Minxian Lai, G2103336A

1. **Among categories of human intelligence, where are we now at Artificial Intelligence? How effective are these Artificial Intelligence?**

For the different categories of human intelligence (ie, Planning, Common Sense, Analogy, Reasoning, etc.) Artificial Intelligence (AI) is most adept in the area of Perception through using Deep Learning as a go-to algorithm with the highest level of success.

Perception is defined as the process by which AI identifies and interprets sensory information that is used to represent the environment.

An illustration of this would be an infant identifying his/her care-givers’ faces.

With Planning as another category of Human Intelligence, there has also been a great deal of attention paid to it - in terms of, for example, decision-making for autonomous vehicles, which is a critical determinant.

Planning is defined as the process by which AI devises action plans that trade off multiple, often conflicting objectives.

An illustration of this would be a child climbing onto a chair in order to open the top compartment of a refrigerator.

Therefore, through Deep Learning - the algorithmic direction by which AI utilises and achieves Computer Vision –these 2 forms of Human Intelligence, Perception and Planning, can be attained, but it would be difficult to surmise the same for the rest of the other categories of Human Intelligence as aforementioned.

Even for Deep Learning itself, we are still in the early stage of adoption (ie, early market) and there exists a gap before it could enter into the mainstream for adoption by the majority, where the idea of “monetisation through the majority” could be realised, maybe in the next several years.

The reasons are three-fold: Firstly, the immaturity of AI Technologies as shown by the dominance of R&D initiatives over production-ready deployments; secondly, the lack of system integrators to help mainstream customers extract value from AI technologies; thirdly, the lack of ROI frameworks to help managers interpret the risks and rewards of their investments in AI technologies.

Therefore, factoring in the above, the level of Artificial Intelligence in mimicking human-level of Intelligence is still quite far-fetched in spite of the hype.

The hype could roughly be attributed to industrial successes centring mostly on supervised learning that is contingent on having massive quantities of human lablelled data defining relevant high-level abstractions, geared towards learning relatively superficial clues and not deep-level enough beyond the testing-training data context.

For unsupervised learning, where the next frontier really lies and where generally humans excel through real-world observations and learning, Artificial Intelligence is still considered to be quite a long way off before being able to be truly considered as on par to Human Intelligence.

That said, when it comes to mental tasks, computers do it far better than humans, such as in the area of chess.

For example, IBM’s Deep Blue chess-playing system defeated world champion Garry Kasparov in 1997.

Devising a successful chess strategy is a tremendous accomplishment, but the challenge is not due to the difficulty of describing the set of chess pieces and allowable moves to the computer

Chess can be completely described by a very brief list of completely formal rules, easily provided ahead of time by the programmer.

Therefore, there are clearly areas in which AI is effective and where it is not so effective.

1. **In Machine Learning, please explain how to balance the Bias-Variance trade off.**

In Machine Learning, the raison d’etre of using models (eg, Decision Trees, Linear/Logistic Regression, KNN, K-Means Clustering, etc.) is to really ensure the best model is chosen that will find the optimal balance in apportioning between in-sample error and model stability.

This is known as the classic trade-off between ‘model bias’ and ‘model variance’ in this field.

Bias is defined as the portion of the generalization error attributable to the simplifying assumptions made in the model. Hence, it naturally declines as model complexity increases. Model bias can be reduced by increasing the model complexity through adding more parameters to the model we can reduce Training error to better fit historical data (i.e. make the backtest look better). But this can lead to ‘over-fitting’ and likely result in much larger Testing errors.

On the flip side, variance is the portion of generalization error generated through the fitting patterns found in any finite/limited sampling of a population. Variance increases as model complexity increases.

In view of this inherent tension, regularisation technique can help to optimise this trade-off by reducing variance from the model, making it more robust to handle multicollinearity (ie, an occurrence whereby independent variables are too highly correlated with one another), having natural feature selection, making more accurate prediction on new data through minimising overfitting on the training data, and through easier interpretation of the output or results.

For regularisation, there are basically 2 approaches:

L1 Regularization aka Lasso Regularization that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).

L2 Regularization aka Ridge Regularization that uses a type of shrinkage estimator called a *ridge estimator*. Shrinkage estimators theoretically produce new estimators that are shrunk closer to the “true” population parameters.

Elastic Nets – a combination of both L1 and L2 regularization, plus adding of a hyperparameter –can be effectively used to manage and mitigate this aforementioned trade-off and keep it to the minimum.

Even in the event of applying these approaches, it must be disclaimed that virtually all cases of financial forecasts can only be modelled to a certain degree. The ultimate goal is really for one to understand what kinds of distributions are relevant to the “real world” that a model can be based upon, and what kinds of ML algorithms perform well on data drawn from the kinds of data-generating distributions.

1. **In Big Data, what are the limitations and potential challenges for the data scientists? How to overcome those limitations and challenges?**

When it comes to Big Data, which by itself is made up of many constituent parts such as Social/Sentiment, Satellite-derived, geo-location, web-data, credit card and so on, the single biggest constituent confronting data scientists today is Social/Sentiment data.

Social/Sentiment analysis-derived data is a very popular type of alternative data. Think Twitter, news and blog. Sentiment data is typically cheaper to acquire as compared to other alternative datasets (e.g. credit card data).

There exist many challenges for data scientists to integrate different social media feeds, because of the following reasons.

Firstly, these data include different formats: each data source may give a different interface and a different format for delivering data.

Secondly, they are not updated in real time: polling websites for updates is complicated by varying latency, for example.

Thirdly, there is always the danger of duplication: many of the data sources report the same activity multiple times; so there is a need to carefully remove duplicated data.

Fourthly, there are always different written styles with Social/Sentiment-derived data: if implementing the natural language processing (NLP) process, the language styles differ across sources. For example, Twitter language differs in style from SEC documentation.

Since the implementation of the EU’s General Data Protection Regulation (GDPR), governments globally have begun to pass more rigorous data protection laws, including the state of California in the US (California Consumer Privacy Act or CCPA) and, more recently, China (People Information Protection Law or PIPL). Outside of government regulation, consumers have increasingly grown wary of how companies collect, store, and use their data. Data scientists have to be mindful of collecting these data from citizens/residents based in these countries.

In terms of overcoming these challenges, data scientists will need to practise the following when it comes to Big Data.

They have to increase their dataset size so as to improve prediction, because massive data sometimes can lead to lower estimation, variance and therefore better predictive performance.

They also need to ensure robust effective non-linear modelling with massive high dimensional data is being used.

To skirt around the privacy laws, data scientists ought to consider the data of people not found in these countries, such as in the developing world that would avail them the critical mass to achieve scalability and actionable insights from their AI/ML approaches.

That said, the data scientists have today a huge libraries of existing data on platforms like Kaggle from which they can comfortably get their data points, plus government sites like data.gov.

1. **How have Cybersecurity companies adapted and responded to the evolution of IT and security architecture, e.g. “The cloud has no wall”?**

Cybersecurity companies have pivoted to a more decentralised/distributed form of network, beyond the traditional network, by recognising that security must perforce be prioritised by breaking down data silos and adopting an identity-centric model (for both users and machines).

As part of the pivot, analytics as well as automation/orchestration has become paramount in this new model, where data from various security controls must inform security operations as a whole and facilitate speedier response to any incidents arising from nefarious cyber activities.

These mean that companies are now increasingly looking at relying and adopting the third-party services offered by companies with expertise in the following areas: Generation Endpoint Security Analytics, Security Orchestration, Automation and Response, Security-as-a-Service, Next-Generation Identity and Access Management (IAM) and Encryption and Application Security.

To define each of these.

Generation Endpoint Security Analytics incorporates real-time analysis of user and system behavior to analyze executables—allowing users to detect fileless “zero day” threats and core advanced technologies prior to and during execution, and take immediate action to block, contain, and roll back those threats.

Security orchestration is the process of integrating a disparate ecosystem of SOC tools and processes to automate tasks for simpler, more effective security operations. Security operations teams typically have dozens of cybersecurity security tools in place to prevent, detect and remediate threats.

Automation Response often coupled with the above refers to technologies that enable organizations to collect inputs monitored by the security operations team. SOAR tools allow an organization to define incident analysis and response procedures in a digital workflow format.

Security as a service (SECaaS or SaaS) in which a service provider integrates their security services into a corporate infrastructure on a subscription basis more cost-effectively than most individuals or corporations can provide on their own when the total cost of ownership is considered.

IAM typically refers to authorization and authentication capabilities like: ... Single sign-on (SSO) so you can give users the ability to sign on once and with a single set of credentials to gain access to multiple services and resources.

Encryption and application security means encrypting data within the application, and not depending on the underlying transport and/or at-rest encryption.

A paradigm shift in the companies’ mindset is also happening in that all large organisations have now assumed that they will be attacked, since they cannot keep out all the attacks, and have contingency incident response plans in place to deal with these attacks and to be resilient.

Increasingly, companies now recognised that information security isn’t capable of predicting or preventing targeted attacks, particularly in the context of an innovating attacker and a mercurial attack surface.

In view of this, they have also invested in enhancing detention and response capabilities, plus endpoint monitoring as opposed to prevention-centric security measures.

1. **How have EU and U.S. regulators done respectively to protect consumer data and privacy, at the same time promote innovation? What they can learn from each other?**

EU, as the world’s foremost implementor of data privacy/protection laws by way of GDPR (General Data Protection Regulation), has really done a lot in this regard and set an excellent precedent for countries all around the world to follow, in terms of its twin objectives: Enhance individuals' control and rights over their personal data and to simplify the regulatory environment for international businesses operating within and within the EU.

Since it was implemented in 25 May 2018, the GDPR has created and enshrined these aforementioned data protection/privacy rights and apply them to government and private-sector treatment of personal data in general, across different institutions and contexts of use.

This inevitably means companies that deal with EU citizens/residents, especially within the digital realm, will have to adhere to the following guidelines as set forth in the GDPR:

Official Privacy Policy: A privacy policy is a statement or a legal document (in privacy law) that discloses some or all the ways a party gathers, uses, discloses, and manages a user’s data.

Official Cookie Policy: A cookie policy is a declaration to your users on what cookies are active on your website, what user data they track, for what purpose, and where in the world this data is sent.

Approved Cookie Consent Manager: Cookie consent manager is the technology that enables your users to decide for themselves which categories of cookies and tracking they wish to consent.

Subject Access Request: Data collection form that links from the footer of a website to allow users the opportunity to make requests for information the company collects and utilizes.

Through the application of all these, it is evident that GDPR is built on a privacy-by-design premise, which would have forced companies to develop new methods and technologies so they can ensure users’ rights and privacies with full transparency on how their data will be processed.

It would have availed a perfect opportunity to re-evaluate how a company’s products and services are created and marketed, thereby, increasing their competitiveness and differentiations.

Whereas the EU has its GDPR, the US’ privacy laws, by contrast, don’t provide for such broad rights for their citizens/residents. US-equivalent legislation, the Privacy Act of 1974, offers limited protections to certain categories of federally held personal data.

Other national legislation is “Sectoral" - applying different restrictions to personal data held in different settings, such as consumer credit or healthcare delivery.

That said, the GDPR does add a lot of compliance costs for businesses tapping on the EU markets. This might be something that is bad about the GDPR.

The US light-touch approach rather than strict enforcement could be something for EU to learn from, especially in channelling efforts to concentrate on investing in areas of their business that could meaningfully reduce the risk of consumer harm yet focus on the UI/UX aspects in the meantime.

Seen in this regard, the EU clearly has a lot to teach the US about in the above regard, and vice versa.

1. **Please explain which application areas that AI/ML/Big data can gain strategic advantages in Legal Tech (Legtech), and why?**

In terms of application areas, Legal Tech can benefit from AI/ML/Big Data through the following areas at large.

Due diligence, where litigators perform due diligence with the help of AI tools to uncover background information; including contract review, legal research and electronic discovery in this section. This is in relation to one of the primary tasks that lawyers perform on behalf of their clients in the confirmation of facts and figures, and thoroughly assessing a legal situation. This due diligence process is required for intelligently advising clients on what their options are, and what actions they should take. Two notable systems are Kira Systems and eBrevia.

Prediction technology, where an AI software generates results that forecast litigation outcome. This is predicated on the use of statistical model that are driven by AI/ML to forecast outcomes with a view on accuracy.

Legal analytics, where lawyers can use data points from past case law, win/loss rates and a judge’s history to be used for trends and patterns. Case documents and docket entries provide supplementary insights during litigation by lawyers. Current AI tools claim that today’s software products are able to extract key data points from these documents to support arguments. One key software is Premonition, which claims to be the world’s largest litigation database. It claims to have invented the concept of predicting a lawyer’s success by analyzing his win rate, case duration and type, and his pairing with a judge at an accuracy of 30.7% average case outcome. The qualifier is that it requires a lot of big data to work, plus its model has been called out for being “exceptionally complicated.” That’s because it needs almost 95 variables supported by almost 4,000 randomized decision trees to predict a judge’s vote.

Document automation, where lawyers use data points from past case law, win/loss rates, and a judge’s history to be used for identifying trends and patterns from previous legal cases. Some law firms are also beginning to adapt such technology by drafting documents through automated software. Many such software companies claim that the final document, which could take days by manual human drafting, is generated in a matter of minutes. One case in point is PerfectNDA, which works by shortening the nondisclosure agreement (NDA) process by offering templates selected by AI according to a user’s scenario. The user answers questions and a pre-filled template is then generated.

Intellectual property, where AI tools guide lawyers in analyzing large IP portfolios and drawing insights from the content. Securing patents, copyrights and trademarks is often best left to a lawyer’s expertise. However, the entire patent application process can be long and arduous. Traditional trademark and patent search, for example, involves looking into hundreds, if not thousands, of results through manual research. This takes so much time, which is ironic considering that patent applications are time-sensitive. Case in point is TrademarkNow, a company taking on some of the manual knowledge work of intellectual property application using AI.

Electronic billing, where lawyers’ billable hours are computed automatically. Electronic Billing platforms provide an alternative to paper-based invoicing with the goal of reducing disputes on line items, more accurate client adjustments, (potentially) more accurate, timely reporting and tracking, and reduced paper costs. Case in point is Brightflag, which offers a centralized legal pricing software that automatically adjusts line-by-line items.