LawTech.Asia

Response to the Personal Data Protection Commission's

Public Consultation on the Proposed Model AI Governance Framework

LawTech.Asia[1]

I. Introduction and structure

1. This paper seeks to:

   (a) Propose a framework/guidelines for implementation of the Proposed Model AI Governance Framework ("**Model Framework")** to the legal technology ("**legal tech**") Industry; and

   (b) Provide additional considerations/suggestions on the Model Framework, specifically in the following areas:

      (i) Internal Governance Structures and Measures;

      (ii) Determining AI Decision-Making Model;

      (iii) Operations Management;

[1] LawTech.Asia is an online publication that aims to drive thought leadership in law and technology matters in Asia. The views set out herein are wholly independent and do not represent the views of any other organisation save for LawTech.Asia.

(iv) Customer Relationship Management.

II. Implementation framework for the legal tech industry (“Proposed Framework”)

A. The legal tech industry

2. This section of the paper seeks to propose guidelines for the implementation of the Model Framework specifically in the legal tech industry (“**Guidelines**”).

3. For the purposes of this paper, we define “legal tech” as “technology that enables a legal services provider to better provide value to anybody involved in understanding or applying the law”.[2]

4. Legal tech can be classified into the following categories and types:

(a) Legal Research and Knowledge Management Software;

(b) Document Management Software;

(c) Document Assembly Software;

[2] Singapore Academy of Law, *Legal Technology Vision 2017*, Accessible at: https://www.sal.org.sg/Portals/0/PDF%20Files/Legal%20Technology%20Vision%20(final%20for%20print).pdf (last accessed 22 June 2019)

(d) Document Review Software (including software for contract review, due diligence and e-discovery);

(e) Practice Management and Billing Software

5. In developing a model through machine-learning, there are at least three considerations which would guide a legal tech developer: [3]

(a) Selecting the right model *family*;

(b) Selecting the right model *form*;

(c) Selecting the *fitted model* which optimizes the parameters after the training process, such that the trained model can be used to make predictive inferences.

6. In the process of selecting the right model *family, form and fit*, the legal tech provider would proceed to process the data. The type of data being processed and how it would be treated would be as follows:

[3] See Wickham, Cook & Hofmann, "Visualising Statistical Models: Removing the Blindfold" *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 8, no. 4, pp. 203–225, 2015; at section 2.1 <http://vita.had.co.nz/papers/model-vis.html> (last accessed 2 June 2019)

Software	**Examples**	**Data**	**Data Treatment**
Legal Research and Knowledge Management Software	LawNet INTELLLEX Oversight	Court decisions	1. Data (in form of legal decisions) is cleaned and/or tagged by humans 2. Data is auto-tagged and auto-categorized by the machine. 3. Humans can override the machine's categorization and impose their own categorization
Document Management Software	NetDocs INTELLLEX	Case Files Firm's internal precedents Client's documents	1. Smart search 2. Cloud Sync and Integration

		Emails	
Document Review (Contract Review, e-Discovery, Due Dilligence etc.)	Luminance Exterro Thought River	Client's documents Contracts	1. Flagging of deviant clauses from standard precedents for review 2. Predictive Coding to identify relevance of documents 3. Recommend actions upon document review
Dispute Resolution	Online Dispute Resolution AI sentencing	Legal decisions	1. Identify trends amongst cases 2. Recommend decisions based on past trends

7. The level of detail of explainability and transparency required of the legal tech solution would increase with the type of impact on the consumers:

(a) **Data Selection and Bias:** For instance, in the context of decision-based legal tech (e.g. Online Dispute Resolution) where decisions made by the software have great impact on the outcome of a case, clients would be more entitled to more detailed explanations as to the method of data selection and data biases which might affect the prediction of cases. In such situations, there would also be concern over the extent to which the human is in the loop – i.e. whether a human decision-maker can override the initial decision recommended by the software.

(b) **Data Protection, Confidentiality and Privacy:** where the type of data dealt with is client data or firm precedent, it would be crucial to explain how data is protected or anonymized, especially in the process of creating training sets. For instance, in the context of contract review, while certain clauses might be flagged for review across the board, where contract review solutions also suggest and propose replacement wordings, law firms would differ in their "house style" phrasing. Each firm's "house style" would be unique and confidential to each firm, and the method of training the contract review legal tech solution, and whether a firm's precedents would be inadvertently shared with another firm or incorporated into the model training which would later be shared with another law firm, would be a paramount consideration to be explained.

In contrast, where software is developed based on more publicly available data (such as legal research tools), confidentiality considerations are less at the forefront.

8. The Proposed Framework presents the following considerations and implications for the Providers and Consumers respectively:

S/N	**Ethical Principle**	**Considerations for Providers**	**Scenarios**
1.	**Explainability and Transparency**	Are clients informed when and how the usage of an algorithm impacts them? Specifically, have clients been informed of the limitations of the product/algorithm's reliability and/or accuracy? If an algorithm's decision affects them, are they provided with information	For example, if an online dispute resolution service has programmed their decisions around a specific data-set, consumers would wish to be warned of any biases inherent in the data-set and how that might affect the decision-making.

		about what information the algorithm uses to make decisions[4]	
2.	**Accuracy and Reliability**	Is there a process for tracking how decisions are made / trends identified? Are there any biases in the data-set and how does it impact the outcome on the consumer? How are errors detected, rectified and minimised? Is there a feedback loop between the customer and the legal tech provider?	For instance, if a contract review legal tech software was trained using a data-set where the majority of contracts used was of a certain type, such that the software mainly identified certain types of potential risks but not others, a client would treasure a feedback loop where users could review and their input would be taken into consideration for the training of the model.

[4] Department of Industry, Innovation and Science of the Australian Government, “Artificial Intelligence, Australia’s Ethics Framework: A Discussion Paper” (5 April 2019) < https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf>, at p 6.

3.	**Fairness**	How was the data selected, processed, used? How were models trained? How does the method of data-selection and use lend itself to certain biases? Is there a process for human review and overriding power in decision-making?	
4.	**Privacy and Confidentiality**	If sensitive client/firm data is used, how are they anonymized for training sets, and/or how are they protected from unauthorized access and usage?	For instance, in the context of a knowledge management system, law firms would be concerned over how client's data is stored and how only certain teams are granted access to the relevant file and data.

III. Additional considerations on the Model Framework

A. Comments to the Preamble and Introduction of the Model Framework

9. **Guiding principles.** The elements of one of the two high-level guiding principles – that organisations using AI in decision-making should ensure that the decision-making process is explainable, transparent and fair - can be further defined.

(1) Making explicit the trade-off between interpretability and completeness in the discussion on explainability

10. The Model Framework states at para 3.17 that an AI solution is said to be explainable if "how it functions and how it arrives at a particular prediction can be explained".[5] The goal of explainability is to "ensure that automated and algorithmic decisions and any associated data driving those decisions can be explained to end-users and other stakeholders in non-technical terms".[6] Explainable AI can be achieved through "explaining how deployed AI models' algorithms function and/or how the decision-making process incorporates model predictions".[7]

11. While these points and clarifications about explainability are helpful, one limitation

[5] Personal Data Protection Commission, "A Proposed Model Artificial Intelligence Governance Framework", January 2019 < https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/A-Proposed-Model-AI-Governance-Framework-January-2019.pdf> ("Model Framework"), at para 3.17.

[6] *Id*, at para 5.3.

[7] *Supra* n 4.

of the Model Framework is that it does not yet sufficiently address the complexity of the explainability conundrum. Hence, in our view, there may be insufficient guidance provided to organisations on the extent of the efforts they should take to provide for explainability.

12. An explanation can be evaluated in two ways: according to (a) its *interpretability*, and (b) its *completeness*.[8] Interpretability refers to how well the explanation describes the way an AI tool works in a way that is understandable to humans. An interpretable system must produce descriptions that are simple enough for a person to understand, while using a vocabulary that is meaningful to the user. On the other hand, completeness refers to the ability of the explanation to describe the operation of a system in an accurate way. An explanation is more complete when it allows the behaviour of the system to be anticipated in more situations.

13. There are several challenges associated with any endeavour to produce an explanation. First, it is a challenge to achieve both interpretability and completeness simultaneously. As noted by the Model Framework at para 3.20,[9] a more technically accurate (i.e. complete) explanation may not always be enlightening (i.e. interpretable), especially to a layperson. The Model Framework also states that implicit explanations of how the AI models' algorithms function may

[8] Leilani H. Gilpin *et al*, "Explaining Explanations: An Overview of Interpretability of Machine Learning" Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory (3 February 2019) <https://arxiv.org/pdf/1806.00069.pdf> at p 2.

[9] Model Framework, at para 3.20.

be more useful than explicit descriptions of the models' logic.[10] An example is to provide an individual with counterfactuals (e.g. "you would have been approved if your average debt was 15% lower"). The need for a trade-off between interpretability and completeness gives rise to ethical dilemmas when building interpretable systems, such as whether it is unethical to manipulate an explanation to better persuade users and how to balance our concerns for transparency and ethics with our desire for interpretability.[11]

14. Second, the Model Framework does not in our view sufficiently address the challenges of achieving interpretability. One big challenge is that whether interpretability is achieved is dependent on the cognition, knowledge and biases of the user. What is interpretable for one user may not be for another with a different level of knowledge about AI.

15. The Model Framework would thus benefit from a more explicit recognition that there is a trade-off between interpretability and completeness. Making this trade-off explicit would also allow for deeper discussions as to how to evaluate explainability and how the evaluation approach would differ depending on the nature of the AI in question. It has been argued, for instance, that proxy methods (i.e. a model that behaves similarly to the original model, but in a way that is easier to explain, such as linear proxy models and decision trees) should be evaluated

[10] *Ibid*.

[11] B. Herman, "The promise and peril of human evaluation for model interpretability", University of Washington eScience Institute (20 November 2017) < https://arxiv.org/pdf/1711.07414.pdf>, at p 3.

based on their faithfulness to the original model while explanation-producing systems can be evaluated according to how well they match user expectations.[12]

(2) Clearer definition of fairness

16. The Model Framework states that a decision is fair towards an individual if it is the same in the actual world and a counterfactual world where the individual belonged to a different demographic group.[13] Fairness involves ensuring that “algorithmic decisions do not create discriminatory or unjust impacts across different demographic lines (e.g. race, sex)”, developing and including monitoring and accounting mechanisms to avoid discrimination when implementing decision-making systems, and consulting a diversity of voices and demographics when developing systems, applications and algorithms.

17. A potential limitation of the Model Framework is that it does not sufficiently address the definition, limitations and challenges of achieving fairness. Consequently, in our view, more guidance could be provided on how organisations could practically ensure fairness in their use of AI tools.

18. As pointed out by Barocas, Hardt and Narayanan, machine learning propagates inequalities in the state of the world through the stages of measurement, learning,

[12] *Supra* n 7 at p 2.
[13] Model Framework, at para 3.22.

action and feedback.[14] One major goal of fair machine learning is to develop an understanding of when these disparities are harmful, unjustified or unacceptable, and to develop interventions to mitigate such disparities.

19. It is difficult to determine whether there is fairness in machine learning based on a single criterion alone. There exist several criteria for determining what is fair. One such criterion is independence, which is defined as requiring the sensitive characteristic in question to be statistically independent of the score. Yet, decisions based on a classifier that satisfies the independence criterion could potentially have undesirable outcomes.

20. Hence, it would be apposite to outline several possible criteria for fairness. In our view, this would provide companies with a more concrete and uniform understanding of what fairness means in the context of this Model Framework.

(3) Explaining the relationship between accountability and transparency

21. While the Model Framework is described as an accountability-based framework, it is presently still insufficiently clear in respect of the stakeholders to whom accountability is owed and how accountability can be practically achieved. It is worth pointing out that private companies do not have the same mandate for public

[14] Solon Barocas, Moritz Hardt, Arvind Narayanan, "Fairness and Machine Learning: Limitations and Opportunities" (unpublished) <https://fairmlbook.org/pdf/fairmlbook.pdf>, at p 31.

accountability that government entities do.[15]

22. Accountability is closely linked to the concept of transparency. This is because it raises the questions of what is to be disclosed and to whom the disclosure should be made. There are various classes of information that may be disclosed about algorithms, including:

(a) **Human element:** The nature of human involvement in developing the algorithm;

(b) **Data:** The quality of the data that drives the algorithm, including its accuracy, completeness, uncertainty and timeliness;

(c) **Model:** The features and variables used in the algorithm and the weights of these;

(d) **Inferences:** The classifications and predictions made by the algorithm;

(e) **Algorithmic presence:** Information about if and when an algorithm is being employed.

23. Depending on the nature of the AI solution, there would be different considerations

[15] Nicholas Diakopoulos, "Accountability in Algorithmic Decision Making" Commun. ACM 59, 2 (Jan. 2016), 8.

as to what and to whom information should be disclosed. In our view, the Model Framework could explore in greater detail what accountability and transparency demands.

B. Comments to the section on Internal Governance Structures and Measures

24. This section sets out comments to the section in the Model Framework on Internal Governance Structures and Measures ("**IGSMs**").

(1) Summary of guidelines under the Framework

25. Broadly, paras 3.3 and 3.4 of the Model Framework set out general guiding principles for businesses to take into consideration when setting up IGSMs for AI deployment.[16]

26. These can be summarised as the following:

(a) IGSMs can be implemented through the adaptation of existing internal structures and/or the institution of new structures. The nature of the structures should be determined by the organisations, which will decide whether an element of de-centralisation is necessary to effectively factor ethical

[16] Model Framework, at paras 3.3—3.4.

considerations into day-to-day operations. The Model Framework states that stakeholder buy-in at the top is critical to the success of the endeavour.

(b) When drawing out IGSMs, organisations have to ensure that clear roles and responsibilities, such as the ones described in the Model Framework, are set out; knowledge transfers are conducted; risk management frameworks and risk control measures described in the Model Framework are implemented.

27. While the Model Framework provides a comprehensive overview of guidelines that could apply to firms in general, there are challenges and suggestions to consider that will provide greater clarity.

(2) The adaptation of existing review mechanisms or implementation of new IGSMs

28. The Model Framework suggests that in creating IGSMs, organisations can adapt existing structures or implement new ones to deal with issues of risks and ethics (among other risks) in AI deployment.

29. Nevertheless, there are some challenges that could arise in the application of these suggestions. Below, where relevant, we have raised examples from the legal and/or legal tech industries to substantiate our points. Suggestions are also proposed below as to how some of these challenges may be mitigated.

30. **Assimilation of IGSMs into existing cultures and systems*.*** The first challenge deals with achieving robust internal governance structures within traditional corporate structures that are predominant in certain industries. For instance, in the legal industry (especially for legal tech consumers which are typically law firms that purchase legal tech solutions including AI-driven ones), the predominant corporate structure used is the equity partnership structure. As these are known for hierarchical and top-down decision-making due to the congregation of positional influence at the top, law firms may find it challenging to achieve stakeholder buy-in or consensus that encourages the adoption and adaptation of robust internal IGSMs. Cultural tensions that may arise in trying to adapt existing review mechanisms and assimilate new IGSMs into existing decision-making models have been commonly noted as a key challenge across industries looking to utilise AI solutions,[17] with ample stakeholder dialogue, participation and involvement being the identified solution.[18]

31. In addition, the Model Framework also notes that organisations should determine the appropriate features in their IGSMs, including a suggestion to use a de-centralised (as opposed to centralised) model to ensure that ethical considerations are factored into operational decision-making.

32. Greater clarity could be achieved if the Model Framework could clarify what some

[17] The European Commission's High Level Expert Group on Artificial Intelligence, "Draft Ethics Guidelines for Trustworthy AI" (18 December 2018) <https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai>.
[18] *Ibid*.

examples of these features could be, or how a de-centralised AI governance system might look through providing context. This is especially pertinent as the Model Framework refers to the "establish[ment] of monitoring and reporting systems" to ensure that the "appropriate level of management is aware of the performance of and other issues relating to the deployed AI". [19] A potential situation could arise where the firm may utilise a decentralised AI governance decision-making structure but is unable to cope with the communication gaps inherent in such a structure, resulting in information asymmetry. Various stakeholders at the operational level who utilise AI solutions for day-to-day decisions may be unaware of sensitive and privileged[20] data regarding the performance of the AI that are only available to decision makers at the top, which would have had a material influence on their decisions at the operational level.

33. In our view, sector or industry-specific illustrations and examples would be helpful in demonstrating the types of features and governance models envisioned by the Model Framework, as certain corporate structures are more prevalent in some industries than others, and some industries handle sensitive data and are subject to stricter regulation. The Model Framework could also provide guiding questions or checklists through which firms in one industry can assess whether they have applied sufficient diligence and consideration to their governance mechanisms (for instance: assessing whether steps to prevent unfair bias in the AI system have

[19] Model Framework, at para 3.4(2)(b).
[20] This point is further elaborated on in the section below.

been taken and listing potential steps to be taken; are the data sets used in training the AI sufficiently diverse and representative; has the algorithm design been stress-tested; were limitations stemming from the composition of used data sets considered; and have sufficient measures been taken to ensure that the AI system is auditable for weaknesses and defects).

(3) Putting a focus on ex post *measures in addition to* ex ante *measures*

34. **Remediation and damage mitigation measures.** Second, the Model Framework has provided considerable guidance on considerations in implementing IGSMs and risk management measures for AI deployment from a proactive, risk-prevention-driven perspective. However, the Model Framework could give further consideration to pointers for remediation and damage mitigation measures in order to provide guidance for firms to determine how to respond in the aftermath of a breach. For example, in the legal tech industry, firms deal with both non-sensitive and sensitive data, ranging from behavioural and consumer preference data (e.g. of consumer preferences relating to key decision makers across legal tech corporate clients) that legal tech solution providers may have, to privileged personal and financial services data that law firms may possess across their practices. Hence, both proactive (risk-prevention-driven) and reactive (solution-driven) perspectives should be adopted.

35. As vast amounts of data will have to be used to train AI systems, care has to be

taken to ensure that proper data governance and protection policies include policies and/or procedures for remediation and damage mitigation. For example, incidents could occur where an AI can re-identify anonymised data and the unintended dissemination of such information would have repercussions that are far-reaching (especially in industries that handle sensitive information). In this respect, the Model Framework could provide guiding breach-response questions for firms to consider in such scenarios (for instance, the type of data that was disclosed, and whether the harm can be mitigated). Breach mitigation measures could include firms providing specific services to corporate or individuals who have been affected, depending on the type of sensitive data disclosed. For instance, in the context of financial data, firms could provide or offer various identity theft protection or credit monitoring services, advise individuals to use multi-factor authentication immediately, and if they have not, for bank account access.[21]

36. To illustrate the significance of these in the real world and draw their application to industry, the Model Framework could provide case studies of past data breaches that have occurred globally,[22] such as the Equifax Data Breach of 2017[23] and the DLA Piper *Petya* cyber-attack.[24] Providing post-mortem analyses of these could

[21] Federal Deposit Insurance Corporation (FDIC), “Breach Response Plan Version 2.6” (26 April 2018) <https://www.fdic.gov/buying/goods/acquisition/data-breach-guide.pdf>.

[22] Department of Industry, Innovation and Science of the Australian Government, “Artificial Intelligence, Australia’s Ethics Framework: A Discussion Paper” (5 April 2019) < https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf>.

[23] Tara Siegel Bernard, Tiffany Hsu, Nicole Perlroth and Ron Lieber, “Equifax Says Cyberattack May Have Affected 143 Million in the U.S.” The New York Times (7 September 2017) <https://www.nytimes.com/2017/09/07/business/equifax-cyberattack.html>.

[24] Barney Thompson, “DLA Piper still struggling with Petya cyber attack”, Financial Times (7 July 2017) <https://www.ft.com/content/1b5f863a-624c-11e7-91a7-502f7ee26895>.

better prepare organisations for designing and executing remediation and damage mitigation measures in the aftermath of unforeseeable breaches.

C. Comments to the section on Determining the AI Decision-Making Model

37. This section examines the Model Framework's recommendation on the necessary considerations when determining an AI decision-making model.

38. First, the Model Framework at para 3.10 identifies three broad decision-making models, namely "human-in-the-loop", "human-over-the-loop", "human-out-of-the-loop".[25] In our opinion, it could be clarified that these decision-making models are better referred to as "levels of human oversight in AI-decision making". The Model Framework should make the above clarification as Model Framework readers might interpret AI-decision making models *per se* to mean the algorithms used in decision making, which is addressed at a later part of the Model Framework.

39. Second, the Model Framework provides the balancing of commercial objectives and risks against a corporate value backdrop as an AI decision-making model selection methodology.[26] In our opinion, while the Model Framework broadly addresses the relevant considerations, to better impress upon readers the importance of the above considerations, it could provide more illustrations how the above considerations interplay in AI deployment.

[25] Model Framework, at para 3.10.

[26] *Id*, at paras 3.5—3.9.

40. In our view, the Model Framework could also explore how these considerations interplay in different sectors and illustrate the consequences when an inappropriate model is implemented. We raise two examples below.

41. The first example relates to the airline industry. In the airline industry, pilots frequently rely on autopilot systems. The commercial objectives in deploying autopilots is to reduce pilot fatigue in long flights. The risks introduced in autopilot deployment are significant, as incorrect decision making may result in plane crashes and the catastrophic loss of life.

42. Many of these autopilot systems can be described as “human-over-the-loop” systems, where the autopilot system makes decisions as to the plane’s orientation and engine power without pilot approval, reacting to data from multiple sensors. However, when necessary (e.g. in the event of sensor failure), pilots can disable or override the autopilot’s decision to prevent disaster.

43. The above decision-making model has worked for aircraft, reducing pilot fatigue while retaining proper safeguards from autopilot failure. However, when the safeguards are incorrectly implemented resulting in the “human-over-the-loop” system becoming more akin to a “human-out-of-the-loop” system, catastrophe can happen. The two recent Boeing 737-MAX crashes, if the ongoing investigations concludes as such, may be examples of the above erroneous implementation. The

pilots in these incidents were unable to override the autopilot systems which, reacting to erroneous sensor input, instructed the plane to point its nose down.

44. The second example relates to the financial industry. In the financial industry, algorithmic trading is used to help banks cut costs and hasten deals by automatically breaking down orders into small pieces and searching for platforms where liquidity is plentiful.

45. However, when trading volumes suddenly collapse or volatility spikes, algorithms are programmed to shut down without human-intervention. Hence, these algorithms are arguably considered “human-out-of-the-loop” systems.

46. Unfortunately, widespread shutdown causes volumes to nosedive, causing dramatic price movement. These movements are known as “flash crashes”. In the first half of 2019, there have been at least two major flash crashes, one in the European Union and another in Japan.

47. Human traders would be able to spot an opportunity from the market turmoil — buying a currency in free fall - which would help to defuse it. Hence “human-over-the-loop” trading algorithms, where controls are implemented to allow humans to restart the algorithms during flash crashes, may be an option trader can explore.

D. Comments to the section on Operations Management

48. This section proposes to examine the Model Framework's recommendations on Operations Management in the context of its applicability and feasibility in the legal industry. In particular, we will look at:

(a) Issues of privilege and confidentiality in using data for model development;

(b) Algorithm audits and review processes in the legal industry; and

(c) Exception handling in the legal industry.

(1) Issues of privilege and confidentiality in using data for model development

49. The Model Framework sets out recommendations for ensuring that the datasets used in an AI solution are not biased, inaccurate or non-representative.[27] While these recommendations are laudable, we are of the view that additional considerations of confidentiality and privilege should be addressed when dealing with datasets.

50. The term "privilege" refers to both legal professional privilege and litigation privilege. Legal professional privilege is a rule of law that provides that all

[27] Model Framework, at paras 3.15—3.16.

communications between a client and his lawyer for the purposes of legal advice are privileged. Along a similar vein, litigation privilege applies to all communications which come into existence for the dominant purpose of being used in aid of pending or contemplated litigation. Information which is privileged cannot be disclosed except with the consent of the client.

51. The Model Framework sets out a number of recommendations which presume that the provenance or the genesis of the data can be easily assessed. For example, the Model Framework recommends for one to "*understand the lineage of data*" and also for one to "*keep a data provenance record*".[28] However, the rules of privilege, as well as the law of confidence generally, may effectively preclude such measures from being taken. As an example:

(a) A technology firm, X, wishes to develop a programme which can help to assess how damages are to be apportioned between the plaintiff and the defendant in medical negligence cases. It approaches a law firm, Y, and asks if it wishes to collaborate on this venture: Firm X would develop the AI solution, and Firm Y would feed its data on past medical negligence cases into the algorithm.

(b) To begin with, Firm Y may not even agree as it has to first seek its clients' consent in order to disclose the information to Firm X. However, even

[28] *Id*, at para 3.16.

assuming that Firm Y manages to seek its clients' consent, the dataset would have to be stripped of sensitive information (e.g. identifiers such as names, ages, and so on). Further, once the AI solution has been trained on Firm Y's data, Firm Y may not agree to Firm X releasing the "trained" AI to the world at large. Firm Y's concern would be, among other things, that releasing this AI solution to the world at large may amount to a waiver of the protection of privilege over its clients' information.

(c) As more information is fed into Firm X's AI solution, the end-user's ability to understand the lineage of the data and the transparency of such data would gradually be shrouded under more layers of privilege and confidentiality. This not only impacts the end-user's ability to keep track of the data provenance record, but also its ability to ensure data quality (see 3.16(a) and (b) of the Model Framework).

52. Moving forward, the Model Framework could set out recommendations on how to balance data transparency on the one hand and confidentiality on the other. This may not be achievable with the broad-brush approach that has been taken in the Model Framework: different industries have different considerations when it comes to confidentiality information, and the advantages that come with data transparency do not always necessarily outweigh the importance of maintaining confidentiality. An alternative solution would be to develop Model Confidentiality Agreements for AI solution-providers to adopt when they obtain information from

third parties. Such a model agreement could provide, for instance, that the AI solution-provider may be given access to the provenance of the data for the purposes of ensuring the reliability of the data. Nevertheless, the terms of the agreement should also ensure that the AI solution-provider does not disclose such data to third parties.

(2) Algorithm audits and review processes

53. The Model Framework proposes that algorithm audits can be carried out as part of developing an intelligent system: see paragraph 3.19. To this end, the Model Framework proposes in Annex A some guidelines for algorithm audits that can be implemented so as to ensure greater accountability and traceability in the AI model.

54. As a preliminary point, we note that "algorithm audits" are not explicitly defined in the Model Framework, although it seems that the "algorithm audits" contemplated in the Model Framework are those "*necessary to discover the* ***actual operations*** *of algorithms comprised in models*". This seems to suggest an in-depth algorithm audit rather than merely a surface-level audit.

55. In addition to what is already set out in the Model Framework on algorithm audits, we suggest that different considerations should apply depending on how the relevant AI is being used, and particularly where AI slots in on the decision-making chain of an organization's processes.

(a) To use a law firm as an example, AI can be used in the process of sifting through a client's documents to find evidence that would support the client's case. This would be relatively "low" in the law firm's decision-making chain as the results from the AI's analysis would only be one facet that the law firm would consider in order to make a decision on how to proceed with the case.

(b) However, AI can also be used "higher up" in the law firm's decision-making chain. For example, given a certain input, an AI could be used to decide whether or not to appeal a judgment, taking into account the commercial risks involved and the costs that would be incurred in doing so.

56. We suggest that algorithm audits should be limited to AI products that are further up on the decision-making chain. This is primarily because it would be a challenge to expect AI companies to open up their algorithm for review in all circumstances, especially since algorithms are proprietary information that form the company's competitive advantage. We note that this has already been considered in the Model Framework: see paragraph 4.2(d) of Annex A of the Model Framework.

57. As for AI products that are "lower down" on the decision-making chain, we suggest that what should be utilized are not "algorithm audits" (as proposed in the Model Framework), but rather "process audits".

(a) Whilst “algorithm audits” suggests an in-depth look into the actual operations of the algorithms being used in an AI model, our proposed “process audits” only look at the *how* the algorithm is used, not the algorithm itself.

(b) For example, a process audit of an AI model would look at: (i) how data is prepared for the AI model; (ii) how the AI model is trained; and (iii) what processes are used in testing the AI model.

(c) Such a “process audit” would not entail looking at the inner workings of the AI algorithm, but rather simply looks at the processes surrounding the use of the AI model. This would protect the proprietary information involved in developing the AI model.

58. All workflows can be described by several high-level components: data, prediction, judgment and action.[29] Some AI tools just sit at the data stage, i.e. those which merely help to process the data, whereas other AI tools help to adjudge the decision. A “process audit” would be more appropriate for AI products that are lower down in the decision-making chain; in contrast, where the stakes are higher at a judgment-making stage, it would be more justifiable to require algorithm audits.

[29] Ajay Agrawal, Joshua Gans and Avi Goldfarb, “The Simple Economics of Machine Intelligence”, Harvard Business Review (17 November 2016) <https://hbr.org/2016/11/the-simple-economics-of-machine-intelligence>.

(3) Exception-handling and the legal industry

59. Case law evolves by having unique and exceptional scenarios come before the Courts. If all the cases brought before the Courts were the same, the common law would not have developed the intricacies that it has now - it simply would not have the impetus to. In other words, there is always some degree of randomness and exceptionality when predicting case outcomes.

60. Exception handling is recognized in the Model Framework as a facet that should be considered when designing an AI Model (see paras 3.22(c) and (d) of the Model Framework). To this end, those designing AI models for the legal industry should bear in mind that AI models would not be entirely appropriate for cases which are wholly different from those that came before it, and that there would always be situations which may fall outside the expected outcomes designed for the AI model.

61. To this end, we agree with the proposition in the Model Framework that in assessing the *repeatability* of the AI model, one must ensure that exception handling is in line with an organisations' policies. Specific to the legal industry, we propose that exception handling can be facilitated in the following manner:

   (a) Where there has been a change in the relevant law or legislation, or even (if possible) changes in the practices used by the industry, to consistently update the AI model with such new developments so that it would be able to deal with them;

(b) Ensure that AI models can recognize when a given set of facts contains new and different variables which it had not previously considered, and that the AI model is able to flag these variables out; and

(c) Make it clear to clients that where an AI model is being used, that it had been trained on previous data which may not take into account new variables present in the clients' case.

E. Comments to the section on Customer Relationship Management

62. This section provides feedback on the Model Framework's recommendations on Customer Relationship Management. In particular:

(a) We recommend firmer and clearer regulation in the communication of AI to consumers.

(b) We recommend providing more detailed explanation on how the proposed factors for implementing and managing communication strategies actually go towards building consumer trust in AI.

(c) In respect of the specific recommended strategies, we recommend including a factor on managing consumer trust, confidence and relationships in the event of a crisis involving AI systems.

(d) More broadly, we recommend explaining in greater detail what it means to have consumer trust and confidence in AI systems.

(1) The importance of maintaining consumer trust

63. At present, the majority of consumers trust AI systems that have been deployed. In a study conducted by Salesforce, 67 percent of customers say that they recognise the good that can come from AI, while 61 percent believe the technology presents positive opportunities for society.[30] In fact, customers have embraced a variety of AI-powered technologies, such as chatbots, credit card fraud detection, email spam filters, as well as voice-activated personal assistants like Apple's Siri or Amazon's Alexa.[31] We also recognise that psychologically, humans are likely to trust AI more as it becomes more prevalently used in society (as long as this is largely without incident).[32]

[30] Vala Afshar, "New Research Uncovers Big Shifts in Consumer Expectations and Trust", Salesforce Blog (5 June 2018) <https://www.salesforce.com/blog/2018/06/digital-customers-research.html>.

[31] Vala Afshar, "In the age of AI, trust is the most important core value", ZDNet (5 September 2018) <https://www.zdnet.com/article/in-an-ai-powered-economy-trust-must-be-your-companys-highest-core-value/>.

[32] R Parasuraman and D Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration" (2010) 52 *Human Factors* 381 DOI: 10.1177/0018720810376055. See for example, the NationalTransportation Safety Board's finding that the car driver's inattention due to overreliance on vehicle automation was a contributing factor to the fatal Tesla crash on 7 May 2016 near Williston, Florida, https://www.ntsb.gov/news/press-releases/Pages/PR20170912.aspx.

64. That humans will trust AI more as its use becomes more prevalent in society is a situation that cannot be expected as a matter of course. Although the use of nuclear plants in Japan had been widely accepted as a fact of life, a single natural disaster in 2011 – resulting in the Fukushima nuclear plant meltdown – caused public trust in nuclear power to fall drastically – leading to a huge setback for the nuclear energy industry around the world. In addition, that consumers broadly trust AI today cannot be assumed to be a fact in all industries. For instance, a recent survey of over 1,000 car buyers in Germany showed that only 5% would prefer a fully autonomous vehicle. People also continue to be skeptical of AI-enabled medical diagnostics systems.

65. Hence, we recognise the importance of proper customer relationship management in the Model Framework, and we support its inclusion. Nevertheless, we propose four improvements to the section below, which we can hope can be taken into consideration in making the Model Framework more relevant and robust for industry.

(2) Firm and clear regulation for the communication of AI to consumers

66. **First, we recommend firm and clear regulation for the communication of AI to consumers. In particular, we are of the view that the self-regulation model of AI (in respect of customer relationship management) is an untenable one**

in the long run, and that clear rules will be needed to avoid a race to the bottom in this aspect. There are companies that will resort to marketing tactics in order to communicate to their clients in such a way that it makes customers believe that AI (a) will be used responsibly and (b) will result in positive effects on the customer. Where the technology ends up failing for some reason, the loss of trust by the consumer will not just be in the company, but in AI systems in general. This would run counter to the Model Framework's aim of building broad-based consumer confidence in AI.

67. In order to ensure consistency in the level of responsibility adopted by companies towards consumers in communicating about AI, we are of the view that a self-regulation model (which the Model Framework is currently based on) will not be sustainable in the long run. Ultimately, certain "harder" regulations that prescribe rules of fair play in communicating about AI to consumers will be needed to ensure that consumers are protected, and that the thin line between responsible communication and marketing is not breached.

(3) Greater explanation on how the proposed factors for implementing and managing communication strategies build consumer trust

68. **Second, we recommend providing more detailed explanation on how the proposed factors for implementing and managing communication strategies actually go towards building consumer trust in AI.** While the Model Framework

has been very helpful in recommending certain communication tools and strategies that companies can take to better manage customer relationships when deploying AI, we think that first-order question of "what builds trust" should first be set out.

69. To elaborate, the Model Framework should provide shed some light on how the various factors shared (such as general disclosure and increased transparency) actually go towards building trust (based on studies of what builds trust), and to highlight which are some of the more important or key measures in developing trust. In our view, this would provide companies with a better understanding of *why* these particular factors have been proposed to companies, and which of these factors are recommended to be prioritised (given that each company will face limitations in the extent to which they are able to implement these recommended factors).

70. To that end, what builds trust? In human interaction, trust is the "willingness to be vulnerable" to the actions of another person.[33] Trust is also key to reducing "perceived risk", a combination of the uncertainty and seriousness of a potential outcome involved.[34] In the context of AI, perceived risk stems from giving up control to the AI system. To build trust in AI, three factors are crucial to gaining trust:[35]

[33] Roger C. Mayer, James H. Davis, F. David Schoorman, "An Integrative Model of Organizational Trust", The Academy of Management Review, Vol. 20, No. 3 (Jul., 1995) at p 729.
[34] Ellen Enkel, "To Get Consumers to Trust AI, Show Them It's Benefits", Harvard Business Review, (17 April 2017) <https://hbr.org/2017/04/to-get-consumers-to-trust-ai-show-them-its-benefits>.
[35] *Ibid*.

(a) **Performance:** The AI system performs as expected. One important element of this is the operational safety of the AI system. It has been noted that since AI technologies usually result in the delegation of control and responsibility from a human to an AI system, the AI system will not be trusted if its operation is flawed.

An additional element of performance is the AI system's usability. This is in turn influenced by the intuitiveness of the AI system, and its perceived ease of use. Usability testing with a targeted consumer group can be an important first step towards creating this ease of use. This is also linked to the concept of trialability, which reflects the idea that people who are able to first visualise the concrete benefits of a new technology via a trial run would reduce their perceived risk and thus their resistance towards the technology.

(b) **Process:** The user has an understanding of the underlying logic of the technology. A related element is data security, where the user trusts that data being used to train / operate the system is used and managed in a secure manner.

(c) **Purpose:** The user has faith in the design's intentions. In this regard, it is important to take note of the concept of cognitive compatibility, or what people feel or think about an innovation as it pertains to their values. According to research from the Harvard Business Review, users tend to trust automation if the algorithms are understandable and guide them towards achieve their goals.

These affect the perceived predictability of the AI system, which is in turn a foundation of trust.

71. As for strategies to foster trust, the same study by the Harvard Business Review shared that ensuring stakeholder alignment, transparency about the development process and gradual introduction of the technology are helpful approaches that can be taken to help build trust in AI in consumers. In addition, ensuring proactive communication and openness in the early stages of introducing the AI system to consumers can influence the company's perceived credibility and trustworthiness, which in turn positively influences attitude formation in the tools.

72. In this regard, we are of the view that many of the communication strategies already set out in the Model Framework are aligned the "three Ps" for developing trust in new technologies. For instance, the recommendation to increase transparency through appropriate disclosure of how an AI system's decision could affect consumers would support efforts to ensure that the AI system performs as expected. In addition, the recommendation to develop a policy to provide explanations to individuals on how the AI system works in a decision-making process supports efforts to ensure that users have an understanding of the underlying logic of the AI system.

(4) Including recommendation(s) on customer relationship management in a crisis

73. **Third, in respect of the specific recommended strategies, we recommend including a factor on managing consumer trust, confidence and relationships in the event of a crisis involving AI systems.** In our view, a crisis of confidence in AI can undo much of the goodwill currently built up for consumers in AI. In such crises, other priorities (such as the safety of personal health, property, data or reputation) can supersede the interest in using AI, regardless of its purported benefits. It is also unclear how the recommended communication tools would hold up in the event of a crisis – for instance, having a feedback channel, as helpful as it is in normal operation, may not be sufficient to deal with a drastic loss of confidence in AI due to a crisis.

(5) Greater explanation on what it means to have consumer trust and confidence in AI systems

74. **Fourth, at a broader level, we recommend explaining in greater detail what it means to have consumer trust and confidence in AI systems.** In other words, we recommended providing greater detail on what the PDPC envisions for consumers to be confident and have trust in using AI.

75. In particular, we are of the view that the Model Framework could set out in greater detail what is the envisioned end-goal where consumers trust and are confident in

deployed AI tools. To that end, there are different degrees of trust and confidence a consumer could have in any product (including an AI tool). For instance, while one customer may fully trust the facial recognition technologies within a smartphone, using the tool to unlock the phone, make payments and access third-party applications, another customer may choose to use it only to unlock the phone (due to a lack of confidence that her personal data (i.e. her facial information) could be conveyed to third parties).

76. In our view, this is useful as it allows organisations to understand the end vision that they should be ideally targeting when carrying out their customer relationship management strategies, and to decide the extent to which they wish to pursue their strategy.

LawTech.Asia

24 June 2019