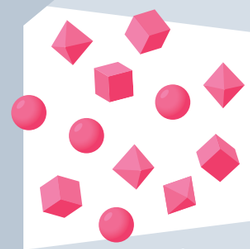
MACHINE LEARNING BIAS

BY PETER STOYKO

Machine learning recognizes patterns and anomalies in data in order to make better decisions. For example, it can make predictions based on patterns in historical data. It can group people or identify individuals based on patterns of similarity and difference. It can even spot suspicious behaviour and sound the alarm. Despite the benefits, machine learning is triggering alarm bells of its own. The logic behind decisions is not obvious, raising accountability concerns. Reliance on data raises privacy concerns. Then there are worries about **bias**. There have been high-profile cases of machine learning discriminating unfairly, making offensive comparisons, and blithely ignoring important distinctions. So what is going on?

:: agent ::

An agent is anything that can interpret data and act on it, such as a robot with physical sensors (camera) and actuators (arms). Or a "smart" household appliance. Or just software, as with tailored recommendations from an online store or an app that identifies the content of photos to organize them.

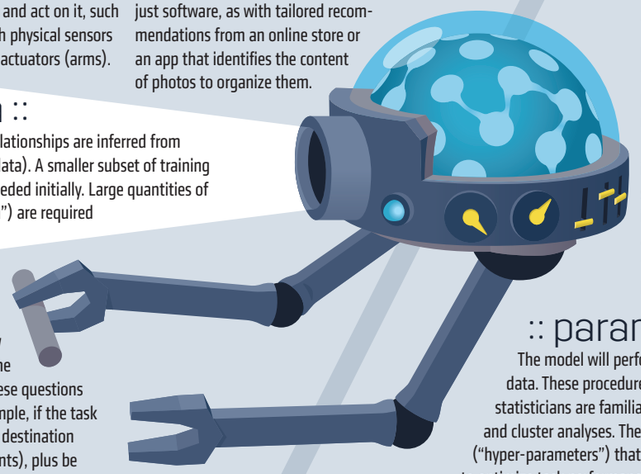


:: data ::

Patterns and relationships are inferred from observations (data). A smaller subset of training data may be needed initially. Large quantities of data ("Big Data") are required ultimately.

:: task ::

What problem is the agent to solve? How is that problem defined? What parts of the environment are relevant? Answers to these questions define the task to be performed. For example, if the task is driving a car, the agent must grasp the destination (goal) and the rules of the road (constraints), plus be able to control the vehicle and recognize objects along the way.



:: model ::

Typically, algorithms tell a computer what to do for each situation it faces. Predetermined pat instructions do not guide machine-learning algorithms. Instead, a statistical model identifies patterns and relationships in data, which then inform decisions. Insofar as assumptions are embedded in the model, the agent is said to lack "autonomy."

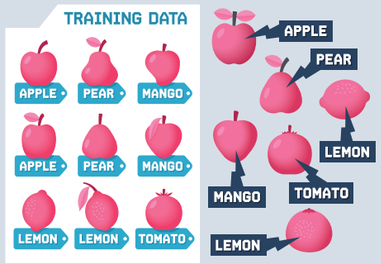
:: parameters ::

The model will perform statistical procedures on data. These procedures are variations of ones that statisticians are familiar with, such as regression- and cluster analyses. These procedures have settings ("hyper-parameters") that have to be tested and tuned to optimize task performance.

:: learning methods ::

The way an agent learns can be grouped into four basic methods.

L1

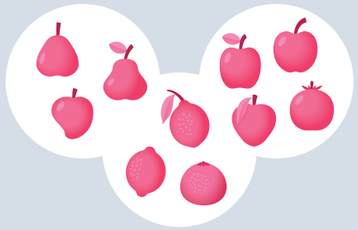


SUPERVISED LEARNING

The agent learns to associate category labels with various examples that are labelled (training data). The patterns found in those examples help the agent infer what category a new (non-labelled) object falls into. That inference can then be used to make decisions and complete a task. Some mistakes will happen, especially if training data was not available for an object.

Labelled validation data can be used to check the model and tune it. Humans often check complex models and label ambiguous data too.

L2



UNSUPERVISED LEARNING

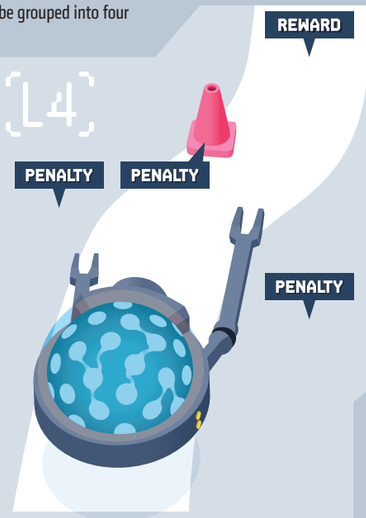
There is no training data. The model measures similarities and differences of objects based on various features. Objects are then clustered into groups. Many photographic applications rely on this type of sorting, for example.

L3

SEMI-SUPERVISED LEARNING

A combination of labelled and unlabelled data may be used. A promising approach is "cooperative" learning that asks for help when it is not confident and adds labels to unlabelled data when it is highly confident.

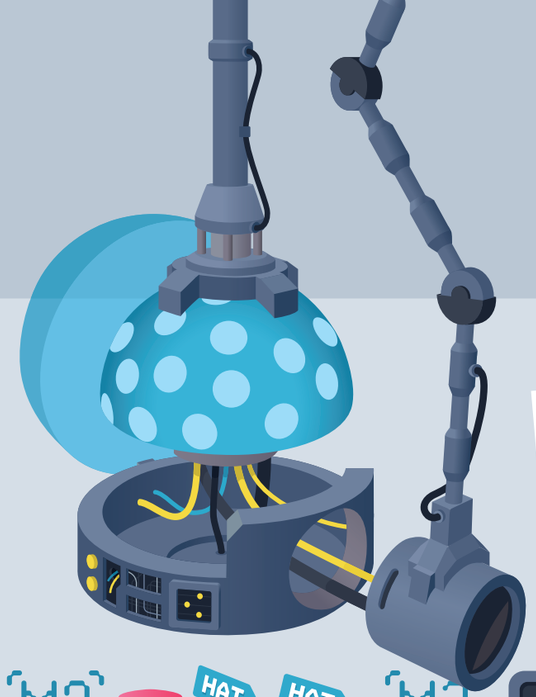
L4



REINFORCEMENT LEARNING

In dynamic environments, the agent can learn with exploratory trial and error. Initial forays are more or less random. A reward function is then used to apply rewards and penalties as the agent works towards a goal. If the task is complex but intervals to reward progress are long, more human guidance ("reward shaping") is involved. That undermines the ability of the agent to apply learning to new environments.

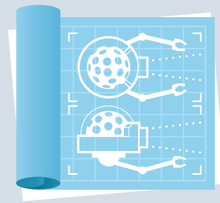
The list of biases builds on Harini Suresh & John V. Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning," *ArXiv*: 1901.10002, 2019.



:: bias ::

We want people in different groups treated fairly and equitably, especially as agents have a greater bearing on our lives. What counts as unfair and inequitable bias is a matter of dispute. Nonetheless, it helps to know where differential treatment can creep in.

:: model bias ::

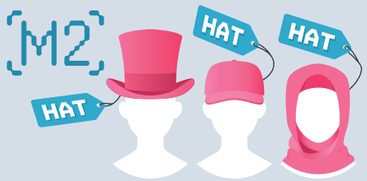


M1

BIASED PROBLEM-FRAMING

What is considered a problem and how it is understood may be biased. Task performance may involve trade-offs that are resolved in biased ways. Unfair, unanticipated downstream consequences may result. Regardless, task performance will not be perfect. It may be a demonstrable improvement over the alternatives. Thus, some expectations have to be managed.

M2



AGGREGATION BIAS

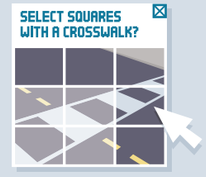
The labels used to describe the data may not be evenly applicable. A one-size-fits-all model will be biased and overall performance of the model will suffer.

M3

EVALUATION BIAS

If benchmarks used to evaluate the model are not representative of the broader reality, any tuning of the model will only be optimized for a limited set of circumstances.

M4

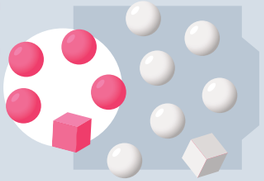


HUMAN-IN-THE-LOOP BIAS

Humans may evaluate results for error or label ambiguous data. Slapdash checks may add bias. Some even use trickery (e.g., are disguised as bot tests).

:: data bias ::

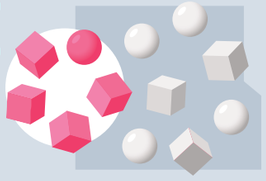
D1



HISTORIAL BIAS

Data is an accurate reflection of reality but that reality reflects long-standing biases in society, which then get entrenched and amplified by machine learning.

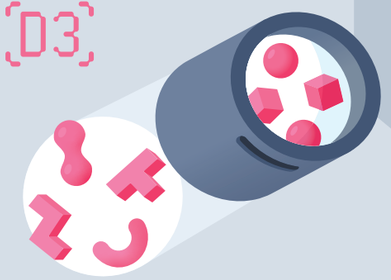
D2



SAMPLING BIAS

The database does not accurately capture the larger reality because the collection method did not offer an equal opportunity for inclusion.

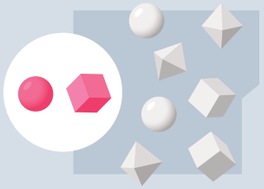
D3



MEASUREMENT BIAS

In practice, "Big Data" is full of quality problems. Some machine learning applications even rely on data sources known to contain dubious data and data of uneven quality. Bias happens when the quality or granularity of the data varies from one group of people to another. Bias can also be baked into the data when the classification schemes used are oversimplifications that favour or disfavour a particular group.

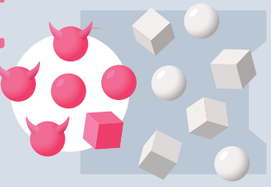
D4



INDUCTIVE BIAS

Training data does not fully reflect the larger reality. That bias is inherent to generalization but is worse if the model tracks too closely or loosely to sparse training data.

D5



ADVERSARIAL BIAS

Hostile actors can deliberately contaminate the database with biased data. Uncurated data fed directly into the agent is a big (and common) source of vulnerability.