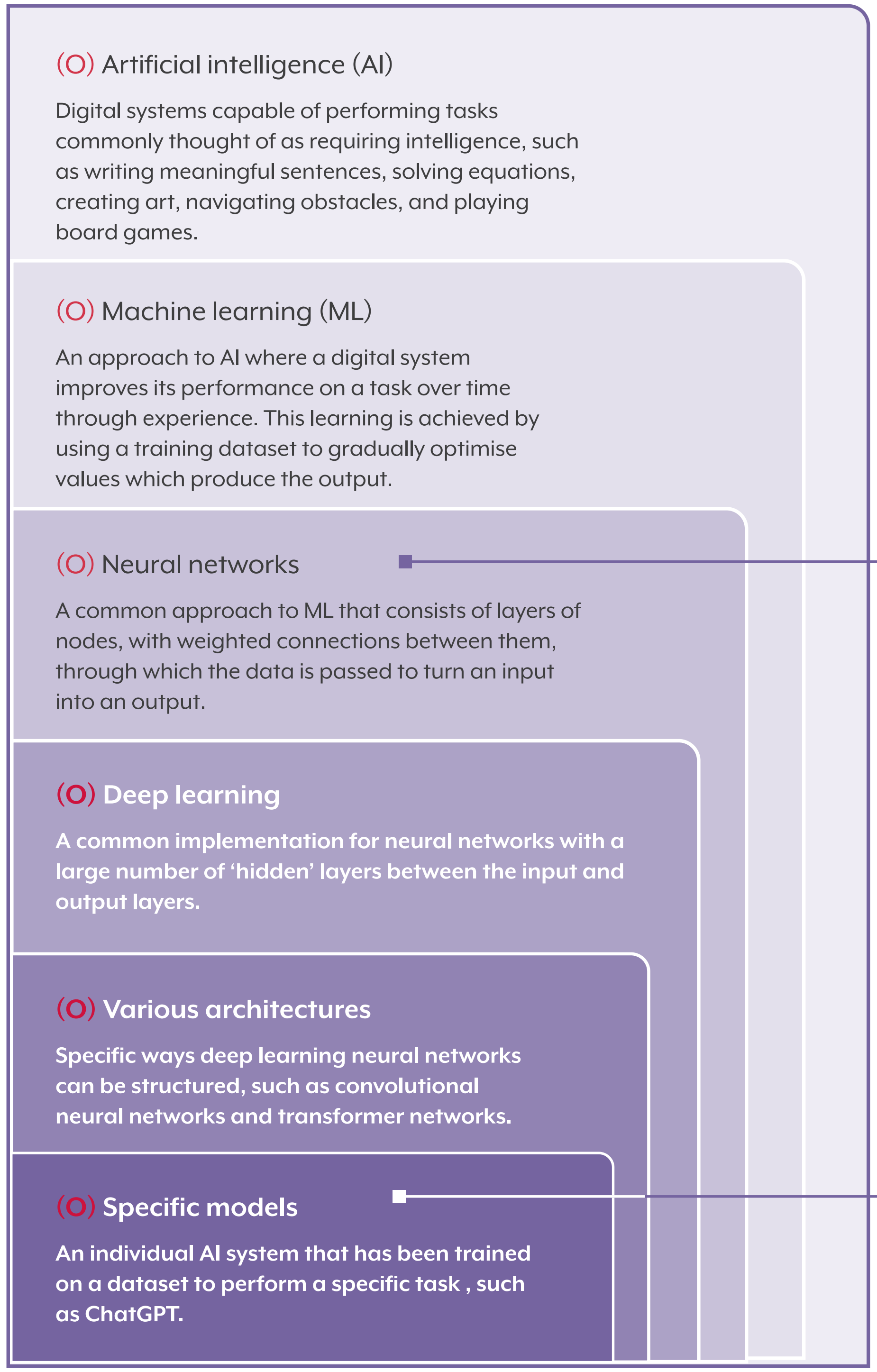
ASD

(O) CONVOLUTED LAYERS: AN ARTIFICIAL INTELLIGENCE (AI) PRIMER

(O) Rapid advances in AI, along with public releases of AI products, have prompted governments, businesses and criminals to accelerate efforts to incorporate this new technology into their operations. These actors are looking to capitalise on the advances in a particular type of AI known as deep learning neural networks - such as ChatGPT. Take up of AI by businesses, governments and malicious actors will enhance existing cyber threats and enable new threat vectors. While it will also be used to enhance cybersecurity, those same systems will themselves become targets of cyber threats.

(O) This placemat provides definitions for some of the most commonly encountered AI terms in cybersecurity and a brief typology of cyber threats that will arise from AI.



(O) Artificial intelligence (AI)

Digital systems capable of performing tasks commonly thought of as requiring intelligence, such as writing meaningful sentences, solving equations, creating art, navigating obstacles, and playing board games.

(O) Machine learning (ML)

An approach to AI where a digital system improves its performance on a task over time through experience. This learning is achieved by using a training dataset to gradually optimise values which produce the output.

(O) Neural networks

A common approach to ML that consists of layers of nodes, with weighted connections between them, through which the data is passed to turn an input into an output.

(O) Neural networks are not the only approach to machine learning. Other approaches include support vector machines, Bayesian networks, and linear regression.

(O) Deep learning

A common implementation for neural networks with a large number of 'hidden' layers between the input and output layers.

(O) Various architectures

Specific ways deep learning neural networks can be structured, such as convolutional neural networks and transformer networks.

(O) Specific models

An individual AI system that has been trained on a dataset to perform a specific task , such as ChatGPT.

(O) ChatGPT is an application based on the GPT3/3.5/4 models, which are deep learning neural networks with a transformer architecture.

(O) Model

(O) An AI system that has been trained to do a particular task. 'Foundational' models are intended to be used with further training to refine their performance. For example ChatGPT is a text generating application based on the GPT 3.5/4 foundation model.

(O) Algorithm

(O) The mathematical process that transforms input data into the output. For ML neutral networks, the algorithm is developed by adjusting the weighting of the connections between nodes during training.

(O) Training

(O) The process where the algorithm is adjusted in response to feedback as the AI uses a data-set to learn how to perform its task. Inference is a form of training to make finer adjusts to models that have already been trained.

(O) Parameter

(O) Variables in the algorithm whose values are adjusted during training and determine how input is transformed into output. 'Hyperparameters' are those set by the human before training.

(O) AI cyber threats

(O) Threats from AI

(O) AI can be used to enhance existing cyber threats and to enable new threat vectors. For example, a malicious actor could use generative AI to enhance spearphishing material.

(O) Threats to AI

(O) As digital systems, AI models themselves can be the target of cyberattack. For example, an AI used for malware classification could be disrupted to enable access by a malicious actor.

(O) Accidental threats

(O) AI can threaten cybersecurity inadvertently. For example, a bug in an AI model could reveal data entered by one user to another.

(O) Threats via AI

(O) AI models and associated datasets and files can be used as a vector for cyberattacks. For example, malicious code could be hidden in an open source AI model that users then download.