The responsible AI framework

Strategy

Design

Development

Operating AI

Strategy

1. Aligning with your strategic goals

2. Don’t expect magic

3. Clear about your partners

4. Opening up to scrutiny

5. Demonstrating regulatory compliance

6. Organisational structure

1. Aligning with your strategic goals

It’s vital to align AI innovation with core strategic objectives and performance indicators, rather than allowing a scattered series of initiatives to operate in isolation. In our experience, a lot of organisations have set various pilots in train. What most aren’t doing is taking a fundamental look at how AI could disrupt their particular business and then determining the threats and opportunities this presents.

Design

1. Opening up the black box

AI applications can communicate with customers and make important business decisions. But a lot of this is carried out within a black box, with the lack of transparency creating inherent reputational and financial risks. It's important to ensure that the software is designed in a way that is as transparent and auditable as possible.

Proper governance and protection include the ability to monitor component systems. It would also include the ability to quickly detect, correct and, if not, shut down 'rogue' components without having to take down whole platforms. Related priorities include identifying dependencies and being able to make modifications with minimal disruption if regulations or some other aspect of the operating environment changes.

2. Creating a compelling user experience

Many AI applications deploy highly subjective user experience performance metrics akin to IQ, personality, and predictability. Even though the bulk of development may focus on the analytics, the success of the product will be determined by an emotional response. This subjectivity means that frequent feedback is required between product owners and developers to make sure evolving expectations and functionality are properly managed. Often it makes sense to bring in specialist user interface vendors or use your in-house digital team alongside the core analytics team.

AI may excel and often surpass humans at particular tasks or in certain subject domains, but is generally incapable of extending these skills or knowledge to other problems. This is not obvious to people who have to interact with AI, especially for the first time, and can cause frustration and confusion.

Branding and persona development ('functionality framing') are therefore key design considerations. Get it right and very basic software can appear human. Get it wrong and users will give up.

Some of the analysis performed by AI will inevitably be probabilistic based on incomplete information. It's therefore important that you recognise the limitations and explain this to customers. Examples might include how you present recommendations on investments from robo-advisors.

It's vital to align AI innovation with core strategic objectives and performance indicators, rather than allowing a scattered series of initiatives to operate in isolation.

3. Embedding the control framework

The most effective controls are built during the design and implementation phase, enabling you to catch issues before they become a problem and also identify opportunities for improvement.

An important question is who designs and monitors the controls? Both the breadth of application and the need to monitor outcomes requires engagement from across the organisation. Control design requires significant input from business domain experts. Specialist safety engineering input is likely to be required for physical applications.

A key part of implementation is breaking the controls down into layers ('hierarchical approach').

At a minimum, there would be a hard control layer setting out 'red lines' and what to do if they're breached. Examples might include a maximum transaction value for a financial market trading algorithm. In more complex applications such as conversational agents, you could introduce a 'behaviour inhibitor' that overrides the core algorithm when there is a risk of errors such as regulatory violation or inappropriate language.

These core controls can be augmented by 'challenger models', which are used as a baseline to monitor the fitness and accuracy of the AI techniques or look for unwanted bias or deviations as the models learn from new data. Moreover, this approach can be integrated with continuous development to improve existing models or identify superior models for system upgrades.

Development

1. Rethinking programme management

2. Managing data dependency

3. Taking the time to test and train

4. Setting confidence thresholds

1. Rethinking programme management

Applying conventional planning, design and building to such data-dependent developments is destined to fail. Innovating and proving the concept through iterative development is needed to handle the complexity of the problems encountered and requires a high level of engagement from the product owners.

Operating AI

1. Curbing unintentional bias

As more information becomes available and your model matures, it's important to guard against unintended bias against particular groups. Transparency is vital to be able to spot these biases. For systems that learn through customer interactions, periodic functional monitoring, perhaps based on a set of standardised interactions, is recommended to catch any adverse 'training drift'.

2. Guarding against attacks

Machine learning (especially deep learning) models can be duped by malicious inputs known as ‘adversarial attacks’. It is possible to find input data combinations that can trigger perverse outputs from machine learning models, in effect ‘hacking’ them. This can be mitigated by simulating adversarial attacks on your own models and retraining models to recognise such attacks. Specialist software can be developed to ‘immunise’ your models against such attacks. This should be considered in the design phase.

AI is only as effective as the data it learns from.

3. Recognising the role of data as your key intellectual property

AI is only as effective as the data it learns from. Maintaining high quality data and continuously evaluating the effectiveness of the model will be key to a successful AI platform. As data and technology applications move on to the cloud, commercial advantage will be driven by the magnitude and scale of the ‘IP’ you hold.

Partnership with a vendor may inevitably involve data exchange – i.e. intentionally or unintentionally passing on valuable IP. It’s therefore important to understand the value of the data you’re sharing and closely monitor and manage its supply and use.

4. Looking out for systemic risks

The flash crash that hit financial markets in 2010 demonstrates what can happen when AI interact happen when multiple AIs interacts in unintended ways and this isn’t sufficiently monitored. Safeguards should include scenario planning, understanding of your own vulnerabilities and how to respond quickly.

[1] ‘PwC’s Dr Anand S Rao explores the ‘Five myths and facts about artificial intelligence’ in Predictive Analytics and Futurism, Society of Actuaries, December 2016