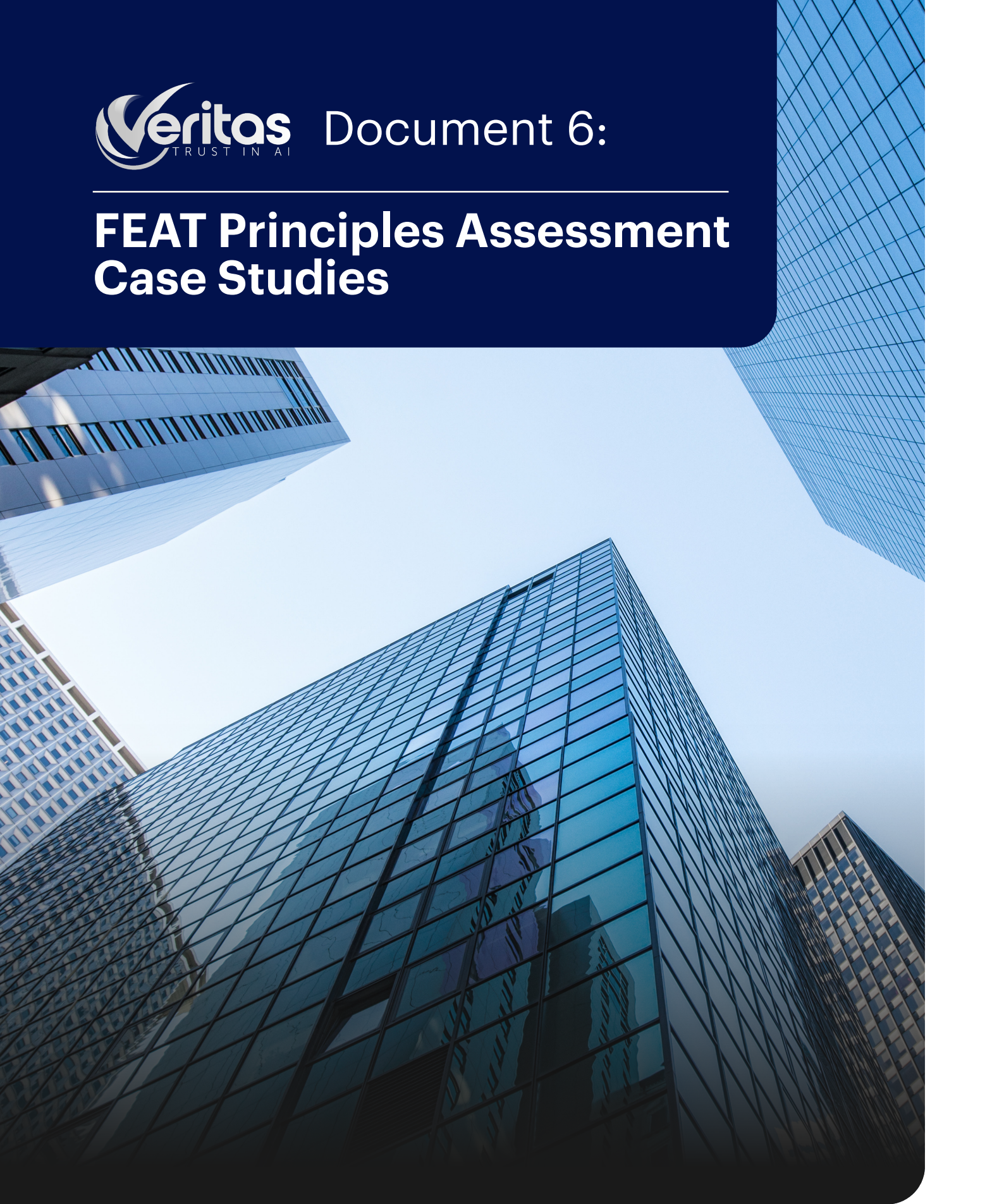




Document 6:

FEAT Principles Assessment Case Studies



Monetary Authority
of Singapore



Swiss Re

Table of Contents

Case Study 1: Google Pay	4
1. Background of Google Pay India Use Case	5
2. Google's AI Principles, and Application of FEAT Ethics and Accountability Principles Assessment Methodology	7
• Discussion on Principle 8: Firms using AIDA are accountable for both internally developed and externally sourced AIDA models	
3. Application of FEAT Transparency Principles Assessment Methodology	17
4. Application of FEAT Fairness Principles Assessment Methodology	19
• Discussion on Principle 2: Use of personal attributes as input factors for AIDA-driven decisions is justified	
5. Considerations and Recommendations	22
 Case Study 2: Swiss Re	 23
1. Background	24
2. Use Case	25
3. Overview of Transparency Assessment Methodology.....	27
4. Transparency Assessment and Explanations	28
Step 0 - Internal Standards (Transparency)	
Step 1 - Define System Context and Design	
Step 2 - Prepare Input Data	
Step 3 - Build and Validate AIDA System	
Step 4 - Monitor AIDA System	
5. Reflection	38
 References	 40

Application of Fairness, Ethics, Accountability and Transparency (FEAT) Assessment Methodologies to Detection of Fraudulent Payments



Google aspires to create technologies that solve important problems and help people in their daily lives. We are optimistic about the incredible potential for artificial intelligence (AI) and other advanced technologies to empower people, widely benefit current and future generations, and work for the common good. We believe that these technologies will promote innovation and further our mission to organize the world's information and make it universally accessible and useful. We recognize that these same technologies also raise important challenges that we need to address clearly, thoughtfully, and affirmatively.

With that objective in mind, Google has been participating in the Veritas consortium facilitated by the Monetary Authority of Singapore (MAS) to promote fairness, ethics, accountability and transparency (FEAT) in the use of AI and data analytics (AIDA) in Singapore's financial sector. Following our engagement in Veritas, Google has applied the FEAT Principles Assessment Methodologies published by Veritas in February 2022 to Google Pay's detection of fraudulent payments in India. This case study summarizes our experience of mapping our AI Principles governance, processes and practices against the FEAT Principles Assessment Methodologies, our learnings for Google Pay's detection of fraudulent payments in India, and our recommendations to Veritas on how it may consider refining the FEAT Principles Assessment Methodologies.



1. Background of Google Pay India Use Case

Unified Payment Interface (UPI) is a convenient and secure mode of making digital payments in India without sharing bank account details, similar to PayNow in Singapore. Google Pay India is a UPI-enabled digital payments application (app) that processes over 3 billion transactions worth almost ₹4.8 trillion (US\$59 billion) a month, according to data published by the National Payments Corporation of India (NPCI) in March 2023. Unlike digital wallets, Google Pay India allows users to send and receive money to other users and merchants directly into bank accounts without having to go through a digital wallet. The app also enables non-users to receive payments.

We note that most of the case studies conducted by Veritas thus far were based on either hypothetical use cases or actual use cases that have yet to be operationalized. In contrast, this is a 'live' use case to examine the complexities of applying the FEAT Principles Assessment Methodologies to existing AIDA systems.

We selected a use case in India for two reasons. First, the use of AIDA for Google Pay India is in a more mature stage, due to the volume of transactions compared to in Singapore. Second, while [NITI Aayog has developed principles for responsible AIDA in India](#)¹, there are no regulations or guidelines specific to the financial sector, unlike in Hong Kong and Singapore. This presents an opportunity for Google Pay India to contribute to the development of an industry framework by demonstrating Google's thought leadership in responsible AI.



In this use case, Google Pay India uses AIDA to detect and deny potential fraudulent payments, warn users of suspicious transactions, and flag users who are deemed to be perpetrators of fraudulent payments to human reviewers for potential account suspension. Diagram 1 below summarizes the process for fraud detection:

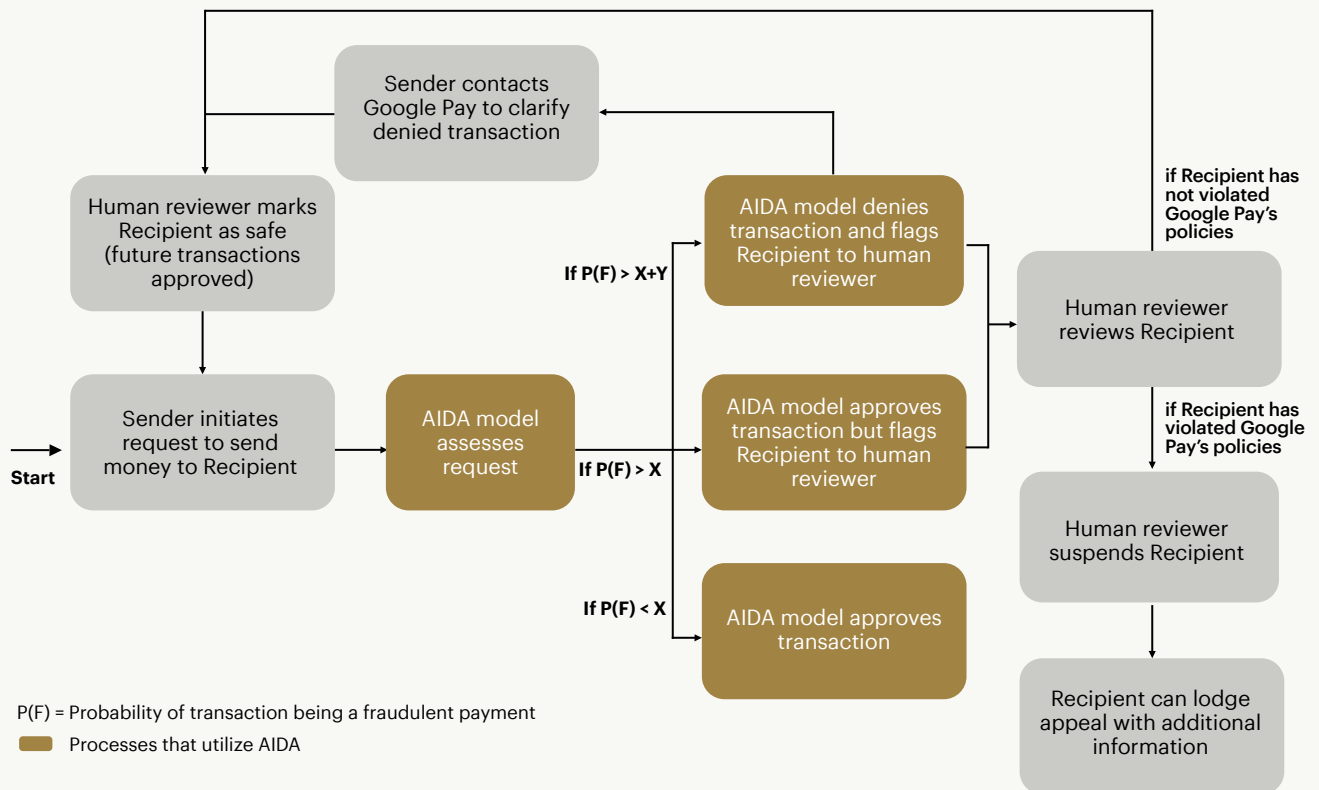


Diagram 1: Google Pay India Fraud Detection Process Flow Chart

For the purpose of this illustration, we focus only on the scenario in which both the Sender and Recipient are users of Google Pay India, and the Recipient is the suspected perpetrator of fraudulent payments since the Recipient is much more likely to be the beneficiary of such transactions. In practice, Google Pay India also addresses other scenarios beyond this.

In the subsequent sections, we share our experience applying the FEAT Principles Assessment Methodologies published by Veritas in February 2022 to Google Pay's detection of fraudulent payments in India. We start with the Ethics and Accountability Principles as these are more focused on organizational implementation with broader implications, followed by the Fairness and Transparency Principles, which are more specific to the use case.

2. Google's AI Principles, and Application of FEAT Ethics and Accountability Principles Assessment Methodology

Principle 5: Use of AIDA is aligned with the firm's ethical standards, values and codes of conduct.

Google has established value-based AI Principles governing our assessment and use of AI. These principles set out Google's commitment to develop technology responsibly. Google has also established specific application areas that we will not pursue because of those applications' inherent conflict with Google's AI Principles. For example, Google will not design or deploy AI where the principal purpose will be to cause harm or directly facilitate injury to people.

Google's AI Principles, listed below, reflect Google's belief that AI should:

- **Be socially beneficial.** As Google considers potential development and uses of AI, we consider a broad range of social and economic factors, and proceed where we believe that the overall likely benefits substantially exceed the foreseeable risks and downsides.
- **Avoid creating or reinforcing unfair bias.** Google seeks to avoid unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious beliefs that may be brought about through use of AI algorithms or datasets.
- **Be built and tested for safety.** Google continues to develop and apply strong safety and security practices to avoid unintended results that create risks of harm.
- **Be accountable to people.** Google subjects our AI technologies to appropriate human direction and control. We design our AI systems to provide appropriate opportunities for feedback, relevant explanations, and appeal.
- **Incorporate privacy design principles.** Google gives opportunity for notice and consent, encourages architectures with privacy safeguards, and provides appropriate transparency and control over the use of data.
- **Uphold high standards of scientific excellence.** Google works with a range of stakeholders to promote thoughtful leadership in this area, drawing on scientifically rigorous and multidisciplinary approaches.
- **Be made available for uses that accord with these principles.** Google works to limit potentially harmful or abusive applications.

Through a range of governance and review processes described in the rest of this section, Google Pay ensures that its use of AIDA, including to detect fraudulent payments in India, adheres to these AI Principles. Subsequent sections of this paper also illustrate how a few of these AI Principles are applied to the specific Google Pay India fraud detection use case.

Principle 6: AIDA-driven decisions are held to at least the same ethical standards as human-driven decisions.

As described above, Google's AI Principles focus on accountability to people, as well as avoiding the creation or reinforcement of unfair bias. These principles reflect the same ethical standards that underpin Google's human-driven decisions. Additionally, our AIDA models are also subject to appropriate human direction and control.

Moreover, Google approaches our use of AIDA with the belief that AIDA-driven decision-making should be benchmarked against comparable human decision-making. This is particularly relevant when setting minimum performance standards. For instance, Google expects our AIDA systems to match or exceed the same accuracy and fairness standards that apply to human-based approaches. Like any system, including human-based processes, AIDA systems are not perfect. They do, however, offer the opportunity to dramatically improve on current human-based decision making. Thus, the operational benchmark for AIDA systems should not be perfection, but instead the performance of comparable current processes (if existing) or an available human-powered alternative.

There is a real risk that innovative uses of AIDA could be precluded by demanding that AIDA systems meet a standard that far exceeds that required of human-based approaches. Sometimes this may be deliberate due to artificial protectionist constraints. More often though, it is likely due to a lack of understanding about hidden flaws in existing human-based decisions, as well as people's natural tendency to be more forgiving of mistakes made by a human versus a machine.

Context is also key. There may be good reasons to deviate from an expectation of parity. In some situations, a lower level of accuracy may be acceptable, such as if an urgent response is needed, the cost of inaction is high, and there are simply not enough qualified people on hand to do the job (eg, helping triage medical screening in crisis settings). In other contexts, the reverse may hold, such as if there are plenty of qualified people willing to do the work, using an AIDA system might only be justified if it was shown to perform significantly better (eg, self-driving cars that have far fewer accidents than human drivers). Similarly, fairness is a vital consideration; even if an AIDA system performs more accurately and reliably across the general population, that may not be sufficient to justify its use if it performs significantly worse than existing approaches for certain subgroups.

Within the context of Google Pay India, AIDA-driven decisions that detect a potential fraudulent transaction and deny it are automatically escalated to a human reviewer to assess if the Recipient had violated Google Pay's policies or participated in any financial fraud or scam, which could result in the suspension of the Recipient's user account. A random sample of these human reviews are then subject to audit by a team of senior human reviewers.

When viewed in its entirety as an AIDA system (instead of just the AIDA model), this Google Pay India use case illustrates how, when AIDA is used to complement humans, AIDA systems involve both AIDA-driven and human-driven decisions. In such situations, it is not feasible to have a set of ethical standards for AIDA-driven decisions and another set for human-driven decisions; both must be held to a single set of ethical standards.

Principle 7: Use of AIDA in AIDA-driven decision making is approved by an appropriate internal authority.

Principle 9: Firms using AIDA proactively raise management and board awareness of their use of AIDA.

As reflected in Diagram 2 below, Google maintains a three-tiered governance structure to ensure responsible use of AI.

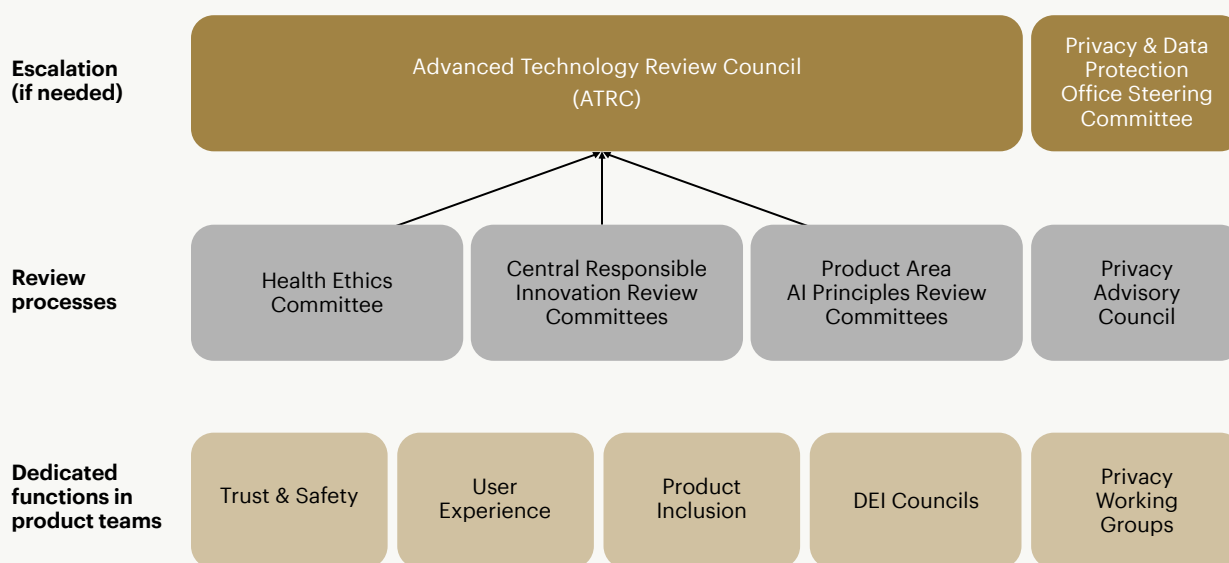


Diagram 2: Google's AI Principles Governance Structure

- **First Tier – Product Teams:** Google's product teams include dedicated user experience, privacy, and trust and safety experts, who provide deep functional expertise consistent with Google's AI Principles.
- **Second Tier – Review Bodies:** Google has established a set of dedicated review bodies and expert teams to serve as a resource for product teams. This includes specialized reviewers with expertise in specific areas, such as Cloud, with expertise in enterprise AI offerings, and Devices and Services, with expertise on how AI is used responsibly within hardware. There is also a Health Ethics Committee that reviews health and medical-related AI research and development across the company. All Google employees are encouraged to engage with the AI Principles review processes throughout the project development lifecycle. In addition, the Privacy Advisory Council (PAC) reviews all projects for potential privacy concerns, including (but not exclusively) issues related to AI. Last year, Google expanded the Responsible

Innovation team, which is the central operations team for AI Principles implementation across Google's product development lifecycle, and moved it into our company-wide Office of Compliance and Integrity for more centralized governance across all Google product areas. This move reflects the growing maturity of our governance strategy.

- **Third Tier – Escalations:** Google has also established the Advanced Technology Review Council (ATRC), a rotating committee of senior product, research, and business executives representing a broad cross-section of Google. The ATRC addresses complex cases that could affect multiple products, or set a precedent because the technology is so new. It also makes decisions on urgent escalations, and establishes company-wide policies impacting multiple product areas. This involves making challenging decisions that require deeply considering the trade offs-between ethical risks of certain new applications and potential business opportunities, prioritizing social benefit.

It is important to note that, in line with our risk-based approach, the reviews and decisions get fewer as we move up each tier of this structure. Google's objective is to embed our AI Principles into the product development process at the onset. The product teams are therefore equipped with the requisite functional expertise to ensure responsible use of AI. Where the situation calls for a deeper review beyond the expertise of the product team, the project gets brought to the relevant review body. In very limited cases where the project is highly complex and a decision could set a policy precedent, it is escalated to the ATRC. Notwithstanding this process, any employee of Google can request an escalation of a project.



Principle 10: Data subjects are provided with channels to enquire about, submit appeals for and request reviews of AIDA-driven decisions that affect them.

Principle 11: Verified and relevant supplementary data provided by data subjects are taken into account when performing a review of AIDA-driven decisions.

Google's AIDA systems are designed to provide appropriate opportunities for feedback, appropriate explanations, and appeal. Google strongly endorses the notion that developers of AIDA systems have a responsibility to provide guides and documentation to deployers and users to ensure they can operate products safely and responsibly, especially when deployers seek to use those systems in a high-risk application. That said, the benefits of these disclosures must be balanced against protecting innovation and the intellectual property of an AIDA system. Moreover, it is essential to consider the nuances of implementing practical pathways for recourse by data subjects when working with AIDA systems, especially those systems whose benefits derive from significant speed and scale.

Google Pay processes over 3 billion transactions each month in India alone. Given this scale, we deem it appropriate and ultimately beneficial for the AIDA model to handle certain tasks without data subject engagement (eg, denial of individual transactions), while providing engagement opportunities – which may involve time-intensive human review – for tasks with potentially greater impact. Specifically, data subjects may submit appeals and request reviews of account suspensions. This avenue allows data subjects to submit additional materials to support their appeals, as seen through two sample mechanisms in Diagram 3 below.

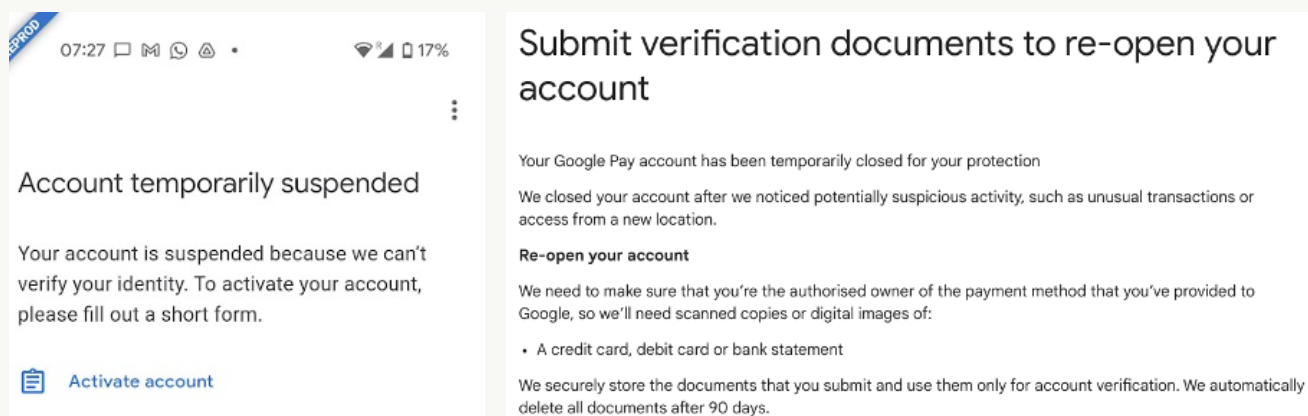


Diagram 3: Notification of Account Suspension & Appeal Process

Discussion on Principle 8: Firms using AIDA are accountable for both internally developed and externally sourced AIDA models

While the principle that firms using AIDA should be accountable for both internally developed and externally sourced AIDA models is sound, it becomes trickier and far more complex when one considers the execution of responsibilities. Phase 2 of Veritas alludes to this challenge, specifically in the FEAT Fairness and Transparency Principles Assessment Methodologies, which recommended the treatment of AIDA usage via third-party service providers as a topic for future consideration. In particular, when a FSI (ie, the AIDA system owner) leverages a third-party service provider (ie, the AIDA system developer) for AIDA, the third-party AIDA system developer will not have full visibility into the AIDA system owner's use of AIDA due to privacy and security obligations, nor will the AIDA system owner have full visibility into proprietary systems and services of the third-party AIDA system developer.

This challenge is not unique to Singapore. In Hong Kong, the Guidance on the Ethical Development and Use of Artificial Intelligence issued by the Office of the Privacy Commissioner for Personal Data explicitly states that the Guidance "primarily provides recommendations for organizations that develop their own AI systems". The Guidance adds simply that organizations using AI systems developed by third-party service providers "should take appropriate steps to ensure that the principles and practices recommended in this Guidance are adhered to".

However, AIDA system owners that have relied on AIDA models developed by third-party AIDA system developers and have attempted to adhere to various AIDA-related regulations or guidelines, be it in Singapore, Hong Kong or the EU, would have immediately realized that such a general approach is untenable. Instead, AIDA system owners should adopt a risk-based approach and be allowed to negotiate roles and responsibilities by contract, when appropriate.

Using the transparency principle for illustration, if regulators require the AIDA system owner to have full transparency into AIDA models that the AIDA system owner uses but are developed by a third-party AIDA system developer, these requirements could expose the AIDA system developer's intellectual property and business confidential information, which may then constrain innovation as AIDA system developers will be less incentivized to serve the market and hinder the development of AIDA. There are many instances where the AIDA system owner would not require full transparency into the AIDA models developed by third-party AIDA system developers.

We must recognize that transparency is not an end in itself; it is a means by which to enable accountability, empower users, and build trust and confidence. In designing transparency requirements, regulators should consider what they are trying to achieve and how best to meet those goals in a given context.

For example, different stakeholders will require different forms of transparency under different circumstances. Requirements should be tailored to ensure that the information is actionable and presented when stakeholders want it, in terms they can understand, and without extraneous details that can be distracting or confusing. In some contexts, detailed information about individual decisions may be important; whereas in others, general information about how systems work or how they are developed may be more beneficial or appropriate.

Regulators should also carefully balance the desire for transparency with other important objectives, such as efficiency, safety, security, and privacy. Importantly, some forms of transparency that are intuitively appealing can carry some of the most significant risks, and provide little actual benefit in terms of enabling accountability and building trust.

For example, disclosure of source code or individual user data may provide little insight into how a system works or why it made a given decision, but could enable abuse or exploitation of systems, and carries significant risks to user privacy and intellectual property.

Google plays the roles of both a FSI – through Google Pay – and a third-party service provider – through Google Cloud. This is why we are deeply committed to shaping an equitable framework that satisfies the interests of the key parties, including the AIDA system owners (ie, the FSIs in this case), the AIDA system developers (ie, the third-party service providers), and the regulators.

As a starting point, we could borrow the shared responsibility framework that is widely adopted by the cloud computing industry. Under this framework, the responsibilities for securing every element of the cloud environment are clearly delineated between the cloud service provider and the user organization, with that delineation differing across various types of cloud services.

	On-premise data center	Infrastructure as a service	Platform as a service	Software as a service
Content	●	●	●	●
Access policies	●	●	●	●
Usage	●	●	●	●
Deployment	●	●	●	●
Operations	●	●	●	●
Access & authentication	●	●	●	●
Network security	●	●	●	●
Operating system	●	●	●	●
Audit logging	●	●	●	●
Network	●	●	●	●
Storage & encryption	●	●	●	●
Hardware	●	●	●	●
Physical security	●	●	●	●

● FSI ● Service Provider

Diagram 4: Shared Responsibility Framework for Cloud Computing

Most regulators adopt an outsourcing regulatory approach to cloud computing and hold their regulated entities accountable for the latter's IT systems, including elements that are outsourced to cloud service providers. However, even as they hold their regulated entities accountable, regulators recognize the shared responsibility framework. Using the regulator's right to audit as an example, regulators typically do not expect their regulated entities to be able to unilaterally grant the regulators access to the outsourced physical infrastructure, which is the responsibility of the cloud service provider.

The underlying principle of this delineation of responsibilities is who has *control* over each specific element. Take Infrastructure as a Service (IaaS) as an example: Since the cloud service provider has *control* over the data center, servers, network and other infrastructure, the cloud service provider should be responsible for the physical security of these elements. Similarly, since the user organization retains *control* over the data – even if it is stored in the cloud service provider's infrastructure, the user organization should be responsible for security of the data in the cloud.

We can also apply this underlying principle to the context of externally sourced AIDA models. Take general-purpose AIDA solutions as an example, which are “off-the-shelf” AIDA solutions typically developed by AIDA system developers and then deployed by AIDA system owners to support a wide variety of applications and uses with a range of potential benefits and risks. In such situations, the AIDA system developer has control only over the base AIDA model, while the AIDA system owner has control over how the AIDA model is deployed for a specific use case, how the AIDA model is retrained, and what data is used to retrain the AIDA model. This is more than a nuance because the AIDA system owner will be the one identifying how to use the solution, and the attendant benefits and risks.

The AIDA system developer should therefore be responsible for providing documentation and transparency artifacts, such as [AIDA model cards](#)², which may include guidance on how the AIDA model should or should not be used. The AIDA system owner can then update these documentation and transparency artifacts to ensure they appropriately reflect how the AIDA model is deployed for a specific use case, how the AIDA model is retrained, and what data is used to retrain the AIDA model.

In short, the AIDA system developer is responsible for pre-deployment compliance obligations (where it has control), while the AIDA system owner is responsible for post-deployment compliance obligations (where it has control). Using the cloud computing analogy again, just as how regulators do not expect the FSIs to hold the keys to the cloud service providers' data centers, they should similarly not expect the FSIs to possess the source codes to AIDA models developed by third-party AIDA system developers.

We therefore encourage regulators to provide clear guidance that ensures that the relevant requirements are allocated appropriately to the entity who is best suited to control the risks that these requirements seek to address. Such requirements include the provision of clear and appropriate explanations, when required. Such regulatory guidance should also allow requirements to be allocated through any contractual agreement between the AIDA system owner and the third-party AIDA system developer.

Borrowing the fundamental principle of shared responsibility from the cloud computing industry, we illustrate at a very high level what a shared responsibility framework could look like in the context of AIDA. There are at least three broad categories of externally developed AIDA, as set out in the FEAT Transparency Principles Assessment Methodology:

1. **AIDA-model as a service**, where the FSI purchases an AIDA model developed by a service provider for use within the FSI (eg, an anti-money laundering system). This is similar to the aforementioned general-purpose AIDA example, where the service provider has control only over the base AIDA model, while the FSI has control over how the AIDA model is deployed for a specific use case, how the AIDA model is retrained, and what data is used to retrain the AIDA model.
2. **AIDA-processing as a service**, where the FSI outsources a service to a service provider, which uses AIDA to deliver that service (eg, contract review by a professional services firm). In this scenario, AIDA is deployed as part of the process of making a decision. The service provider has full control over the AIDA model, including how it is deployed and the data used to train the AIDA model (assuming the service provider does not further outsource the development of the AIDA model). However, as the outsourced service is merely part of a broader process, the FSI retains control over the rest of the process, including the customer data that is used as input into the service provider's AIDA model.
3. **AIDA-output as a service**, where the FSI purchases the output of an AIDA model from a service provider (eg, a credit score). In this scenario, the service provider has full control over the AIDA model, including how it is deployed, the data used to train the AIDA model, and even the customer data that is used as input into the AIDA model. In fact, the decision is heavily dependent on the service provider, and the FSI has minimal visibility into and control of the process.

To illustrate the difference across these three categories of externally developed AIDA, we rely on a simplified use case of a home loan application. The bank adopts a general-purpose AIDA model developed by a third-party service provider to calculate the interest rate for each loan application, which is determined by a range of factors such as the maximum loan amount, the value of the property, the interest rate environment, and the bank's risk appetite. Although the service provider had trained the base AIDA model, the bank has to retrain the model with its own data so that the model is applicable to the bank. We call this **AIDA-model as a service**.

Before it assesses any application, the bank wants to filter out applicants with poor credit history to reduce non-performing loans and to minimize administrative costs. It therefore subscribes to a service by a credit bureau that provides it with creditworthiness scores, which the credit bureau derives using an AIDA model that it developed itself and using data that it collects through other sources. The bank only receives the score for each applicant and has no visibility into the AIDA model and the underlying data used. We call this **AIDA-output as a service**.

For the applicants that meet the threshold creditworthiness score, the bank engages a real estate consultancy to conduct valuation of the property associated with each application, which is one of the input features in the bank's interest rate AIDA model. The bank shares some data of the applicant with the consultancy, which then uses an AIDA model that it developed itself to conduct the valuation. We call this **AIDA-processing as a service**.

Based on these three broad categories of externally developed AIDA, we developed the illustrative shared responsibility framework below. However, it is important to note that this framework is heavily dependent on the specific use case, as the delineation of responsibilities between the AIDA system owner and the third-party AIDA system developer is a function of the control each party has, which in turn depends on the exact application of the AIDA model.

	Internally developed AIDA	AIDA-model as a service	AIDA-processing as a service	AIDA-output as a service
Proactive disclosure to data subjects of use of AIDA	●	●	●	●
Provide clear and appropriate explanations on <u>what</u> data is used to make AIDA-driven decisions upon request by data subject	●	●	●	●
Provide clear and appropriate explanations on <u>how</u> data is used to make AIDA-driven decisions upon request by data subject	●	●	●	●
Provide clear documentation and appropriate explanations on the underlying AIDA model upon request by regulator, especially for material use cases	●	●	●	●

● FSI ● Service Provider

Diagram 5: Illustrative Shared Responsibility Framework for AIDA

We strongly encourage Veritas to build on this initial illustrative shared responsibility framework to develop a more comprehensive framework that can then form an important reference for AIDA regulations in the same way the cloud computing shared responsibility framework has become an important reference for outsourcing regulations.

3. Application of FEAT Transparency Principles Assessment Methodology

Google Pay India utilizes gradient-boosted decision trees (GBDT) for the fraud detection AIDA model, which is an inherently interpretable architecture. The graph below shows the relative importance of anonymized input features used in the AIDA model.

Global Explainability: Relative Important of Input Features for Google Pay India Fraud Detection AIDA Model

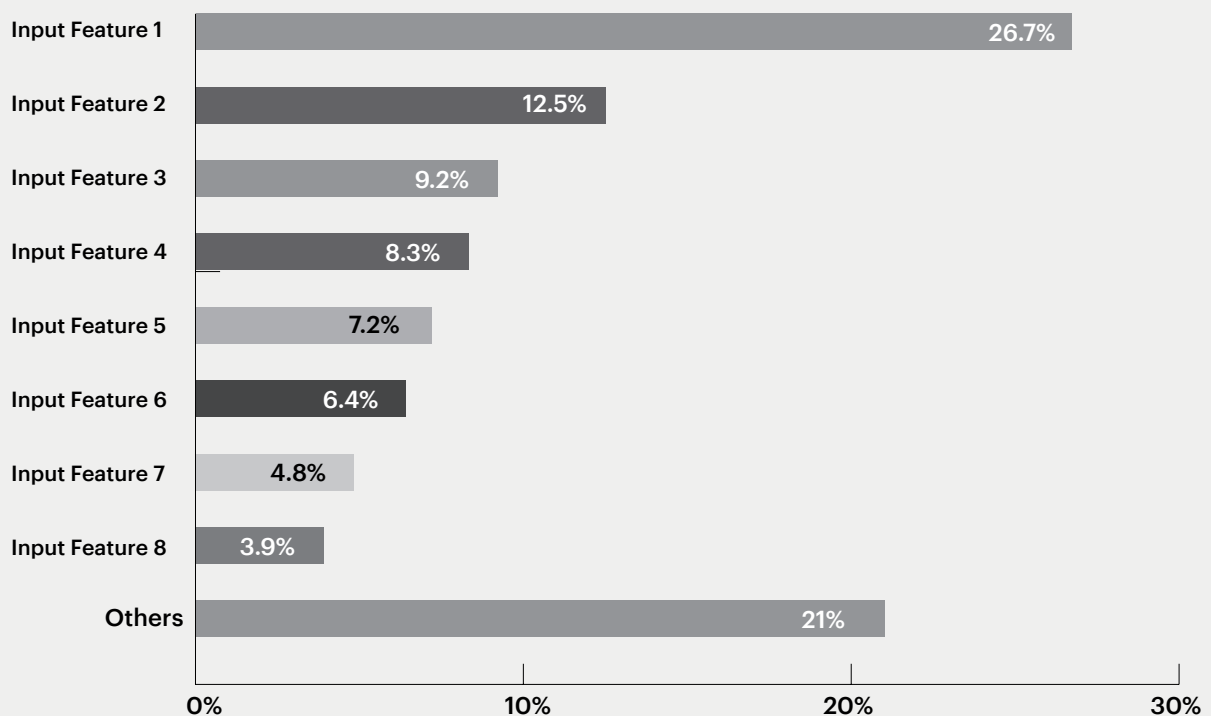


Diagram 6: Importance of Input Features for Google Pay India Fraud Detection AIDA Model

As noted above, Google believes that transparency is not an end in itself; it is a means to enable accountability, empower users, and build trust and confidence. As noted in the FEAT Transparency Principles Assessment Methodology, in designing transparency requirements, FSIs should consider what they are trying to achieve and how best to meet those goals in a given context. In other words, transparency requirements must reflect the context of a particular AIDA application. In this specific Google Pay India fraud detection use case, we demonstrate in practical terms how transparency requirements could, in fact, be harmful, a risk highlighted several times in the FEAT Transparency Principles Assessment Methodology.

Principle 12: To increase public confidence, use of AIDA is proactively disclosed to data subjects as part of general communication

Since 2017, [Google has made a shift from a mobile-first to an AI-first strategy](#).³ Today, AIDA is embedded in all our products and everything that we do. It is therefore not feasible for us to proactively and meaningfully disclose the use of AIDA to data subjects, especially at every instance (eg, each time a search is made on Google, each time a user sees a list of content recommendations on YouTube, each time an enterprise runs a workload on Google Cloud, each time a user makes a payment on Google Pay).

This challenge is not unique to Google. With the rapid advancements in technology, computing will similarly shift from a mobile-first to an AI-first world. Policymakers should adopt a forward-looking approach and consider if it would still be meaningful to proactively disclose the use of AIDA in an AI-first world.

Principle 13: Data subjects are provided, upon request, clear explanations on what data is used to make AIDA-driven decisions about the data subject and how the data affects the decision**Principle 14: Data subjects are provided, upon request, clear explanations on the consequences that AIDA-driven decisions may have on them**

Although one of Google's AI Principles is to "be accountable to people" (which includes providing appropriate opportunities for feedback, relevant explanations, and appeal), we need to balance this with the other AI Principles, such as to "be socially beneficial". In this fraud detection use case, while we recognize the importance of transparency – specifically external transparency, we are also very mindful of the risk of abuse by fraudsters to game the system if they were to understand how our AIDA model detects fraudulent transactions. Disclosure of this information could severely undermine our objective to make Google Pay a faster and more secure way to pay. The FEAT Transparency Principle Assessment Methodology highlighted this as a potential argument against transparency, with a caveat that one-off explanations may not meet the high bar in making this argument against transparency.

Once again, context is key. Google Pay India's data shows that the user accounts of fraudsters tend to be short-lived, as our AIDA model typically flags accounts associated with potentially fraudulent transactions after some time. To get around this problem, a single fraudster is therefore likely to have multiple user accounts, while each fraudster is likely to operate in a syndicate comprising multiple fraudsters. If we were to view transparency requests by each user account as one-off explanations, we would be divulging critical information about the inner workings of our AIDA model, which would enable criminal syndicates to game the system by accumulating knowledge over multiple fraudsters with multiple user accounts each.

4. Application of FEAT Fairness Principles Assessment Methodology

Principle 1: Individuals or groups of individuals are not systematically disadvantaged through AIDA-driven decisions, unless these decisions can be justified

Google Pay India voluntarily collects minimal personal data of our users out of respect for user privacy. The only data we collect that could be considered personal attributes are location, language preference, and device type (Android or iOS).

Given this limited data, we are unable to conduct comprehensive fairness assessments. Nonetheless, we analyzed AIDA-driven decisions against the Sender's location and found the geographical distribution of Senders whose transactions were approved was not statistically different from the geographical distribution of Senders whose transactions were denied. This may suggest that our users are not systematically disadvantaged through AIDA-driven decisions.

However, this is by no means conclusive evidence. We discuss in greater detail the implications of this challenge under Principle 2 below.

Discussion on Principle 2: Use of personal attributes as input factors for AIDA-driven decisions is justified

The principle that the use of personal attributes as input factors for AIDA-driven decisions should be justified suggests that the use of these personal attributes as input factors is likely to undermine fairness. While this may be true in general, our experience in applying the FEAT Fairness Principles Assessment Methodology to the Google Pay India use case illustrated how, in fact, not using these personal attributes could undermine fairness. We summarize below the challenges we faced with this principle of justifiability, some of which had been touched on briefly in the FEAT Fairness Principles Assessment Methodology.

- 1. Not using or collecting personal attributes creates a significant challenge in assessing fairness:** Google Pay India uses minimal personal attributes in the fraud detection AIDA model, be it in the actual operation or the training of the model. In fact, such data is not even collected by Google Pay India, as it is neither a requirement nor industry practice for payments service providers to do so, unlike other FSIs, and is also consistent with data minimization principles. This creates a significant challenge for Google Pay India in assessing the AIDA model against fairness metrics, as the requisite data for such assessment is almost non-existent. This challenge was highlighted in the FEAT Fairness Principles Assessment Methodology as "fairness by unawareness".
- 2. Not using personal attributes does not guarantee fairness:** One may argue that the aforementioned challenge is not an issue because if personal attributes are not used as input factors, the AIDA model cannot be biased against any specific group or individual. However, the underlying data that is used to train the AIDA model could be inherently biased (eg, perhaps there are more Google Pay India users of

one demographic segment than others), meaning the use of such data for training could reinforce systemic disadvantages. The FEAT Fairness Principles Assessment Methodology also highlighted this challenge as a potential justification for the use of personal attributes. FSI therefore should be mindful not to simply rely on an AIDA system that is blind to personal attributes, as that in itself does not guarantee fairness.

3. **Privacy laws could limit collection of personal data:** Even if FSIs are cognizant of the aforementioned challenges and therefore wish to collect personal data to facilitate fairness assessments, privacy laws could limit the collection of such personal data. If FSIs are required to seek consent from data subjects for the collection and use of personal data, FSIs may not have the requisite data for fairness assessments. This challenge of “consumer consent” was also highlighted in the FEAT Fairness Principles Assessment Methodology. Moreover, if only personal data for which consent has been obtained is used either for the training or operation of the AIDA model, it could reinforce systemic disadvantages because data subjects that withhold consent may be more likely to be subject to societal prejudices (and therefore fear that the disclosure of their personal data would be used against them).

Notwithstanding the aforementioned challenges, we utilized whatever data we had to assess the Google Pay India use case against fairness. Based on this limited assessment (as described under Principle 1), we did not detect any potential fairness issues in the AIDA system. In addition, our assessment of the materiality of the AIDA system concluded that the system’s risk of harm is fairly low because of the following reasons:

- **Low probability and severity of potential harm on affected users:** Less than 0.1% of total transactions are denied automatically by the AIDA model (though the absolute number is still large due to the total volume of transactions). More importantly, the severity of potential harm on users whose transactions have been denied is arguably very low, as a false positive presents primarily inconvenience to the affected users. This should be balanced against the benefits of processing literally billions of transactions in a faster and more secure way.
- **Available options for recourse:** As explained in Diagram 1, users whose transactions have been denied have a recourse option available to them by contacting a human reviewer to clarify that the transaction is legitimate. Once this is verified, the human reviewer will mark the user as safe, which would then overwrite the AIDA model and approve future transactions, albeit only for a limited period.
- **Human-in-the-loop throughout AIDA system:** When viewed holistically as an AIDA system (instead of just the AIDA model), automated decision-making is limited only to the denial of suspicious transactions, which has a low probability and severity of potential harm, as explained above. These automated decisions are then used as input factors for human decisions on the suspension of user accounts, which have higher severity of potential harm on users whose accounts have been suspended. This use case is an illustration of Google’s approach to leveraging AIDA to support and complement human-driven processes, especially those underlying services at scale.

The challenges we faced in assessing and ensuring fairness in this Google Pay India use case are not unique to AIDA; almost all of them apply to human-driven systems too. The proliferation of AIDA has however shone a light on these challenges, and presented an opportunity for us to address them.

However, until we can find a way to address these challenges, regulators may need to be realistic about how fairness objectives can be achieved, and avoid a one-size-fits-all approach. Google firmly believes in the importance of using AIDA in a manner that does not create or reinforce unfair bias. It is precisely because of how important this principle is that regulators should approach it carefully and thoughtfully, as unrealistic expectations of fairness could, in certain situations, harm precisely those whom we are trying to protect.

Principle 3: Data and models used for AIDA driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimize unintentional bias.

Principle 4: AIDA driven decisions are regularly reviewed so that models behave as designed and intended.

To ensure proactive monitoring of the AIDA model's deployment and decision-making process, Google Pay India has put in place the following mechanisms as part of a multi-tiered governance system:

- 1. Pre-deployment checks:** Before an AIDA model is deployed or any new features are introduced, we conduct extensive rigorous checks on the model and perform numerous dry runs.
- 2. Using human decisions to refine AIDA model:** As explained in Diagram 1, based on the AIDA model's assessment of the probability of a transaction being fraudulent, the AIDA model may flag the account of the transaction's Recipient for review by a human reviewer. The outcomes of these human reviews not only provide insight into the precision of the AIDA model, but also are fed back into the AIDA model to refine its decision-making process. It is important to note that the decisions taken by human reviewers take precedence over those taken by the AIDA model. For example, if the AIDA model denies a transaction and flags the Recipient's account for review, but the human reviewer determines that the Recipient has not violated Google Pay's policies, the Recipient's account will be marked "safe" such that the AIDA model's decision will be ignored for a certain number of days (and not more so that the Recipient does not get perpetual "immunity").
- 3. Minimizing unintentional human bias:** However, we recognize that human reviewers may not be perfect and could be prone to unintentional human bias. We therefore have a team of senior human reviewers to constantly evaluate the human reviewers by conducting audits of a random sample of decisions made by the human reviewers. Once again, the outcomes of these audits are fed back into the AIDA model to refine its decision-making process.
- 4. Ensuring quality assurance using key performance metrics:** We also have a separate quality assurance team that monitors a set of key performance metrics to ensure that the entire AIDA system, including the human-driven processes, performs as designed and intended. For instance, this team tracks the AIDA model's daily and monthly transaction denial rates to ensure consistency, as well as monitors the overall quality of the human review process based on key metrics from the aforementioned audit process. There is also an alert system that flags anomalies in these performance metrics (eg, sudden spikes in denied transactions), which would trigger further inspection by the quality assurance team.

5. Considerations and Recommendations

Based on our experience in applying the FEAT Principles Assessment Methodologies to the Google Pay India fraud detection use case, we have highlighted the following considerations as potential areas of improvement for Google Pay India:

1. As explained in Principles 10 and 11, there is a mechanism for users to submit appeals and request reviews in the event of accounts suspensions. However, through this exercise, we realized that the mechanism to do the same for denial of transactions is less user-friendly, though we note that the severity of potential harm is far lower. Nonetheless, Google Pay India may wish to consider improving the user interface.
2. While we note that it may not be feasible for Google Pay India to proactively and meaningfully disclose the use of AIDA to data subjects, it may be worth exploring the benefit of such disclosure, perhaps as part of Google Pay's terms of service.

In addition, we would like to make the following three broad policy recommendations for consideration by MAS and Veritas:

1. The FEAT Principles Assessment Methodologies are especially useful for FSIs who may be new to AIDA, but regulatory supervision of FEAT should be kept at a principles level as FSIs may have different ways of implementing the FEAT Principles. This is especially true for FSIs who have existing principles, governance and review processes and practices that may not be fully aligned with the exact FEAT Principles Assessment Methodologies, but are still able to adhere to the FEAT Principles.
2. Notwithstanding the above, the FEAT Principles should incorporate learnings from Veritas' application of the FEAT Principles Assessment Methodologies and be periodically reviewed. One potential learning that came out of this use case is the need to update the FEAT Transparency Principles, which focus exclusively on external transparency for data subjects, while the FEAT Transparency Principles Assessment Methodology highlights the importance of internal transparency and the need to consider external transparency for other stakeholders, such as regulators.
3. Relevant regulators may wish to study if existing regulations under their jurisdiction should be refined to support FSIs' implementation of the FEAT Principles. As highlighted in this use case, MAS' Outsourcing Guidelines may benefit from incorporating a shared responsibility framework for AIDA to address externally sourced AIDA models. The Personal Data Protection Commission may also wish to refine the Personal Data Protection Act or the corresponding Advisory Guidelines to address the tradeoffs between privacy and fairness.

Veritas Phase 3 – Transparency Use Case Whitepaper

Transparency Assessment for Predictive Underwriting



1. Background

This document is the output of Veritas Phase 3, developed with the foundation of the Transparency Assessment Methodology and use cases for banks and insurance applications from Veritas Phase 1 and 2 (the Phase 2 insurance use cases applications covered fairness, Ethics & Accountability). This use case whitepaper aims to illustrate the application of the transparency assessment methodology for the insurance sector. This document starts with a summary of the predictive underwriting use case (PUW) and the Transparency methodology, followed by the transparency assessment and explanations structured according to the AIDA lifecycle, and then concluded with key learnings and reflections on additional value transparency could bring to customers and insurers when deploying such solutions.

In Phase 3 of the Veritas project, Swiss Re (Swiss Re Asia Pte. Ltd.), leading the workstream, Great Eastern (Great Eastern Life Assurance Singapore) and TruEra (Truera, Inc.) have partnered to apply the Transparency Assessment Methodology and validate the outcomes produced in the use case document from a synthetic dataset as in [Phase 2 Fairness use case](#).⁴

The Monetary Authority of Singapore (MAS) has published a set of guiding principles on Fairness, Ethics, Accountability and Transparency in 2018, aiming to provide guidance to financial institutions on the responsible use of Artificial Intelligence and Data Analytics (AIDA) solutions. The FEAT Principles defined by MAS consists of a set of 14 principles around Fairness, Ethics, Accountability and Transparency, of which P12 to P14 illustrated below are connected to Transparency.

The following figure show the Principles of Transparency as defined in the Monetary Authority of [Singapore's FEAT Principles published in 2018](#)⁵:

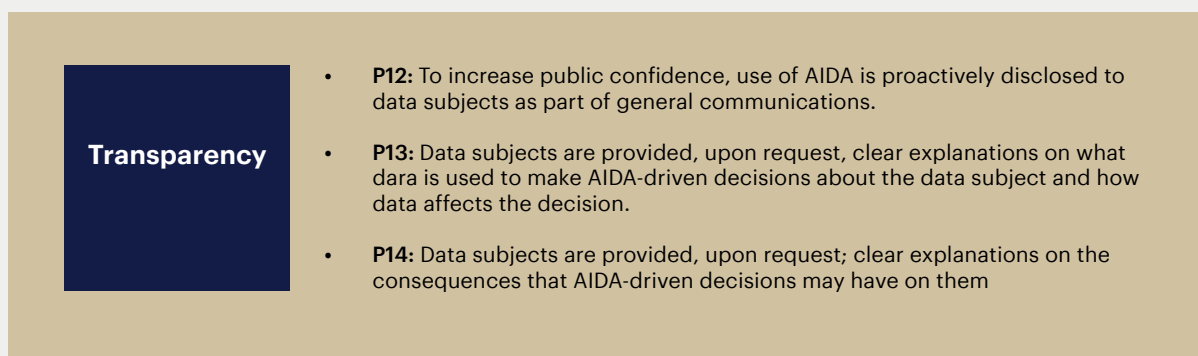


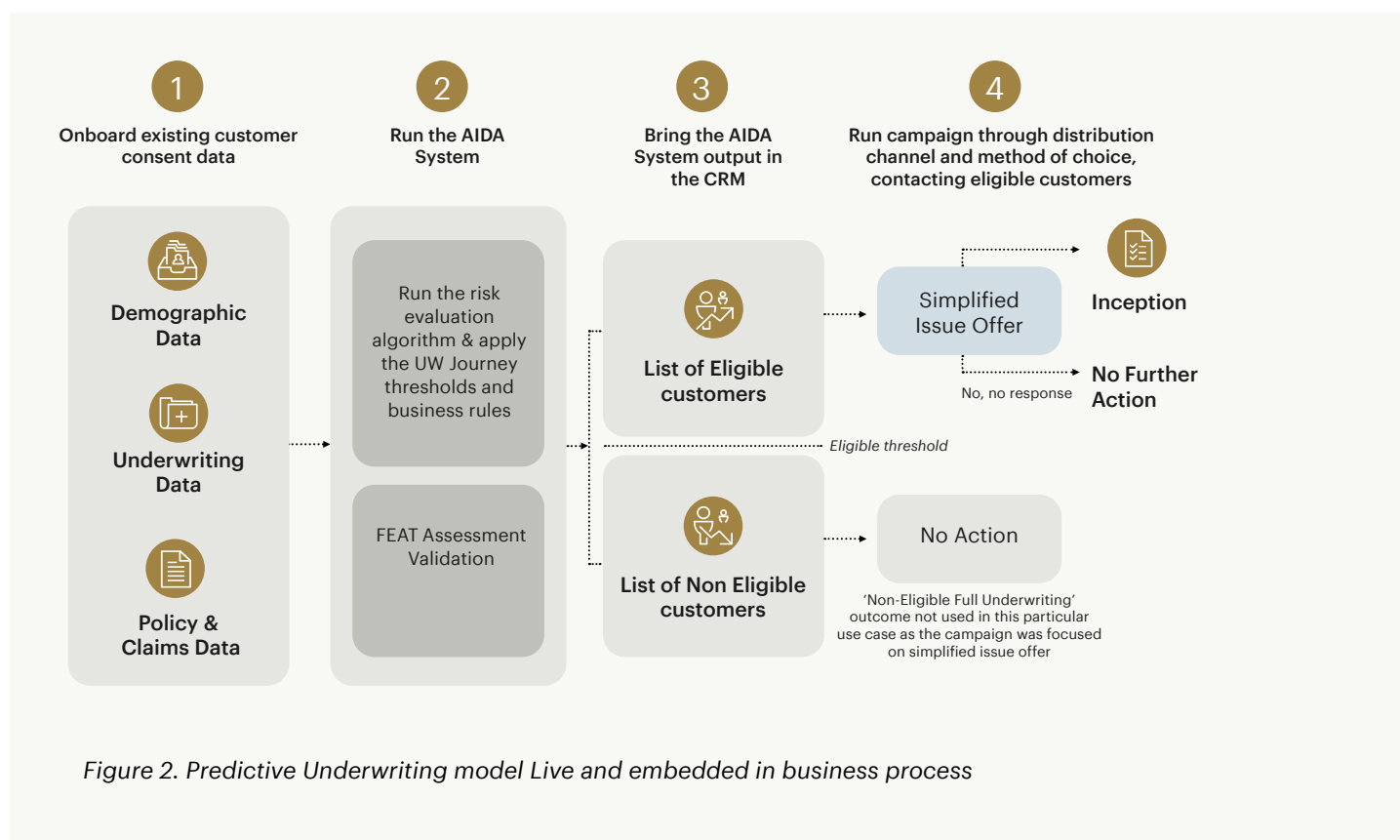
Figure 1. MAS FEAT Transparency Principles (2018)

2. Use case

Use Case Description

This whitepaper illustrates the application of the Transparency Assessment Methodology applying it to the same insurance use case selected for fairness assessment in the [Phase 2 Veritas Paper 4](#).⁶

This use case considered the scope of predictive underwriting model to cross sell life insurance products to existing customers in the Singapore market. Predictive Underwriting (PUW) helps life insurance companies segment and underwrite risks using data and domain expertise from underwriters. The system aims to predict if an insurer's customer is eligible for undergoing a simplified underwriting process to obtain a life insurance product based on associated risk levels. In our use case, the model is used for a cross-sell campaign - an active effort by the insurer to reach out to existing customers pre-identified as eligible by the model outcome (cross sell campaign), rather than a passive one where the insurer waits for existing customers to request life insurance before checking their eligibility. If the eligible customer accepts the simplified underwriting offer, they receive an expedited approval (simplified underwriting process with no price loading) and do not have to go through the full underwriting process. The ineligible customer may be able to receive the same insurance product if they contact the insurer and go through the full underwriting journey. For this use case, gender-specific thresholds were adopted with fairness, accuracy and precision considered - meeting both the firm's primary commercial objectives and keeping the fairness metrics within acceptable fairness threshold. Details of the gender-specific thresholds are documented in [Phase 2 Fairness Use Case Whitepaper](#).⁷



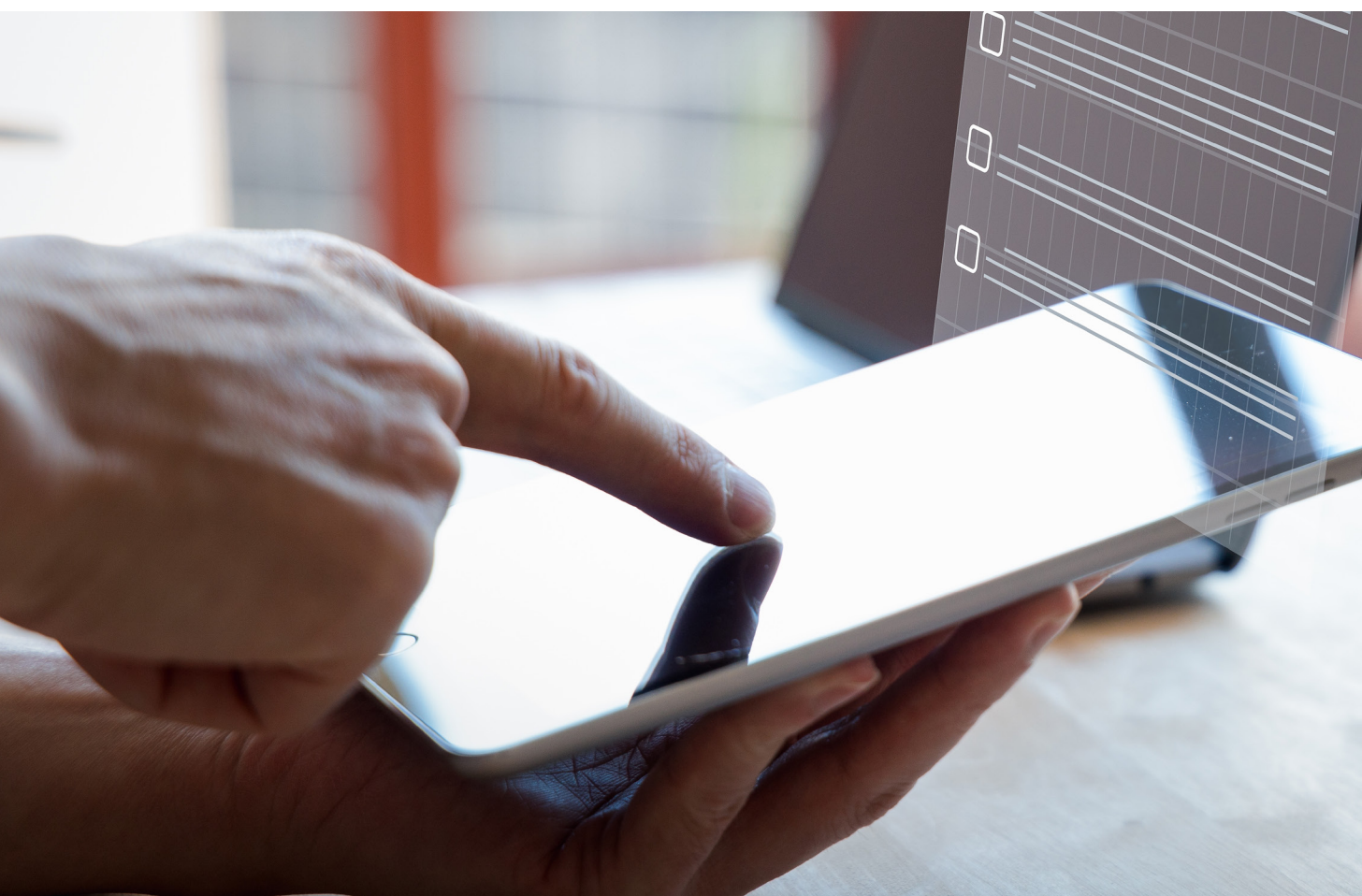
Insurer Description

The predictive underwriting use case and the application of Transparency Assessment Methodology have been conducted for a hypothetical insurer using a synthetically generated dataset to preserve data privacy. The intention of which is to give Financial Service Institutions (FSIs), and particularly insurers, a worked example of a populated transparency assessment document.

The company “Insurer A” is a hypothetical Singapore insurer that is conducting a transparency assessment on its predictive underwriting model.

Insurer A has defined transparency standards for its organisation, taking into account its organisational principles and values as well as all relevant regulation, and these standards have been approved through its internal governance processes. They are used consistently for AIDA systems across the organisation.

Note: It is up to the FSI to determine which standards are best for its organisation considering market practices, sociocultural views, organisational values, regulatory guidance and the legal requirements in its jurisdiction on related concepts that could impact transparency.



3. Overview of Transparency Assessment Methodology

This overview is based on Section 3 from the Phase 2 Transparency Assessment Methodology whitepaper. The transparency assessment consists of a set of questions mapped to the steps in a typical AIDA lifecycle, outlined in the diagram below.

As illustrated in the [Transparency Assessment Methodology Whitepaper⁸](#): The original FEAT document outlines the Principles to be used to guide FSI's work on Transparency (Principles 12, 13 and 14). FSI's should first translate these Principles into appropriate internal standards that are not specific to individual use cases but applicable to the FSI as a whole. Teams accountable and responsible for individual AIDA systems should then adopt and operationalise the appropriate standards during development, validation, deployment, and ongoing monitoring of their AIDA systems.

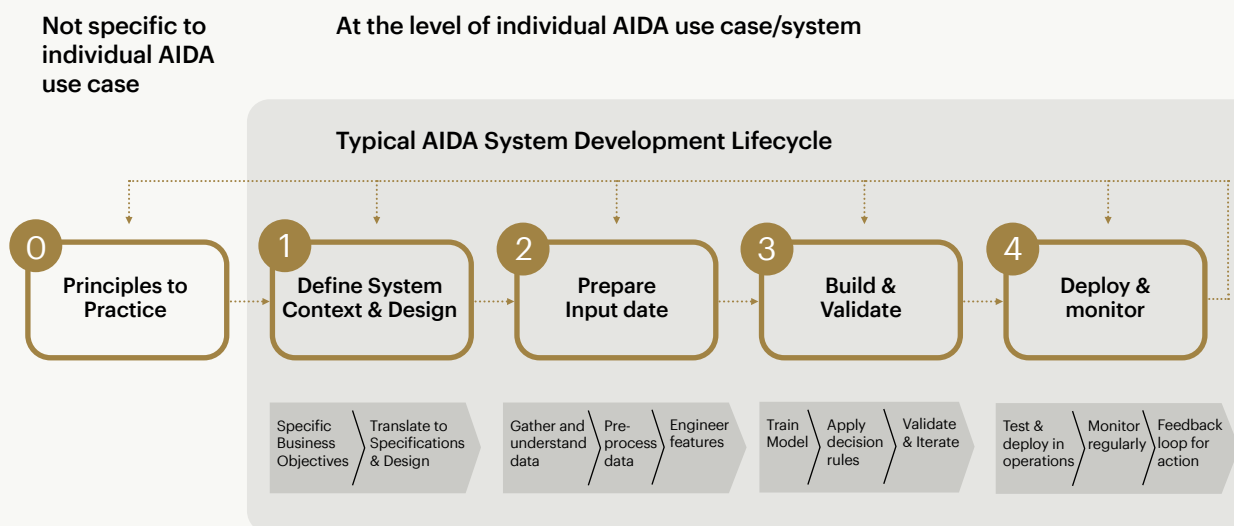


Figure 3. AIDA System development lifecycle

An AIDA system encompasses not just the AIDA model(s), but also other parts of the decision-making process that are required to get to the final AIDA-driven decision such as human contributions, other governance or rules that affect decisions, and actions taken based on the AIDA system's output. The steps involved in the typical AIDA lifecycle are illustrated in the above diagram, and the Methodology whitepaper has the questions applicable at each step.

4. Transparency Assessment and Explanations

This section demonstrates how the transparency assessment is conducted using the Transparency Assessment Methodology, defined in Veritas Phase 2, mapped to a typical AIDA System Development Lifecycle, and applying it to the specific use case of predictive underwriting for cross-sell in Singapore.

Step 0 – Internal Standards (Transparency)

(T1) Has the FSI defined the factors it will use to determine whether external (customer facing) transparency is essential for a particular AIDA use case?

Existing requirements of Insurer A around treating customers fairly and respecting the rights of individuals around their personal data already cover transparency. In addition, specific Responsible Advanced Analytics principles have been recently defined by Insurer A and communicated throughout the company. These include external and internal transparency guidelines and their applicability. Furthermore, the decision is also informed by the prevailing regulatory and legal disclosure requirements in the relevant market.

(T2) (Where an FSI has chosen to provide external transparency) At each stage of the FSI's customer lifecycle, has the FSI determined what proactive or reactive communication may be needed, and the standard templates/interfaces for the same?

Insurer A's guidelines emphasize the trade-offs that need to be made between disclosing confidential information and customers' rights. Insurer A identified four types of information that could be delivered depending on the use case, context and jurisdiction: proactive, reactive, informational, and action-oriented. For definition of proactive and reactive communication, please refer to Veritas Phase 2 Transparency Assessment Methodology paper. Any communication must adhere to applicable laws and regulations while respecting principle of transparency.

(T3) Has the FSI defined the factors it will use to determine the extent of, and audience for internal transparency for individual AIDA use cases?

Internal transparency is aimed at a large audience, with focus on business users of the AIDA system, the business model owners, the AIDA system developers, and the model auditors.

According to Insurer A's analytics governance, the business users are offered an intuitive understanding of the system's outputs to gain confidence on its performance. The business model owner, the person who is ultimately accountable for the analytics model's outcomes as defined in Insurer A's analytics governance and is selected per use case – typically the key business expert with sufficient decision-making and resources-allocating power, is provided with model performance monitoring dashboards to ensure the AIDA systems run as expected. The AIDA developers and auditors use an extensive array of transparency tools to maintain, debug and improve the system, especially if it is handed over to a new team.

(T4) Has the FSI defined an acceptable set of AIDA ML explanation method(s) for use within the FSI?

Insurer A is developing transparency requirements and guidance on approaches, including benefits and drawbacks of specific methods. An internal process to review advanced analytics models from both the modelling and coding perspectives ensures adherence to best practices, including that proper explanation methods are chosen for each use case, for instance depending on the type of data, e.g., structured numerical or categorical, or unstructured such as free text. The internal process and set of best practices are periodically reviewed and updated, to incorporate findings from use cases and stay current with progress in research and industry standards.

(T5) Has the FSI set minimum accuracy standards for such explanation methods?

Assessments are required to validate the robustness of the explanations generated by various approaches, however specific metrics and corresponding thresholds need to be defined on a case-by-case basis. The assessments check whether the explanations change significantly when applying small changes to the training data or to the model parameters.

Step 1 – Define System Context and Design

(T6) Has the AIDA use case team determined whether there is a need for external (customer facing) transparency? Apply standards from T1 to help answer the question.

Yes. For this use case, external transparency is required as the activity relates to proactive cross-selling efforts to address potential customer needs via the preferred communication channel of the customer. Use of AIDA would be proactively disclosed.

Furthermore, transparency will be provided also to those customers who have not received a cross-selling offer and are requesting to buy. The insurance company broadly shares with the customer what data is used to analyse her/his patterns and provide an avenue for the customer to provide more accurate or updated information. There are very few cases in our scenario where this could happen as the cross-sell offer is limited only to eligible existing customers. One example is when a family member is eligible and willing to buy a similar product for the rest of the family members (and those are as well existing insurer customers). In case other family members are not eligible, they can still purchase similar products going through the full underwriting journey, with the insurer providing key data information (for example, via the Customer Relationship Management tool or via the Point-of-Sales tool embedded in the automated Underwriting engine) to the agent or directly to the customer on key data/reasons why they are not eligible in accordance with the transparency principle. This engagement with the customer could also be an opportunity for the insurer to verify the veracity of the data used by the model and provide an opportunity to the customer to update this data in case of change or mistake in a truthful manner (leveraging behavioural economics techniques for example in order to not provide an incentive to customers to lie or non-disclose on requested data that could harm future claim presentations.). The provision of this information/data must also be balanced with the need to protect the trade secrecy or know-how of the AIDA system and from potential adversarial usage in accordance with all applicable laws and regulations.

(T7) If yes, has the team identified the proactive and reactive communication needed at each stage of the customer lifecycle, and the form of such customer facing communication? Apply standards from T2 to help answer the question.

Yes. See answer above.

(T8) Has the team determined the level of internal transparency needed, and the audiences for the same? Apply standards from T3 to help answer the question.

Yes. Internal transparency is aimed at both business end-users, distributors (e.g., agents/brokers/service team), and model developers. Dashboards, featuring the relevant information specific to the audience groups within the business context, are provided to offer insights on global model behaviour and explanations of individual predictions.

(T9) Has the team selected a suitable explanation method for this specific use case from the approved list in T4?

Yes. Model-agnostic explanation methods have been selected for this use case to enable potential comparison across models. The selected explanation methods are becoming industry best practices: Partial Dependence Plots, Shapley values and counterfactual examples.

- Dependence plots illustrate the overall effect of a single feature on the predictions made by the model.
- Shapley values quantifies the marginal contribution (i.e. feature importance) that each feature brings to the prediction.
- Counterfactual examples provide the minimum change in features' values required to flip the model's decision (for binary classification, extensions exist for multi-class scenarios as well).

Note: It is up to the FSI to determine which standards and methods are best for its organisation considering market practices, sociocultural views, organisational values, regulatory guidance and the legal requirements in its jurisdiction on related concepts that could impact transparency.

(T10) Has the team ascertained that the chosen explanation method/implementation meets the minimum accuracy requirement for this specific use case (based on T5)?

An independent peer review validated the choice of transparency approaches and their sound application to the AIDA system for the predictive underwriting use case. Several robustness and soundness checks were performed to ascertain the validity of the explanations, e.g., assessing stability versus changes in the training dataset, and in-depth reviews by subject matter experts.

Step 2 – Prepare Input Data

Data Sources Description

To showcase how the assessment process can be followed in practice, a real insurance model development dataset was used to produce synthetic data – data that is close enough to real data to meaningfully illustrate the application of the methodology, but different enough to preserve privacy. The synthetic data generation was conducted using generative network methods including generative adversarial networks (GANs) and variational auto-encoders (VAEs), optimising to meet KPIs for similarity, utility, and privacy.

This synthetic dataset has been used throughout the use case described below. The synthetic dataset Insurer A used for the predictive underwriting model development consists of more than 20,000 records /observation and 21 attributes. The synthetic dataset does not include personal data (i.e., information that would allow individuals to be identified, such as address and name). As per common practices, a manual risk labelling process was used to label records as 1 and 0 representing an eligible and non-eligible risk case respectively. The ratio of eligible to non-eligible risk cases in the dataset is approximately 5:1.

In the context of this hypothetical use case, Insurer A had secured a consent from existing customers to use relevant data for the purpose of cross selling before including in the model development dataset, in accordance with Singapore laws and regulations.

Step 3 – Build and Validate AIDA System

Components of the AIDA system:

The AIDA system, for our illustrative use case, consists of one pre-processing business rule, one logistic regression model and two post-processing business rules.

1. Model approach: a logistic regression binary classification model was built to predict the eligible/ineligible risk labels from the training data with a confidence score. We then fine-tuned the model and underwriters decided on a threshold to make the final classification decision (simplified offer or full underwriting) based on the underwriting philosophy and the financial risk willing to onboard. The threshold(s) depended on model performance metrics on training and testing data and on the risk appetite of both reinsurance and insurance parties.
2. Business rules for predictive model: Certain business rules, agreed with underwriters and actuaries, were applied and where appropriate overruled the model's predictions to further control the risk and business context. For example, depending on the product, there are age or claims restricted offers and these individuals would need to go through the full underwriting process.

3. Human manual overrides: In some cases, humans override the decisions of AIDA systems, such as when the model's outcomes are particularly uncertain. The AIDA system described in this use case includes no such manual overrides. Instead, strict business rules are used to exclude cases that would typically be sent for manual review (e.g., high-cost claims). In this simplified example, only three business rules are used, but a real predictive underwriting model would typically have many more.

(T11) Have internal transparency dashboards/reports been implemented in line with the requirements agreed in T8-T10?

Yes, global and local SHAP plots and dependence plots have been generated, as outlined below. These reports and graphs aim to describe the marginal contribution (i.e. feature importance) that each feature brings to the prediction outcome. The explanation methods chosen by Insurer A are model-agnostic to facilitate comparison and the establishment of standards that can be applied to other models.

A simplified and contextualized version is provided to customer-front facing staffs (e.g. agents / brokers) based on their roles and action required during underwriting process, which is integrated into CRM/ automated underwriting engine.

Global interpretability

A global SHAP plot summarizes the effect of all the features on the model output and provides two insights:

1. The ranking of the features in terms of their impact on the model output and
2. The direction of the impact in relation to the feature values.

Each point on the graph represents a case, and the features are ranked by the sum of the absolute SHAP values for all samples. SHAP values represent the marginal contribution that each feature brings to the model output so the larger the SHAP value (absolute), the greater the impact the feature has on the model output. As seen in Figure 4, the SHAP values are also colour coded to show the direction of the impact – red represents larger values of the feature while blue represents smaller values of the feature. Categorical features (e.g. Feature 15) are not coloured/grey in colour as these features are non-numeric (e.g. category A, B, C) and cannot be represented meaningfully with the colour spectrum.

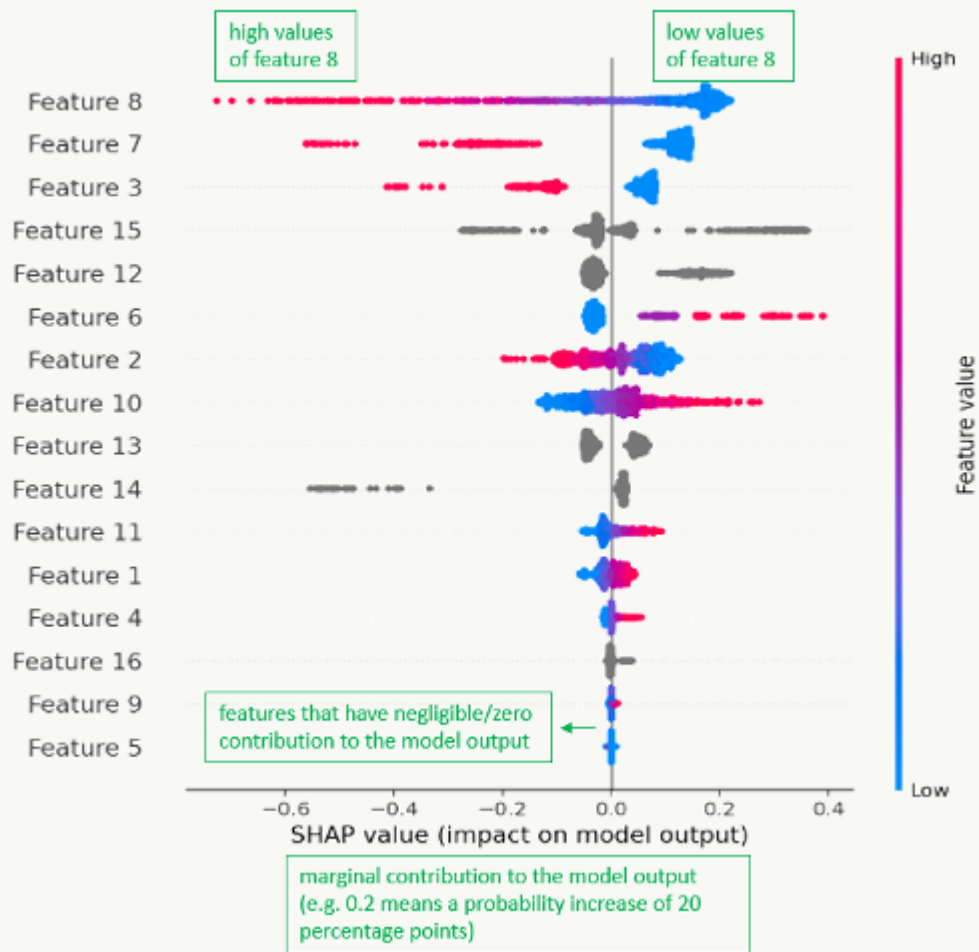


Figure 4: Global model interpretability with Shapley values

(Note: For this illustrative example – sense check of output was completed and is aligned with the expectation of the subject-matter-experts).

Example

Feature	Importance Ranking	Impact on Model Output
Feature 8	1	The higher its value, the less likely the customer will be eligible for the product
Feature 6	6	The higher its value, the more likely the customer will be eligible for the product

Table 1: Global interpretability example

Local interpretability:

Unlike global plots which show the overall feature importance, local SHAP plots show the feature importance for a single observation.

We show the local SHAP plots for two individuals below – the first individual was classified as eligible by the model while the second one was classified as ineligible. The plot shows the marginal contribution of each feature to the model output, which is the probability of the individual being eligible for the product. The baseline value, as seen at the bottom of the plot, is the expected probability of an individual being eligible for the product.

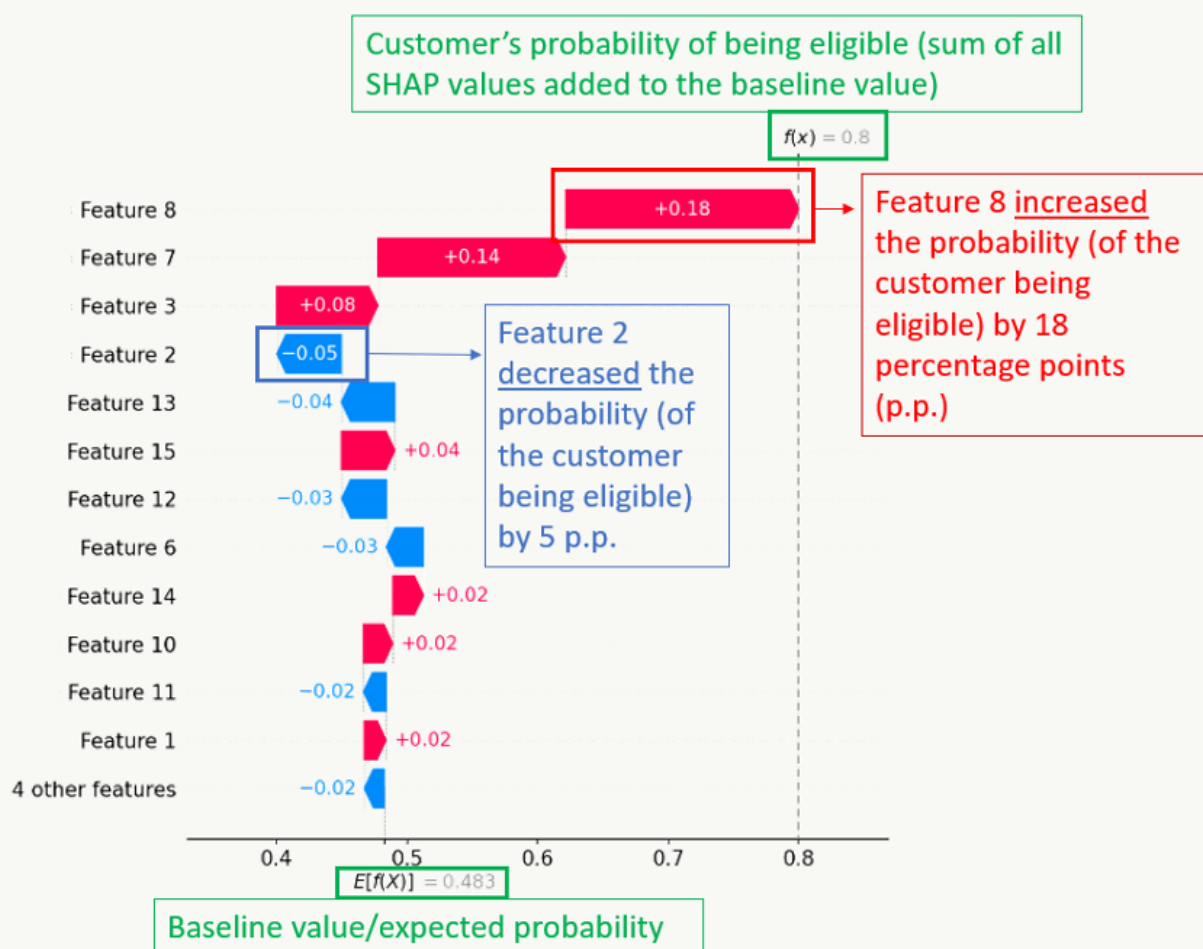


Figure 5: Local model interpretability with Shapley values (Individual 1, Eligible)

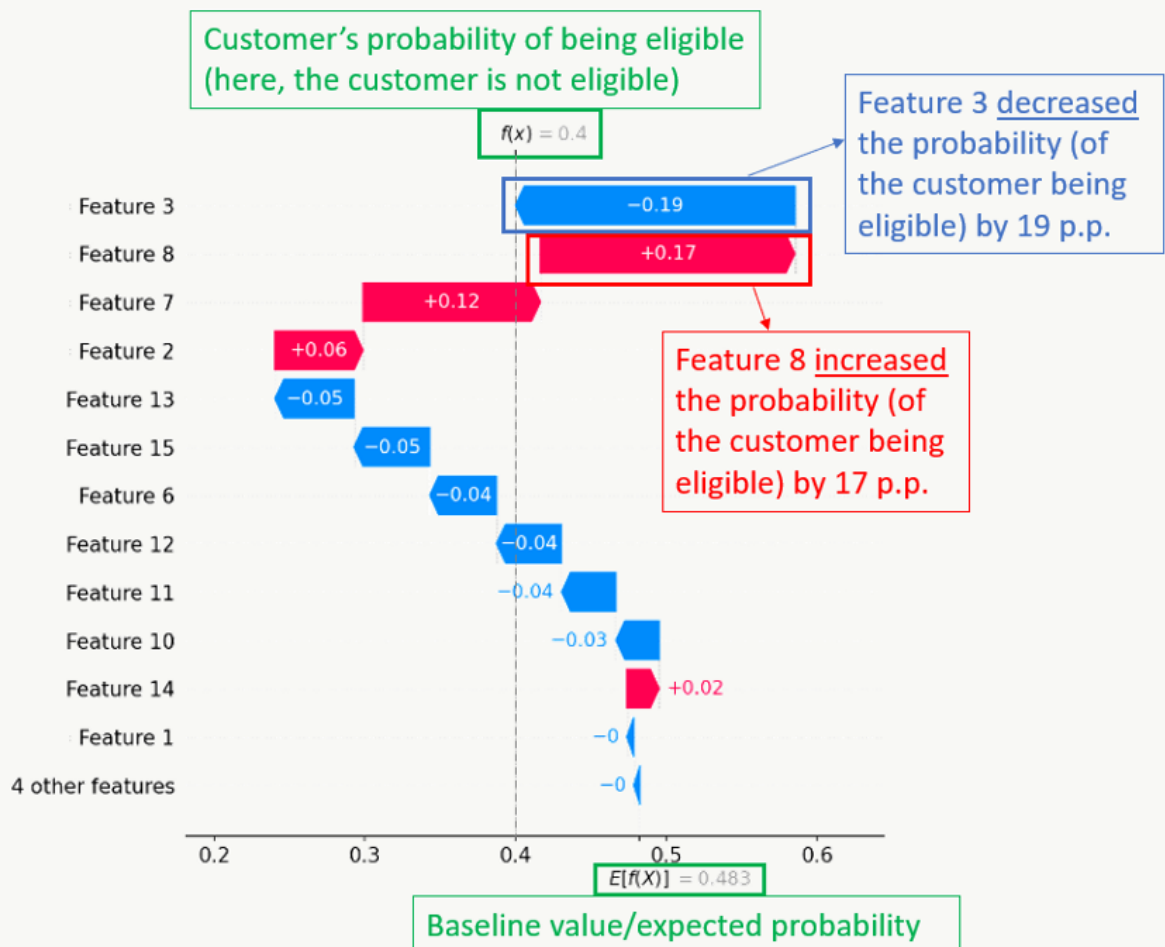


Figure 6: Local model interpretability with Shapley values (Individual 2, Ineligible)

Dependence plots:

Dependence plots show the overall effect of a single feature on the predictions made by the model. They are a simple way to make machine learning models easy to understand. We will be looking at two types of dependence plots.

SHAP dependence plots show how SHAP values (of a feature) change as values of the feature change. SHAP values represent the marginal contribution of each feature to the model prediction. Partial dependence plots are similar in that they show how the model prediction changes as values of the feature change.

Take Feature 8 as an example - Both the SHAP and partial dependence plots for Feature 8 show that the probability of being eligible decreases as the value of Feature 8 increases.

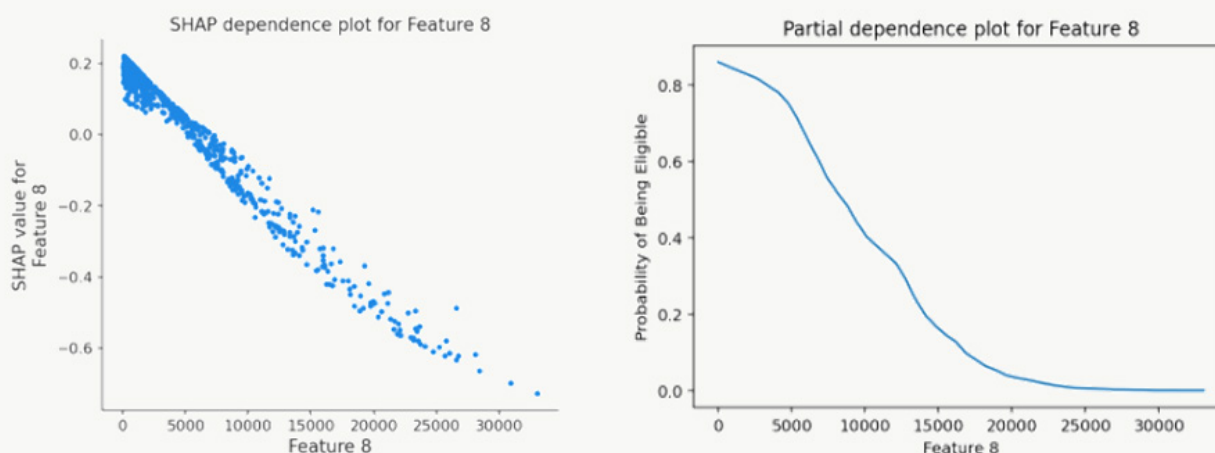


Figure 7: SHAP dependence plots and Partial dependence plots

(T12) Have relevant first and second line control teams – including model validation where relevant – reviewed and approved these outputs (e.g., local and global explanations, conceptual soundness)?

Yes. Relevant first and second line control teams of Insurer A were involved to validate the robustness of the transparency assessment outputs and its compliance status.

Insurer A has implemented the three lines of control model as part of their Analytics Governance and Risk Management standards. As Insurer A's first line of control, an independent peer review validated the choice of transparency approaches and their sound application to the AIDA system. Several robustness and soundness checks were performed to ascertain the validity of the explanations (e.g., assessing stability versus changes in the training dataset), in combination with in-depth reviews by subject matter experts. As Insurer A's second line of control, the Risk Management and Compliance functions provided risk control and advisory to ensure the model is compliant to Insurer A's data framework and internal policy.

(T13) Where the explanations have not met first/second line expectations, have appropriate mitigation actions been taken (e.g., switching to a simpler model despite a reduction in predictive accuracy, dropping difficult to explain features, introducing more human oversight)?

As an outcome of Insurer A's analytics governance evaluation process, mitigation actions have been taken to resolve contradictory explanations. These included revisiting the data pre-processing pipelines, the choice of input featured, modelling approach and model selection (hyperparameter tuning).

(T14) In line with the external transparency requirements agreed in T6-T7, has appropriate system functionality been developed and tested as part of the AIDA system's implementation plan?

Not yet applicable – to be considered in the future.

(T15) Have operational processes such as customer service and complaint handling been modified appropriately to incorporate AIDA customer transparency? Have relevant staff been provided appropriate training to address customer queries?

Not yet applicable. Insurer A will update their processes, personalized dashboard and train their agents to reply to customers questions on data and models used while interacting during the campaign.

(T16) Have customer/website Terms and Conditions been appropriately updated?

Not yet applicable. Terms & conditions of insurer A will be updated to cover the consent of data usage by customers for this campaign.

(T17) Does the AIDA system implementation support the agreed internal and external explanations even after go-live (i.e., not just as a one-off before approval but throughout the lifetime of the AIDA system)?

Not yet applicable – to be considered in the future. Insurer A's internal Responsible Advanced Analytics principles cover transparency monitoring requirements for implemented and deployed AIDA systems, and will be considered for the predictive underwriting system upon its implementation.

Step 4 – Monitor AIDA System

Post-launch monitoring and subsequent documentation is in place on model performance and usage. Context-specific evaluation shall be introduced depending on the business use case.

5. Reflections

Additional insight on model and data quality

The transparency assessment involves evaluating the model and data beyond the typical model performance metrics, which may not trigger any concerns on the suitability of the model. Specifically, this exercise enabled us to uncover and resolve data quality issues which were leading to inexplicable model behaviour, therefore further enhancing our confidence in the overall performance on the AIDA solutions.

Starts the assessment in early stage

Considerations around transparency and responsible AI are connected and intertwined with all stages of the AIDA system development lifecycle. The transparency assessment therefore cannot be deemed as a one-off evaluation to be conducted after the model development. For each stage in the development lifecycle, the relevant standards and transparency questions should be reviewed, followed by the respective model or data improvement needed. This practice will enable to identify and resolve potential modelling or data issues at an early stage, and avoid much more costly corrective measures at a later phase.

Building digital trust

Digital technologies may bolster the trust quotient in insurance. Trust between insurers and customers is not always easily realised - customers are worried about opacity, paying a higher price, policy or claims being denied, or being penalised for submitting claims. Ultimately, insurance is a promise to compensate policyholders for losses arising from specific eventualities. (More details on [May2022 SRI Decoding digital trust](#)⁹ – an insurance perspective)

Providing specific and potentially actionable feedback on decisions as outlined in the Transparency Assessment Methodology paper (2.2), must be: 1. It must be consistent with human intuition. 2. Its level of complexity must match the expertise level of the humans affected. 3. It must use domain specific (i.e., specific to the use case) vocabulary.

For example, Swiss Re has applied transparency principles to processing personal data in its [Data Protection and Privacy framework](#)¹⁰. We recognize that data subjects may have rights with regard to our processing of their data. While the nature and extent of these rights will differ from location to location, we have processes in place allowing for the timely response of any valid request, such as access, rectification, portability, object, prevent marketing, restrict processing and erase of the data, in any case where we use customers' data to make decisions solely by automated means (including using data to build a profile about data subjects), we will inform the customers that we are doing this and make sure that they are able to contest any such decision. Any new profiling activity or automated decision-making activity we carry out is subject to a robust assessment aimed at mitigating risks to data subjects. This assessment is carried out before the processing commences.

Testing and managing transparency in practice with the end-customer / agents a key for success

Care must be taken with end users that the individuals building the models and explanations don't presume understanding of customer knowledge, options for redress and potential for manipulation. Here it will be key to engage with customers to best understand the need based on the context, drivers and understanding of target customers for the case in hand. This requires a more customer centric development and continuous feedback approach in order to minimise potential 'backfire' effects where the well-meaning transparency does more harm than good. We suggest that rigorous behaviourally informed A/B testing become commonplace in the development of these systems. This requires greater collaboration between model developers, behavioural scientists and customers experience professionals in designing effective explanations that bring both value to the customer and ensure accurate disclosure.

Balancing customer transparency with non-disclosure risk, confidential information & adversarial usage protection

As covered in T6, bringing more transparency on data that drives an AIDA underwriting decision, should not come at the cost of providing incentive to a customer or its agent for non-disclosure, i.e., not providing information about existing health conditions or untruthful data in order to obtain a positive outcome from the AIDA system. However, if the insurer discovers that the customer failed to declare a pre-existing health condition, they can reject the health insurance claim and even void the entire plan.

Therefore, it's crucial when bringing more transparency on data usage and AIDA system decisions that the customer/agent is incentivised to be truthful at the time of underwriting. Recently, many insurers and reinsurers have developed strong methods supporting disclosure such as behaviour economic techniques, providing personalised protection, limiting price loading and rejection that could be tested and implemented according to each scenario. This would further strengthen the case for testing and using behavioural informed insights as well as monitoring of systems to minimise the risk of non-disclosure, more details on [May2022 SRI Decoding digital trust](#)¹¹ – an insurance perspective – Section reliability Ease of Use.

Providing more transparency on AIDA decision making processing assists in compliance with all applicable laws and regulations, balanced with the need to protect the confidential information of the AIDA system owner(s) and the potential for adversarial usage. We are still in the early days of AIDA adoption at scale and the area of protection versus adversarial usage and other logarithmic risks is very much developing worldwide. The adoption of best practices, tools and standards by the industry are yet to be defined and remain an area that will help shape the application of the transparency principles and its assessment by the financial service industry for the benefits of the industry and the end-customers.

References

1. <https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf>
2. <https://modelcards.withgoogle.com/about>
3. <https://blog.google/technology/ai/making-ai-work-for-everyone/>
4. <https://www.mas.gov.sg/news/media-releases/2022/mas-led-industry-consortium-publishes-assessment-methodologies-for-responsible-use-of-ai-by-financial-institutions>
5. <https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/FEAT>
6. <https://www.mas.gov.sg/news/media-releases/2022/mas-led-industry-consortium-publishes-assessment-methodologies-for-responsible-use-of-ai-by-financial-institutions>
7. <https://www.mas.gov.sg/news/media-releases/2022/mas-led-industry-consortium-publishes-assessment-methodologies-for-responsible-use-of-ai-by-financial-institutions>
8. <https://www.mas.gov.sg/-/media/MAS-Media-Library/news/media-releases/2022/Veritas-Documents-3C---FEAT-Transparency-Principles-Assessment-Methodology.pdf>
9. <https://www.swissre.com/dam/jcr:f8f4c8b1-73b4-45d6-830b-b582df02488c/may2022-sri-expertise-publication-decoding-digital-trust-en.pdf>
10. <https://www.swissre.com/about-us/data-protection-brochure/our-role-as-data-controller.html>
11. <https://www.swissre.com/dam/jcr:f8f4c8b1-73b4-45d6-830b-b582df02488c/may2022-sri-expertise-publication-decoding-digital-trust-en.pdf>

Acknowledgements:



Chester Chua

Project Lead and AI Subject Matter Expert

Addie Cooke

AI Subject Matter Expert

Matthew Ichinose

AI Subject Matter Expert

Will Carter

AI Subject Matter Expert

Arati Deo

Payments Subject Matter Expert

Monu Agrawal

Payments Subject Matter Expert

Ramanan Balakrishnan

Payments Subject Matter Expert

Niket Choudhary

Payments Subject Matter Expert

Nirupama Jayasimha

Payments Subject Matter Expert



Yannick Even

Lead Subject Matter Expert

Louise Lau

Project Coordination

Yvonne Tan

Data Scientist

Luca Baldassarre

Lead Data Scientist

Farooque Ahmed

L&H Underwriting Expert

Jasmine Ong

Lead Risk Management

Tomas Abelovsky

Legal Counsel

Mackenzie Clavin

Behaviour Economist Expert



Shameek Kundu

Subject Matter Expert

About Accenture

Accenture is a global professional services company with leading capabilities in digital, cloud and security. Combining unmatched experience and specialized skills across more than 40 industries, we offer Strategy and Consulting, Technology and Operations services and Accenture Song — all powered by the world's largest network of Advanced Technology and Intelligent Operations centers. Our 721,000 people deliver on the promise of technology and human ingenuity every day, serving clients in more than 120 countries. We embrace the power of change to create value and shared success for our clients, people, shareholders, partners and communities. Visit us at www.accenture.com.

About Google

Google's mission is to organize the world's information and make it universally accessible and useful. Through products and platforms like Search, Maps, Gmail, Google Cloud, Google Pay, Chrome and YouTube, Google plays a meaningful role in the daily lives of billions of people and has become one of the most widely-known companies in the world. Google is a subsidiary of Alphabet Inc.

About SwissRe

The Swiss Re Group, founded in 1863 and Headquartered in Zurich, Switzerland, is one of the world's leading providers of reinsurance, insurance and other forms of insurance-based risk transfer. Through our work with clients and partners we fulfil our vision of helping to make the world more resilient. We have a distinct mission: Together, we apply deep knowledge, intelligent data and capital strength to anticipate and manage risk. That's how Swiss Re powers progress for our clients, helping the world rebuild, renew and move forward.

Swiss Re's internal governance recognises that while technology offers many business opportunities, it also creates new risks. Swiss Re has developed a Digital Governance Framework (DGF) that aims to balance the often competing needs for fast business innovation and effective risk management. Designed to be comprehensive, risk-based and user-friendly, it makes requirements transparent and positions governance as a fundamental pillar of digitalisation. Our team of experts across the world will continue to further develop our DGF and digital responsibility frameworks to align with the development and adoption of AIDA systems, technical advancement, and regulatory changes in order to create value for individuals, society and the industry.

Disclaimer

This content is provided for general information purposes and is not intended to be used in place of consultation with our professional advisors. This document may refer to marks owned by third parties. All such third-party marks are the property of their respective owners. No sponsorship, endorsement or approval of this content by the owners of such marks is intended, expressed or implied.