Building Trust and Explainable AI

Sim Xinming[1] and Joey Pang[2]

[1]AI Singapore, Singapore
[2]DBS Bank, Singapore

I. Introduction

The Oxford Online Dictionary Trust defines trust as being the belief in the reliability, truth or ability of something or someone. Stakeholder trust refers to confidence in either the organisation or the AI system, both of which have become increasingly intertwined in the digital environment.

Regardless of the type – whether it is a project, product, or service – the importance of trust in any stakeholder relationship is vital. That's because distrust in AI systems can lead to animosity and suspicion towards the organisation that promotes the use of such AI systems.

Beyond establishing trust with external stakeholders, maintaining trust within the organisation is equally essential to imbue employees with a sense of mission and meaning in the work they do. While there are many ways to earn trust, all such measures rely on a common principle: open and honest communications that focus on the stakeholder's best interests.

II. Building Trust

Often, stakeholders do not fully understand the logic behind AI-assisted decisions. This gap in understanding could is often filled with the unfounded fear that's causes mistrust in the AI systems. There is no universal approach to determine how and what information is required to earn trust. A few due diligence factors are always applicable, such as striking a balance between the information

needs for each stakeholder group, the level of explainability that can be supported by the AI model, and the level of detail that can, or should be divulged.

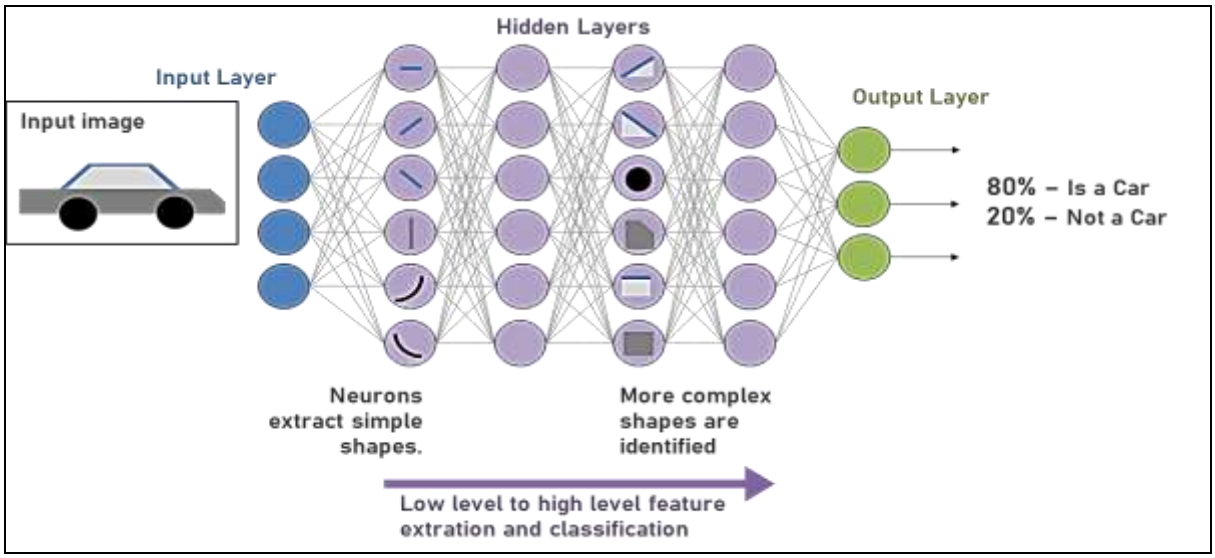
The "why" and "how" would relevant questions to identify information pertinent to different groups of stakeholders. How effectively the organisation can build trust with its internal and external stakeholder groups depends on a multitude of factors, including cultural, political, environmental, economic, and others. For details on these issues, refer to Chapter 2 of the BoK.

Bridge Knowledge Gaps

People fear what they don't understand, and often misinterpret what they partly do. Some people plug their knowledge gaps from potentially misleading or unrealistic sources, such as movie representations of AI. Organisations, therefore, should note that different stakeholders will have varying levels of understanding of AI. Before offering any explanations on the decision-making attributes of AI, organisations may first need to bridge fundamental knowledge gaps in their intended audiences.

Where possible, offer a more straightforward explanation. Strive to explain key AI concepts in a concise and easily understood manner, and where appropriate, provide a non-textual description, such as infographics.

- **Example**: An AI technique used for analysing images is convolutional neural nets. Explaining the mathematics behind the DL (Deep Learning) neural network model (such as weights, bias, activation functions, and others) would require some level of technical competency. However, when communicating with the general public, organisations could use infographics to explain how AI functions. One such infographic:



Convey AI Functionality

Once you have addressed the fundamental knowledge gaps, the next step would be to explain how AI is embedded in specific products or services. You can do that by explaining how the AI decision-making process works, with emphasis on how data (especially personal data) gets used. Plan to provide this explanation at the right juncture, where such description will be the most useful, relevant, or impactful for the user. That could be when the AI algorithm makes a decision, or when data are being collected or processed by the AI system.

- **Example**: An AI engine interprets all "likes" on particular pages that users visit or specific advertisements that users click on, as "user preference" data. That results in a recommendation engine pushing out similar promotions to the user's attention. That has irked many users. As a remedy, some social media platforms have introduced a "Why do I see this ad?" link to help users understand the factors influencing the decisions made by the recommendation engine, and how users can adjust them based on their preferences.

Explain AI Outcomes

Stakeholders need to know why the AI system makes specific recommendations before they can begin to trust it. Explaining AI outcomes is a critical and delicate aspect of communications. While internal stakeholders could be less critical, external stakeholders can

be unforgiving and put an organisation into disrepute in the event of missteps or, misperception, misrepresentation or misunderstanding.

- **Example**: A social media organisation published the results of two AI chatbots negotiating with each other. Despite being an otherwise good scientific experiment, some media outlets dramatised parts of the research with misleading content, suggesting that the company stop further research, alleging that the AI chatbots "invented their language" and that "robot intelligence is dangerous".

Organisations should broadly aim to provide reasonable explanations on how the AI algorithms arrive at decisions. Key stakeholders may want to know why the AI system is behaving in a particular way, how decisions may vary depending on the data, their accuracy levels, and how decisions may differ from those made by a human decision-maker.

a. Explainability of AI Decisions

All AI-based systems operate on the principle of arriving at decisions by using data and rules (or algorithms) based on parameters set by the AI software developers. There are many use-cases of AI models deployed to boost efficiency; the rules of engagement may be relatively straightforward.

But when a "black box" AI model (one where the inputs, the operations or the algorithms are not understood) gets used, it may be tough to explain how these systems work. A "black box", in a general sense, is an inaccessible system. Even if one can examine the features and the weightage assigned to different data points, there is no way to know how the AI system arrived at a specific decision.

The flip side: AI developers cannot get away with treating all AI models as a black box. With society's focus on AI, the demand for greater explainability in AI decisions and outcomes has grown louder. In Britain, the ICO (Information Commissioner's Office) wants regulations to force businesses to provide explanations about results derived from their use of AI, or face penalties.

How can organisations provide meaningful explanations on their AI systems built on ML (machine learning) models?

- Utilise explainable AI techniques and approaches to explain the input parameters that influence the model's decision.
    - **Example**: Accuracy and interpretability are two dominant features of successful predictive models. Typically, a choice must be made in favour of complex black box models such as recurrent neural networks (RNN) for accuracy versus less accurate but more interpretable traditional models such as logistic regression1. Explainable AI techniques could give organisations higher levels of accuracy and interpretability. It could mean identifying and informing patients which hospital visits or clinical variables influenced the AI towards that decision. (Source: https://arxiv.org/pdf/1608.05745.pdf).
- Use proxy models to explain the outcomes of an elaborate black box AI model. Note that the proxy is an abstraction of the real AI model. Organisations should use proxies where the simplicity of using such them does not compromise the accuracy of how such models function. In other words, work on a virtual copy of a real use-case and ensure that it maintains the accuracy and integrity of the data, the processes, and the outcomes.
- When detailed and precise tracing of algorithmic operations is impossible, organisations should document every aspect of the AI system. That includes datasets used to train the algorithm, assumptions made, the scope of coverage, context, constraints, conditions for the application and its intended uses, and a general tracing of the decision steps.
- Explainability depends on the domain and context in which the AI-enabled product or service gets deployed. There can be significant repercussions in sensitive sectors such

as healthcare, finance, legal, and law enforcement. The context is as important as the outcome.

- **Example**: AI-enabled navigation systems in ambulances and other emergency services would face more scrutiny than similar systems deployed in private-hire vehicles.

b. Explainability of AI Decisions

All AI models are abstractions of the datasets on which they get trained. Where datasets are biased or discriminatory towards specific populations, this will skew the outcomes. Organisations should ensure that the AI system minimises such occurrences. There have been several cases of AI discriminating against individuals based on race, gender or socio-economic class. One way of uncovering discrimination could be by rewarding developers for discovering bias, for example.

One key reason commonly cited in favour of AI is the consistency of its decisions. That's a double-edged sword in that decisions can be consistently right or flawed. When automated decisions get made on a large number of transactions, this consistency can have a significant impact, especially if the outcomes are questionable.

- **Example**: Due to the large volume of applications, one of Singapore's local banks developed an AI chatbot to speed up the recruitment process for the initial screening of entry-level wealth planners. The bank said that the chatbot eliminates certain human biases, such as favouring candidates from a particular educational background, thus ensuring a fair recruitment process.

c. Accuracy of AI decisions

Let's discuss AI decision-making from two aspects – data quality and performance.

- **Data Quality**

The volume of data collected has increased enormously and will continue to grow exponentially, with 5G and IoT (Internet of Things). However, just having more data does not necessarily lead to a more accurate outcome or decision. In cases where the AI model is robust in the quality of its datasets, having more data may not make a significant difference to the outcome. But in cases where the AI model is complicated, having more data could lead to better abstraction or streamlining of the model.

You need to differentiate data quality from data noise. Data quality covers a broad range of attributes, including machine-readability (structured data for machine-reading) and accuracy (the data reflecting the ground reality). Organisations should secure adequate quality data by ensuring that data collection is conducted objectively, consistently and correctly. Data collected must be cleaned for duplicates, missing fields and other inconsistencies. An ML model is only as good as the data that it is trained on; if the input data cannot be trusted, the outcome will be flawed.

- **Example**: An AI developer designs a screening tool to detect child abuse to determine whether the family and the child should be separated. However, the dataset used to train the AI model contained latent discrimination; families with higher income status could conceal abuse through the private healthcare system. That resulted in a particular ethnic group being thrice as likely to get flagged out than others.

- **Outcome Performance**

How can we be confident of the validity or meaningfulness of the outcomes predicted by the AI system? In domains where the impact of the decision is substantial, organisations should factor in additional oversights when designing the AI system, for instance, by adopting “human-in-the-loop” for critical

decisions. That allows the human to be the ultimate decision-maker. For details on "human-in-the-loop", refer to Chapter 4.2.

- **Example**: Human-in-the-loop AI may enable an ideal symbiosis of human experts and AI models, harnessing the advantages of both while overcoming their respective limitations. This study was reported in a paper, "Human-machine partnership with AI for chest radiograph diagnosis" published in Nature on November 18, 2019. (Source: https://go.nature.com/3eyzAR8).

A. Demonstrate Accountability

Accountability drives responsibility. It ensures that there is always one party that is answerable for every aspect in an AI system's lifecycle – including design, development, deployment, and maintenance. It gives all stakeholders confidence in the system as well as in each other.

Accountability strives to make AI-driven decision making explainable, fair and accurate. It drives responsible design and development that is both privacy-preserving and secure so that the integrity of the AI system protects both the organisation and its users. It holds organisations answerable to the users of the AI system to protect its integrity and all sensitive data (personal data, decision outcomes in highly sensitive domains like healthcare, finance, legal and others).

Accountability drives responsible use of the AI system so that malicious intent by users can be deterred or detected. Note that being accountable does not mean that an organisation cannot make mistakes or programming or data errors. Instead, accountability means that organisations can be trusted to do the right thing, including being upfront and proactive when mistakes happen. Hence, being accountable is vital for any organisation that wishes to boost its trust with society and regulators.

The next section elaborates on what specifically organisations could do to build trustworthy relationships with stakeholders.

III. Suggested Best Practices

All AI systems have limitations. That may be due to the type of AI models deployed, the inherent nature of its training data, or both. Most people tend to blame the organisation that deploys or designs the AI system when things go wrong or decisions appear biased.

However, it is often difficult to discern where the fault lies; it may not be clear whether the problem arose from the deployment (flawed implementation) or design (inherent defect) of an AI system. When issues inevitably occur, it is not uncommon to find designers and deployment teams (where these are different parties) blaming each other.

Nevertheless, there are measures that organisations can put in place measures to ensure the safe and responsible use of the AI systems they develop or deploy. These measures can not only protect the organisation but also foster a greater sense of trust between stakeholders, assuring that due diligence processes are in place.

Communications is key to enhancing trust, especially communicating measures done correctly, legally and ethically. Succinctly communicate the rationale of the action taken and its impact to assure stakeholders about the organisation proactively identifying potential issues, evaluating them and addressing them promptly.

Here are some actionable measures which you could adopt to build stakeholder trust:

A. Acceptable Use Policies

Acceptable Use Policies (AUPs) are documents outlining the list of constraints and practices that a user must agree to abide by, to use a product or service. It sets rules and guidelines to inform users on how the AI system can be used responsibly. It includes information such as what users are allowed or not allowed to do and the possible actions which the organisation will pursue in the event of misuse.

AUPs ensure that the AI systems implemented get used within their design limits; it restricts the interactions between users and the AI systems to a well-defined boundary. Doing so

provides the first line of defence to deter against intentional manipulation of the AI system's integrity or performance. That could be due to malicious actions, such as corrupting the input data fed into the models. Developing AUPs is a necessary step, given past instances of AI abuse (such as chatbot systems being manipulated by irresponsible users to produce highly biased responses).

From a policy perspective, organisations should consider listing AUPs to safeguard against intentional manipulation or malicious use, and avoid a potential degradation of trust.

- **Example**: An AI chatbot launched in March 2016 function well initially. However, the chatbot soon began tweeting racist and sexual messages and had to be shut down shortly after its release. Investigations found that the chatbot had turned racist because some users intentionally fed it politically incorrect tweets, which then resulted in the AI chatbot learning and mimicking their undesirable behaviour.

a. Explainability of AI Decisions

Depending on the unique characteristics or requirements of the organisation's AI system, AUPs should address the following key topics:

Section	Description
Introduction & Agreement	This section highlights the AUP's purpose and coverage. Users agree to accept the AUP's general principles for responsible use of the AI system.
Intended Scope & Uses	This section describes the intended uses for which the AI system was developed, including the range of interactions permitted, and specific examples of those prohibited. Explanations should be accompanied by examples or use-cases to highlight the boundaries to the users.
Violations	This section states what constitutes illegal use, the rights of the organisation, how it will enforce the AUP, and the actions it will take upon breach by the user. Make it clear that the AUP applies to all users without bias or discrimination.
Miscellaneous Provisions	This section covers various clauses (ones that do not fall under the other three areas). The provisions here would also include the governing law. State the dispute resolution mechanisms to settle disputes under the AUP. Add details on avenues where users can seek clarification on the AUP.

b. Excessive Restrictions

For explicit prohibitions, such as illegal interactions, organisations should clearly state in their AUPs that such the organisation will report such infractions to law enforcement.

However, for other issues, especially ones near the threshold of anti-social behaviour, striking the right balance is essential. Avoid implementing an overly restrictive policy. That could be particularly problematic for platforms which advocate open and free sharing of information (such as research and peer review platforms).

- **Example**: Some AUPs provide clauses which specifically give acceptable use provisions to support open research across international boundaries, but highlights that any commercial use or for-profits are strictly prohibited.

c. Scalability

Technology typically progresses at a much faster pace than legislation or regulations. In due course, laws will catch up. Therefore, it is ideal that the AUP reflects the best practices required the responsible use of AI systems. Organisations should consider stating their AUPs in a modular structure. That will allow regular updating of the AUP to cater to increased types and range of their AI solutions that may be covered as well as any changes to laws or regulations.

- **Example**: Some AUPs consist of a broad overarching document with complementing sub-policies that address different issues. That makes scaling more manageable, as each sub-policy can be reviewed, amended, and approved individually.

d. Internal Stakeholders

Consider these issues when communicating with internal teams (technical development, legal, senior management):

- The extent of coverage:
  - Has the technical development team identified all the risks or potential misuse?
  - Have these been adequately covered in the AUP by the legal team?
  - Are the restrictions imposed by the AUP proportional to the risks and potential misuse?

- Does the AUP comply with existing government regulations or laws?

- Impact to self-interest:
  - How will the AUP protect the organisation's interests?
  - Are there any areas where the AUP might give a negative perception of the organisation?
- Monitoring and enforcement:
  - How do you determine whether a violation has occurred?
  - Have the criteria set for determining violations been validated?
  - What actions will the organisation take in case of a violation?
  - How will the organisation take action in case of a violation?

e. External Stakeholders

Consider these issues when communicating with external stakeholders (general public, media, NGOs, trade associations, others):

- The precision of communication:
  - Users should understand the language and terms used in the AUP. Keep it simple.
  - Is the AUP concise and simple enough to be understood by laymen?
- Critical points in the AUP:
  - What are the restrictions imposed on users? Are users well-informed about them?
  - Is the AUP explicit about what is considered acceptable or unacceptable use?
  - Is the AUP explicit what constitutes a violation to the AUP?
  - Does the AUP have examples of types of violations to convey the points?
  - Are the legal implications in case of misuse explicitly detailed?

B. Safeguard Against Cyberattacks

AUPs are for users who don't have explicit malicious intent in mind. Cybercriminals are ones that do. It is, therefore, crucial for organisations to safeguard AI systems against cyberattacks

and malware. Organisations should assure external stakeholders that they are taking proportionate measures to protect users (including compliance with recognised security standards).

Cybersecurity is a vast topic. The intent here is to offer some vital hygiene points about cybersecurity. Cyberattacks could come in several forms, the two most common being "input attacks" and "poison attacks":

- **Input Attacks**: These are situations where cyber-criminals manipulate the inputs into the AI system to change its output. Examples include altering digital images with pixel-level deviations that are undetectable by the human eye to modify the AI system's output classification.
- **Poison Attacks**: These work by corrupting various critical aspects of the AI system, such as by manipulating its training data to have the AI system learn or disregard a different pattern, or corrupting its algorithm to alter the logic process.

a. Underlying Flaws

Cyber-criminals or malware attacks an AI system's underlying flaws and vulnerabilities. It is the organisation's responsibility to take proactive action to ensure the integrity of its AI system and data.

- **Example**: In early 2020, experts discovered a severe flaw in a standard operating system that allowed the spoofing of digital signatures. That could potentially lead to malicious code passing off as legitimate software. The software firm promptly released an emergency patch. Had it not done so, an extensive amount of damage would have resulted, leading to a huge trust deficit in its products and reputation.

b. Active Risk Mitigation

Active risk mitigation throughout an AI system's lifecycle is necessary. From a technical perspective, companies can do this with internal protocols to mitigate the risks of cyberattacks. Do this across the AI system's lifecycle – planning, design, implementation and maintenance.

The risk of not doing so could be huge, including loss of revenue, costs of fallout and remediation, loss of stakeholder trust, and damage to reputation. Consider investing in the robustness of AI systems as fundamental that is well worth the extra cost and effort.

c. Internal Protocols

Internal protocols should address issues across the entire lifecycle of an AI system. At each stage, stakeholders (including research scientists, engineers and managers) need to carry out risk identification regularly, as well as mitigation and response for various aspects of the AI system, including data sources, bias, and tested AI models. Here's a checklist:

Considerations	Description
Identifying vulnerabilities	What are the vulnerabilities in with the various aspects of the AI system? Can they be tampered with, corrupted or manipulated by adversaries? **Example**: If an AI model relies on and learns solely from one training dataset, adversaries could corrupt its dataset to prevent the model from learning specific patterns (such as preventing it from recognising the "stop" sign for autonomous vehicles).
Potential damage	What are the potential damages to the integrity of the AI system? That could arise from lack of rectification or external exploits of flaws. **Example**: The computer vision system used in a brand of autonomous vehicles possesses an inherent flaw. It allows the training dataset to be hacked and manipulated. How would stakeholders trust the organisation that provides the technology, the AI system, or the vehicle company?
The current state of play	What are the current techniques which can be reasonably and quickly adapted or embedded into the AI system to boost its robustness against identified risks? **Example**: For applications that place a strong emphasis on security, reasonable levels of "defensive programming" could be incorporated in the development stages to ensure that the system can withstand most types of cyberattacks.
Risk strategy and response	What are the organisation's risk strategies and responses in the event of a cyberattack? How is its effectiveness measured? What contingency plans are in place if these responses fail? **Example**: If an autonomous vehicle's navigation system is compromised, risk responses could be to identify the affected vehicles (through software patch history for instance) and ground all impacted vehicles until the issues get resolved. Whether organisations can build stakeholder trust in times of crises depends on the speed, level of response, and follow-up actions.

d. Ethical Hacking

Some vulnerabilities cannot get identified up during the development phases, and therefore its risk and impact may not be fully apparent. Organisations should assess the security of an AI system through the perspective of a third party or even an adversary. Some ideas worth considering:

- **Red-Teaming**: This involves having independent teams simulate a multi-layered attack on the AI system to detect its weaknesses and vulnerabilities to manipulate the outcome. Ted-teaming can help secure the organisation's network and how it is likely to perform against a real-life attack. Red teams could comprise internal and external experts across multiple domains. Consider each member's skillset when assembling the red team so that every aspect of the AI system and its development lifecycle gets rigorously tested.
- **Open-source engagement**: Open-source development promotes transparency in the system because the source code is publicly available. Bugs and flaws get more easily picked up as there are often many people scrutinising the source code of open-source software. The added scrutiny provides an additional degree of protection.

e. Building Trust Internally

Consider these points to build trust with internal stakeholders to protect against malware:

- What are the potential logic flaws, exploitable bugs or backdoors?
- How could they compromise the AI system, and how are they being addressed?
- What are the potential ways in which data could get manipulated for malicious intent?
- How are they being addressed?
- What are the encryption algorithms used at each stage of the development lifecycle?
- How will their suitability and performance be evaluated?
- What are the data security protocols in place for all types of data, not just personal data?
- Are any penetration tests done to check for vulnerabilities actively?
- How are penetrations tests carried out and evaluated?
- Which KPIs (key performance indicators) are used to evaluate the effectiveness of measures

- What is the response plan in the event of a cybersecurity breach?
- Who will be involved in breach response, and what are their respective roles?
- How will the organisation stay up to date with the evolving cybersecurity landscape?
- How will security know-how and updates be disseminated, and how frequently?

f. Building Trust Externally

Consider these points to build trust with external stakeholders to protect against malware:

- What are the types of personal data collected?
- How is personal data protected and their privacy preserved?
- How does the organisation stay ahead of the curve to mitigate evolving types of cyberattacks?
- How and where can external stakeholders report potential flaws, bugs or errors?
- How does the organisation communicate intent and desired outcomes of security measures?
- What are some of the restrictions required due to stricter security measures?
- How are these stricter measures implemented, and how does it impact users?
- How do you communicate responses before, during and after a cybersecurity breach?
- How do you demonstrate initiative and proactive measures to safeguarding against cyberattacks?
- What has the organisation done to resolve breaches and mitigate the damage done?
- Post-breach, how best should the organisation repair trust lines?
- Do you encourage open-source efforts to enhance the security of the AI system?

C. Safeguard Against Cyberattacks

Building trust within the organisation takes time. Often, years of meticulous transformation and employee mind-set shift are needed. That is where the culture and moral values (or corporate ethics) of the organisation come into play. Here are some tips on how to promote a culture of trust:

a. **Sacrifice Silo**: Shift from a silo mentality to a whole-of-organisation mind-set. In many organisations, technical teams work in silos and have minimal interaction with other groups. That means useful feedback gets missed due to lack of communication channels, lack of oversight on what other teams are doing, and workplace politics. These factors act as barriers to a more collaborative workspace, which limits the confidence and trust each individual has in the organisation. If individuals do not see the bigger picture, it will be difficult to expect them to appreciate their role in the organisation. If the organisation doesn't trust itself, how will it expect others outside the organisation to trust them? Build trust and collaboration within internal teams first and foremost.

b. **Good Governance**: Focus on good governance within the organisation. Get executives from senior management to be on oversight committees. Set up a chain of command to make decisions at each level of the AI development or deployment cycle. Include AI technical experts at every decision-making level, especially in oversight committees. Don't delegate critical technical issues, such as those concerning the use of personal data, down in the chain. From the organisation's perspective, this establishes clear accountability and enables the company to plan and execute its AI strategy effectively. Good governance practices will help build a high level of confidence and trust within the organisation, particularly with senior management. That enables the organisation to deal with external stakeholders, such as government regulators and shareholders.

c. **Robust Reporting**: As an organisation advances in its development or use of AI, there will likely be a slew of AI models in operation in use by different teams for different purposes. In short, the AI ecosystem will probably be increasingly problematic. This issue calls for the creation of standards, protocols or frameworks that to make uniform assessments. Such consistency ensures that crucial information regarding the use of various AI models can be made accessible and used and trusted by different teams internally. Specifically, this standardisation could come in the form of an inventory.

In a paper "Model Cards for Model Reporting", published in January 2019, the authors recommend that AI developers include a "Model Card" listing necessary details about AI models: who the developers are, their contact details, and model versions. It could also include details about data (datasets used for training the model, why they were chosen, how the pre-processing was done), and intended use (what was it developed for, who are the primary users). Knowing what the model is designed or developed is key to trusting its use. (Source: https://arxiv.org/pdf/1810.03993.pdf).

d. **Effective Engagement**: Proactive engagement is an integral part of communicating and building public trust; stakeholders appreciate getting regular updates. Organisations can solicit feedback on available platforms, including suggestions for improvements, or share what is being done, and provide stakeholders with the satisfaction of being heard. You can achieve this through various means, such as maintaining an active online community, roadshows, hackathons, among others. Ultimately, organisations need to exercise judgement on the level of detail that they should communicate to the public. When communicating with the general public, media, and external parties, maintain transparency and openness – welcome feedback and suggestions.

    - **Example**: To better engage the community, a technology company created a tech blog with contributions from employees and management. The blog not only discusses issues relevant to the products and services offered by the organisation, but also technical problems, such as how they rationalise supply and demand through data, and improvements to their ML capabilities, and how the benefits help end-users. This initiative helps foster trust between the organisation and the community. The blog conveys issues to users in bite-sized articles written in simple language. Each concept is well defined and explained before diving into more-advanced discussions. It establishes a line of communication between the organisation's data scientists, engineers and product managers and end-users.

e. **Internal Issues**: Encourage communication, collaboration, innovation and trust within the organisation. Break down key barriers hindering cross-disciplinary teams from working closely together. Office politics and a "blame game" culture fosters a strong sense of mistrust and discourages collaboration within teams. Offering incentives for cross-departmental collaboration could boost innovation since AI can impact the entire organisation. Establishing a specific governance structure of oversight could help AI developers under

D. Communication Tips

It is easy to ramble on and lose the audience, especially with technical explanations. Consequently, it is challenging to build trust with this disengaged audience. When strategising how an organisation should engage with the audience, clarity is crucial. Achieve this by communicating succinctly; get the message across with the least effort while achieving the maximum impact.

a. **Communicate Clearly**: Organisations should be clear about the message they need to convey to their stakeholders. Who are the critical stakeholder groups? What are the pertinent issues that bug them the most? What do they need to know? Taking a broader view, the three aspects (explainability, fairness and accuracy) of explaining AI outcomes are all equally important. However, depending on the situation, individual stakeholders may place greater emphasis on one particular aspect.

    - **Example**: In food delivery platforms, restaurants list their services; users within a certain vicinity can view and order from these listings. However, several restaurants noticed that their listings do not appear to users on some days. Discussions between the restaurants and the food delivery platform found that some listings were "demoted" because some restaurants paid higher commission rates. The AI recommendation engine promoted these higher-paying restaurants to users. The platform could argue that one of the factors influencing listing priorities was the amount of commission paid.

However, restaurants did not know this and felt betrayed. In this instance, all three aspects were lacking – explainability, transparency, and fairness.

b. **Discuss the How**: Besides the "what" of communications, discuss the "how". Write the content based on the target audience. Knowing the audience is, therefore, critical to decide how the content or message will be delivered. During the development process, provide technical teams with the content (explaining how the AI works and how it arrives at decisions) to evaluate technical issues and risks. To the general public, such details may be unnecessary at best, and confusing at worst. Simplify complex concepts and present them as infographics, user-stories, metaphors or analogies if possible.

c. **Be Sensitive**: Stakeholder dynamics evolve, influenced by economic, sociological and political factors. Keep a close watch on stakeholder dynamics. If not managed properly, the timing and tone of communication can result in adverse outcomes with messages being ignored, clouded, distorted or misconstrued. Empathy plays a critical role in building relationships by demonstrating that the organisation identifies with the challenges and worries faced by stakeholders. Empathy helps allay anxiety and fear, gives clarity on the rationale and justification behind actions, and may even help reveal underlying issues that are not verbalised.

    - **Example**: A global technology company wants to market its cybersecurity solutions. Instead of highlighting the impact of breaches and how their products address them, they created a user journey that allows prospects to view the issues from a hacker's perspective. It provides a step-by-step view of how hackers gain network access and the damage they cause. Customers closely relate to this narrative, and the company reassures them that its solutions address the vulnerabilities. That demonstrates that the company shares the worries and fears of its customers, and is mitigating their concerns.

IV. Checklist

A. Internal Stakeholders

Consider the following when communicating between technical teams and management decision-makers:

- Bridging the knowledge gap:
  - Key decision-makers may comprise both technical and non-technical senior managers.
  - Besides basic information, convey technical or unique information about the AI system.
  - Give them sufficient background information to make informed judgements.
- Conveying AI functionality:
  - How is AI being used in the system?
  - What are the requirements to use AI effectively and ensure optimum performance?
  - What are the risks, and how are they being addressed?
  - Convince decision-makers of the need for AI. What business problems does it solve?
  - What is the value-add to the organisation by using AI?
  - How does this compare with similar solutions in the market?
- Explaining AI outcomes:
  - Different decision-makers have different attitudes towards the use of AI.
  - Some are favourable, while others may be resistant to certain aspects of AI adoption.
  - What are their key concerns, and how can they be addressed?
  - How does the use of AI align with the organisation's AI or business strategy?
- On explainability of AI decisions:
  - What is the level of explainability that is supported by the AI system?
  - Does the AI system adhere to government regulations or guidelines on AI explainability?
  - How does the logic of the AI system work?

- What are the parameters that contributed most to the outcome?
- How was the data assessed and cleaned before being used?
- Are there constraints or conditions established for its intended uses?
- What are the critical decisions made at the various decision-making levels so far?
- How will the organisation present this workflow to external stakeholders?

- On the fairness of AI decisions:
  - Protocols to mitigate bias and discrimination in datasets? How are they verified?
  - Performance indicators to show that the outcomes are fair?
  - Would a human decision-maker have arrived at a different decision?
- On the accuracy of AI decisions:
  - Protocols to ensure that data used in the pipeline is accurate and representative?
  - How are they being verified?
  - What is the acceptance criteria for success? How does the organisation determine it?
  - What are the performance indicators showing that the outcomes derived are accurate?

B. External Stakeholders

Consider the following when communicating with external stakeholders:

- Bridging the knowledge gap:
  - External stakeholders range from the general public to government regulators.
  - Don’t assume the general public to all be well-versed on the intricacies of AI.
  - Expect government regulators to have a relatively good understanding of AI and AI ethics.
  - What is the level of understanding of your target stakeholder groups?
  - How much does each stakeholder group need to know?
  - What are the critical knowledge gaps for specific stakeholder groups?
  - What is the best medium to best reach them?

- How detailed should your communications be for specific stakeholder groups?

- Conveying AI functionality:
  - Convey a basic understanding of the AI system. How does the AI work?
  - Where and when is AI used? How is data collected, used, and secured?
  - Convey the impact on external stakeholders, particular end-users.
  - How does the use of AI solve the problem? What is the value-add to them?
  - Convey the safety and security issues to get users to trust the system.
  - Convey sufficient details on the cybersecurity risks and steps taken to mitigate them.
  - Convey details on the robustness of your AI system design.
- Explaining AI outcomes:
  - Convey key parameters on explainability to stakeholder groups.
  - Explain simply and clearly so that stakeholders can rationalise the outcome.
  - Explain how outcomes are arrived at, especially those that are or seem unfavourable.
  - For example, why a loan was denied to a person; empathise with the person's feelings.
  - Suggest options or actions that people can take to change the outcome to a positive one.
  - When disagreements arise, how can users challenge the outcome?
  - Convey the fairness and transparency of outcomes in simple language.
  - Convey that the outcomes are free of bias and discrimination.
  - Convey that the outcomes are accurate and similar to ones if a human were to decide.