suggest that five core principles are used to define organizational accountability:[101]

- Auditability: External bodies should be able to analyze and probe algorithm behavior.
- Accuracy: Ensure good clean data is used. Identification of evaluation metrics, regular tracking of accuracy, and benchmarking should be used to calculate and audit accuracy.

- Explainability: Agent decisions should be explainable in an accessible manner to all stakeholders.
- Fairness: Agents should be appraised for discrimination.
- Responsibility: A single point of contact should be identified to manage unintended consequences and unexpected outputs. This is similar to a Data Protection Officer, but specifically for AI ethics.

First-Time Problems

Real-world data is already demonstrating that type 2 diabetes management has been misguided for the past 50 years. Who should be liable when real-world experience demonstrates legacy treatments have been misled?

Supervised and unsupervised models developed on patient data is creating new concepts in healthcare that are defining AI ethics. First-time problems are a form of unintended consequence. Data can now be used within AI to detect risks of disease, such as type 2 diabetes, hypertension, and pancreatic cancer. If an application was able to tell you about a terminal disease you could do nothing about, would you want to know? Would you even want to know that there was a choice to know? Services like 23andMe give feedback on disease risk based on the patient's genetic profile. For instance, if someone were to be informed they had a reduced risk of developing certain diseases, there may be some individuals that have a propensity for riskier behaviors. If a patient is told their susceptibility to lung cancer is lower than average, would this lead to a higher risk of them becoming smokers or other risky behaviors? Similarly, knowing the risk of developing particular conditions could lead to mental and emotional consequences. The implications on human psychology and behavior of knowing disease risk will be discovered as time and evidence develops.

Defining Fairness

AI is changing the way we interact and engage with products and services. As accessibility to data grows, how do we ensure AI agents treat people justly? Recommendation systems, for instance, can significantly affect our experience. But how do we know they are fair? Are we being biased by the choices we are given? A machine learning model cannot be fair if there is not a clear definition of fairness. There are many definitions of fairness and collaboration between social scientists and engineers, and AI researchers are required to determine a clear recognition of fairness. Fairness dictates that data misuse should have public, legal, and ethical consequences.

How Do Machines Affect Our Behavior and Interaction

AI's rapid evolution is prompting humankind to evaluate just what it means to be a human.

Humanity

As AI bots have become better at modeling conversations and engagements with humans, so too have their prevalence. Virtual assistants such as Alexa, Siri, and Google Home are in the hands of over 100 million people—and voice-based telephony agents are commonplace. Not only are agents able to mimic conversational language, but emotion and reaction modeling is more sophisticated and believable than ever. In 2015, a bot named Eugene won the Turing Challenge, which asks humans to rate the quality of their communication with an unknown agent and to guess whether it is a human or an AI agent. Eugene was able to disguise itself as a human to more than half of its human raters.[102] The blurring between humankind and robot-kind is imploring exploratory interspecies

relationships. In 2017, Chinese AI engineer Zheng Jiajia married robot Yingying, a robot spouse he created to ease the pressure of marriage.[103] Human-replacement sex robots are now common, with connected AI robots developed with human-like skin and warmth, as well as safety measures to mitigate electrical shocks or worse.

Whether applied to banking, healthcare, transport, or anything else, humanity is at the beginning of a juncture where humans engage with AI agents as frequently as though they too are humans. Where humans have limits as to the amount of attention, kindness, compassion, and energy they can expend, AI bots are an infinite resource for developing and maintaining relationships.

Behavior and Addictions

Our behaviors are already influenced and manipulated by technology. Web landing pages, shared links, and user experiences are optimized through multivariant testing. This is a crude algorithmic approach to capture human attention. In the 2017 US Elections, Donald Trump's campaign famously used and tested more than 50,000 ad variations daily to micro-target voters.[104] Human experimentation has been taking place for centuries, so this is nothing new. However, there is now the ability to influence choice and freedom like never before. Organizations have a moral responsibility to analyze the impact of AI and safeguard particular groups of people, including vulnerable adults and children, from manipulative tools like nudging or social proofing. Codes of conduct for behavioral experimentation within technological experiments is required.

Technology addiction is becoming a growing concern. Robert Lustig, a professor at the University of Southern California, discovered that the human brain responds to technology in a similar way to other addictive substances.[105] Internet addiction has been documented in a variety of countries; and adolescents appear to be more at risk, partly due to the prefrontal cortex being the last part of the brain to develop. According

to the charity Action on Addiction, one in three people is addicted to something.[106] In 2015, a 38-year-old man was found dead after playing video games for five days in an Internet cafe in Taipei.[107]

Real-world and academic evidence demonstrates that technology addiction is a mounting issue. AI has the potential to improve productivity and discovery. However, humankind is responsible for taking measures to prevent AI leading to destructive, digital addictions that make us more insular, overdependent, and lethargic.

Economy and Employment

What happens after the end of jobs? Supermarkets already use automated checkouts, and even McDonalds have already replaced the human-to-human food ordering procedure with a digital interface.[108] The hierarchy of labor is concerned primarily with automation, and this is also true of healthcare. Pharmacies, assembly lines, and check-ins are all commonly automated systems. Although this appears a concern for humankind, it can facilitate more complex roles for humans, moving the species more from physical work to cognitive labor.

Healthcare organizations exploring the use of AI are understanding how to retrain or redeploy their staff, rather than make them redundant. As these AI systems become increasingly common and grow in capacity and accuracy, conversations are moving past manual and cognitive labor to how specialisms such as radiology will be affected. Healthcare teams that use image processing AI, for instance, are increasingly confronted with the dilemma of what to do with teams of specialists. AI can be used to redirect resources to amplify patient care, with more opportunities to engage with patients and their treatment. AI can be understood as an efficient and effective tool that can analyze and detect patterns in patient data at a rate and efficiency inconceivable to humans, with little misinterpretation. However, AI should be seen to augment a human's natural knowledge and intellect rather than be left to make a final decision.

Healthcare is unique in that a digital-only approach is inconceivable. It appears that human intervention will always be required to mitigate risk and confirm predictions. Even with the most sophisticated AI, 53% of 1,000 respondents to a Diabetes.co.uk survey stated they would want human verification of a machine-led diagnosis.[109]

Affecting the future

There are many indications that how people are using technology is changing how children and adults develop. Among many things, research now demonstrates too much screen time is associated with structural and functional changes in the brain, affecting attention, decision-making, and cognitive control.[110] Another study of 390,089 people by the University of Glasgow found an association between a high level of time in front of a digital display and poor health.[111] Screen consumption time is considered to be sedentary behavior. Excessive screen time has been demonstrated to impair the brain, particularly the frontal lobe, which is used by humans in all aspects of life. The next few decades will illustrate the implications of connected technology on human physiology, psychology, evolution, and ultimately survival.

Playing God

Engineers from the University of Adelaide developed an AI agent that was able to predict when you were going to die. The AI agent analyzed 16,000 image features to understand the signs of diseased organs.[112] The neural-network-based AI model was able to predict mortality within 5 years with a 69% accuracy rate. Enhancements in data access, AI, and predictive analytics is enabling humans to predict more precise morbidity and mortality risks, which reignites the debate as to whether AI is playing God or enabling humans (facilitated with a data-driven approach) to play God.

Overhype and Scaremongering

The capabilities of AI have been historically overhyped through marketing and simplified media representations. Media scaremongering, along with more public breaches of data, have also not helped the cause of AI. There is a fog of confusion and misconception, particularly over the possible use (and misuse) of data and outputs, which needs to be resolved for stakeholders to seriously consider implementing AI into healthcare systems.

Stakeholder Buy-In and Alignment

It is pivotal to ensure all stakeholders are aligned in the health intelligence a project can provide and how it will be used. For instance, developers of AI for healthcare applications may have values that are not always aligned with the values of clinicians. There may be the temptation, for example, to guide systems toward clinical actions that would improve quality metrics but not necessarily patient care, or skew data provided for public evaluation when being reviewed by potential regulators.

Policy, Law, and Regulation

The next decade will provide a pivotal juncture in the ethics of AI. Regulation in the form of laws, ethical guidelines, and industry policy will influence the direction of machine learning. So too is machine learning influenced in the absence of such policies and where much of the risk is considered to lie. By failing to set a perimeter of acceptability, we open the risk of wider unintended consequences.

Technical governance is needed, particularly for the application of AI. It is vital to engage key opinion leaders, stakeholders, lawmakers, and disruptors to form and nurture these policies to ensure the potential of machine learning is realized without undue restriction. Applications of

AI in non-healthcare sectors have demonstrated that there are potential ethical problems with algorithmic learning when deployed at scale.

Lawmakers and critical stakeholders in the legal decision-making process must understand the complexities of AI and welcome the challenge of the first-time problems it brings. Furthermore, it cannot be assumed that AI will be solely used for benevolent purposes: AI is already used in warfare, with the United States and China spending heavily.[113] (Interestingly, the UK has decided to focus on AI ethics.)[114] Healthcare professionals using AI must understand how deployed AI algorithms are created, assess the data used for modeling, and understand how the agent can safeguard for mistakes. Regulators must be prepared to handle each concern individually, as all AI is individual, akin to humans. Safe and inclusive AI development will only be achieved through a diverse, inclusive, and multidisciplinary team of engineers, humanitarians, and social scientists.

International engagement between organizations and governments on the critical topics raised by AI is a must. Questions of ethics, responsibility, employment, safety, and evolution are examples of conversations that will require stakeholders from the public sector, private sector, and academic sector to collaborate as AI embeds itself in society and to ensure humanity's democratized prosperity.

Data and Information Governance

Implementation of AI in healthcare requires addressing ethical challenges such as the potential for unethical or cheating algorithms, algorithms trained with incomplete or biased data, a lack of understanding of the limitations or extent of algorithms, and the effect of AI on the fundamental fiduciary relationship between physicians and patients.

A data governance policy is a documented set of guidelines for ensuring the proper management of an organization's data—digital or otherwise. Such guidelines can involve policies for business process

management (BPM) and enterprise risk planning (ERP), as well as security, data quality, and privacy. This differs from information governance, which is a documented set of guidelines for ensuring the proper management of an organization's information. An information governance policy would cover the use of predictive analytics outcomes. An ethical code of conduct or ethical governance policy would cover the organization's direction and intent and steer governance.

Is There Such a Thing as Too Much Policy?

There is a possibility of too much policy, hampering progress and disengaging stakeholders. AI has fueled innovation across all industries including digital health. Start-ups have led the way for AI pervasiveness because of less bureaucracy and red tape compared to organizations that are larger and typically slower to respond to environmental change. Audit your policies to ensure space for listening to the user's voice and incorporating it into cycles of ethical innovation.

Global standards and schemas

There are several bodies that exist that share common codes of ethics and responsibility, but no globally enforced schematics. An AI agent that predicts your risk of death has little validation other than you, or someone else it has predicted on, dying. How should such an agent be verified? To which standards must it uphold? Must the agent, or its creator, provide evidence? As yet, AI has no defined industry standards to which technologies are tested against. As such, it is difficult to be assured of the quality of the AI you are engaging with. And in a fast-paced industry such as AI, the concerns presented are novel and immediate. Global standards are required to set a minimum expectation of AI systems.

Healthcare also suffers from a time lag from evidence-based to clinical acceptance. For example, people with type 2 diabetes have placed their type 2 diabetes into remission for well over a decade, but health systems have been slow to be able to register a patient as placing their type 2 diabetes into remission. Thus, patients have historically either been labeled as having reversed their type 2 diabetes or as living with type 2 diabetes. If patients were labeled as having reversed their diabetes, they often would not receive the tests that confirmed their type 2 diabetes was in remission and were at risk of not receiving checks. Equally, if the patient was labeled as having type 2 diabetes (and did not), this has an impact on insurance and is technically untrue. Collaboration on schemas and parallel strides of adoption would help to reduce international disparities.

Do We Need to Treat AI with Humanity?

The basic concept of reinforcement is parallel to how humans and other animals learn. For instance, when training a dog, compliant and expected performance is often rewarded with a treat. Noncompliance by a reinforcement learning system is met with punishment. This is similar to reinforcement learning by an AI agent, where we build mechanisms of reward and punishment. Positive performance is reinforced with a virtual reward, and negative actions are penalized to augment avoidance.

The majority of AI systems are currently fairly simple and reductionist. As AI develops, we can expect them to become more complex and lifelike. Humans are already beginning to develop relationships with robots, in many cases seen as replacements for another human whether for companionship or carnal desire. It could be considered that a penalty input to an AI system is a harmful or negative input. Are genetic algorithms that delete generations that are no longer of use a form of murder? At what point do we need to consider AI's humane treatment is catalyzed when AI systems can mirror both human behavior and appearance?

If AI systems are considered to be able to perceive, feel, and respond to stimuli, it is not a huge leap to consider their place as a species, legal status, or even nationality. An AI agent made in America was even given Saudi Arabian citizenship.[115] Should AI be treated like an animal of intelligence comparable to humans? Can machines truly feel and suffer? The moral questions of how we treat AI agents, mitigate suffering, and the risk of negative outcomes are fundamental to responsibly progress AI toward its immense potential to better the lives of humanity.

Employing Data Ethics Within Your Organization

Many organizations have a code of conduct that serves as an internal communication on expected behaviors, requirements, and engagement and provides external stakeholders confidence and reassurance. Codes of conduct should be shared with employees alongside appropriate training to ensure the code of conduct is understood and observed. All stakeholders of an organization should practice and promote the code of conduct adherence.

Ethical Code

A code of ethics or ethical code is a document that is used to govern the moral conduct of an organization. A code of ethics demonstrates that an organization is committed to responsible business and technological progress.

The code of ethics narrates the behaviors that are promoted by an organization and those that are considered harmful to the organization's own moral compass, reputation, or clients. It may not cover actions that are illegal, but it will typically state the repercussions of noncompliance and how such violations can be reported. Employees should be made

aware of items that are not obvious to them and then facilitated to avoid inadvertent yet potentially harmful actions. A code of ethics should also include a summary of the motivations for using data and its purpose in the organization's ambition—and reflect the profitability, integrity, and reputation of a business.

A code of ethics should be free of technical and philosophical jargon and directly communicate the expectations required of employees. Keep it simple and right to the point. Set the expectations for new employees at the time of joining and develop an organizational culture of adherence. Ensure all stakeholders are aware of amendments and reviews. Rather than only referring to the code of conduct to recruits or clients, refer to the ethical code frequently to engrain the ethical direction of your organization among internal and external stakeholders. Refer to the code of ethics when examining the ethical risks of incorporating new technologies into the decision-making process.

When developing your code of ethical conduct from scratch, consult employees and stakeholders for their contributions. Questions to consider asking include the following:

- What does AI ethics mean to you?
- How can what we do improve humanity?
- How should our organization act responsibly in its ambition to [insert ambition here]?
- What are the potential benefits of what we would like to achieve?
- What are the potential disadvantages of what we would like to achieve?
- How can we improve our ethical code?
- Are there items in the code of ethics that are confusing or require more explanation?

- Is the code of ethics useful in making decisions?
- Are the organization's ethics in alignment with your own ethical viewpoints?

Once all of the preceding questions are answered by different segments of the organization, a consensus can be reached on an implementation plan. An organizational code of ethics serves as a reference point for disciplinary actions for those who fail to meet the standards. The code of ethics provides a solid foundation for identifying and dealing with ethical challenges.

Ethical Framework Considerations

Organizations have a responsibility to be ethically guided in the collection and use of patient data. Organizations require a data privacy approach that prevents breaches and enforces security. Failing to adhere to regulations can lead to fines, reputational repercussions, and the loss of customers. There are various ways to mitigate the risks of data collection while taking advantage of the opportunities big data has to offer. There are several techniques to safeguard ethical data collection.

Collect the Minimal Amount of Data

The first step toward protecting user data is collecting only the data that is required for the task at hand. More data does not always mean more useful data. Although machine learning typically welcomes more data, data collection must be kept deliberate and concise. For instance, there is little to no use in knowing someone's ethnicity and BMI if they are applying for a credit card; but this same data is important when determining an individual's risk of type 2 diabetes.

Identify and Scrub Sensitive Data

All sensitive data should be identified and secured. Data scientists should be aware of what data is deemed personal and how to use such information. For instance, data collected on consumers that is collected without their consent should be scrubbed of personally identifiable data. It is paramount that only individuals with the appropriate permissions can observe sensitive data. Many organizations also discourage the use of USB sticks and external data drives to ensure data remains safe and on-site.

Compliance with Applicable Laws and Regulations

Trust in your organization's approach to data, and information management can be garnered through adherence to appropriate local and national standards, policies, and laws. There are several data and information governance laws and regulations that set the boundaries of legal and appropriate data usage. It is vital that all regulations are abided to. Bodies such as the FDA and MHRA (Medicines and Healthcare products Regulatory Agency) set boundaries for the legal, ethical, and compliant collection, usage, and management of data. Compliance with GDPR and approval from governance-related accreditation bodies demonstrates your organization's approach to ethical data usage and builds trust with your users.

Many national- and international-level standards will have a fair degree of overlap, requiring appropriate data reporting, documentation, transparency, and risk mitigation procedures. Within Europe, for example, adherence to regulations such as GDPR is a necessity and regulates over 27 countries in the European Union. International standards such as ISO 27001 ensure good data and information management standards. ISO 27001 (formally known as ISO/IEC 27001:2005) is a specification for an information security management system (ISMS). An ISMS is a framework of policies and procedures that includes all legal, physical, and technical controls involved in an organization's information risk management processes.

Many data regulations are self-accredited. This means that as an organization, you must collect the documentation required to demonstrate adherence to standards without the requirement for it to ever be verified by an external body. The concerns of self-accreditation are apparent, particularly when applied in healthcare.

Robust data and information governance is the foundation of data ethics.

Thorough data and information governance is the foundation of data ethics. Governance is concerned with how data and information is protected in its collection, use, analysis, and disposal. Good governance covers a variety of topics including data architecture, infrastructure, risk assessments, audit reports, risk treatment plans, design workflows, information security policies, complaint procedures, medical governance, corporate responsibility, and disaster recovery planning. An organizational ethical code steers the direction of good governance.

A Hippocratic Oath for Data Scientists

Data scientists with access to sensitive data are typically required to sign documentation to ensure they comply with risk mitigation and security policies when they begin employment. It has been suggested that to engrain a culture of safety and positive progression in AI, data scientists acting in digital and nondigital sectors pledge allegiance to a Hippocratic oath to do no harm. This would sit alongside regulations, internal manifestos, standards, and codes of conduct operated by an organization.

Auditing Your Frameworks

Conduct regular auditing of ethical, data, and information governance procedures and rehearse worst-case scenarios to ensure that if the worst were to happen, your organization would be able to minimize as much risk in as little time as possible. Risk mitigation drills should be rehearsed at least twice a year.

Review each section of your code of conduct to ensure it represents the values of your organization. Technology organizations move rapidly, so ensure compliance with the latest standards, guidelines, and policies relevant to your industry. Organizations operating with medical data may come up with new methods of prediction, or new AI abilities. Consider aspects of your code of conduct that are missing, particularly if your organization is scaling, or facing new use cases. For example, a historically business-orientated organization that begins to engage with patients will require a new section of engaging with patients.

There is no reason not to engage every member of staff in developing and reviewing your code of conduct. Digital surveys are a quick and easy way to funnel opinion. Diverse contributions make for diverse, inclusive, and multidisciplinary approaches to AI. Ask for comments on new sections or concepts you feel employees should be aware of. Conversations can highlight areas in which employees may need further training, awareness, or demonstration. Develop a dialogue with employees to engrain the organizational culture within the wider team and foster an ethically aware environment. Having the maximum number of stakeholders on board with your organization's code of conduct will likely increase employee motivation and productivity.

An ethical framework in healthcare organizations, particularly for those using AI, provides internal and external stakeholders the security, transparency, trust, and direction required for maximizing engagement. Adhering to industry standards and involving members of your organization in developing and maintaining the ethical framework will accelerate employee and organization-wide adoption. Moreover, good policy documentation, training, and auditing help maintain the integrity of organizational AI.