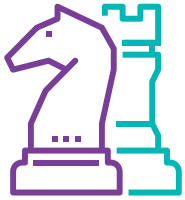


DETERMINING THE LEVEL OF HUMAN INVOLVEMENT IN AI-AUGMENTED DECISION-MAKING

3.8 This section is intended to help organisations determine the appropriate extent of human oversight in AI-augmented decision-making.

3.9 Having clarity on the objective of using AI is a key first step in determining the extent of human oversight. Organisations can start by deciding on their commercial objectives of using AI (e.g. ensuring consistency in decision-making, improving operational efficiency and reducing costs, or introducing new product features to increase consumer choice). These commercial objectives can then be weighed against the risks of using AI in the organisation's decision-making. This assessment should be guided by organisations' corporate values, which in turn, could reflect the societal norms or expectations of the territories in which the organisations operate.

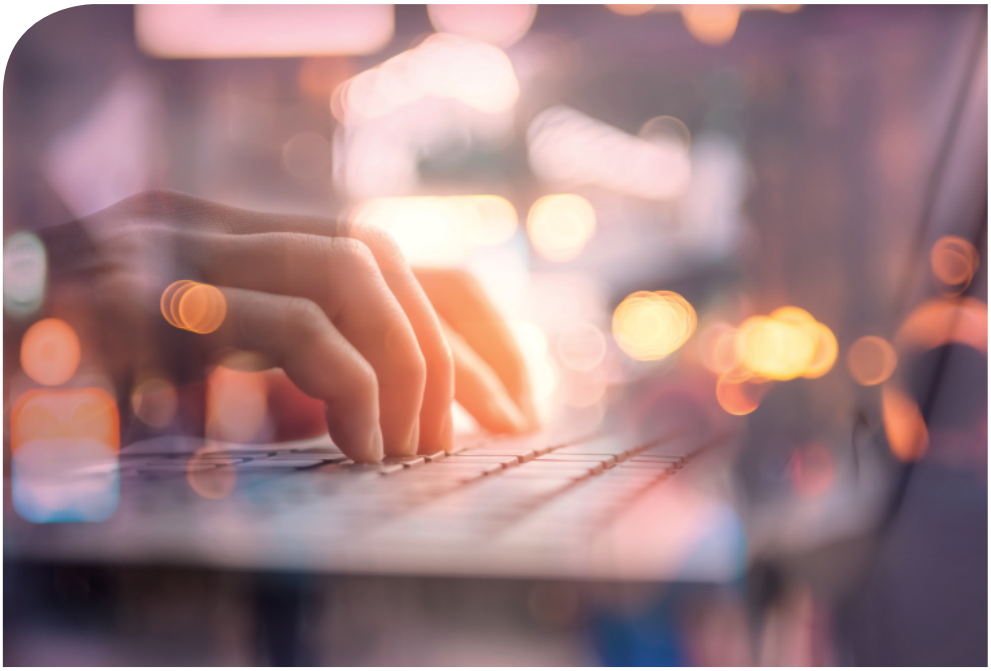
Before deploying AI solutions, organisations should decide on their commercial objectives of using AI, and then weigh them against the risks of using AI in the organisation's decision-making.

3.10 It is also desirable for organisations operating in multiple countries to consider the differences in societal norms, values and/or expectations, where possible. For example, gaming advertisements may be acceptable in one country but not in another. Even within a country, risks may vary significantly depending on where AI is deployed. For example, risks to individuals associated with recommendation engines that promote products in an online mall or automating the approval of online applications for travel insurance may be lower than the risks associated with algorithmic trading facilities offered to sophisticated investors.

3.11 Some risks to individuals may only manifest at group level. For example, widespread adoption of a stock recommendation algorithm might cause herding behaviour, increasing overall market volatility if sufficiently large numbers of individuals make similar decisions at the same time. In addition to risks to individuals, other types of risks may also be identified (e.g. risk to an organisation's commercial reputation).

3.12 Organisations' weighing of their commercial objectives against the risks of using AI should ideally be guided by their corporate values. Organisations can assess if the intended AI deployment and the selected model for algorithmic decision-making are consistent with their own core values. Any inconsistencies and deviations should be conscious decisions made by organisations with a clearly defined and documented rationale.

3.13 As identifying commercial objectives, risks and determining the appropriate level of human involvement in AI-augmented decision-making is an iterative and ongoing process, it is desirable for organisations to continually identify and review risks relevant to their technology solutions, mitigate those risks, and maintain a response plan should mitigation fail. Documenting this process through a periodically reviewed **risk impact assessment** helps organisations develop clarity and confidence in using the AI solutions. It will also help organisations respond to potential challenges from individuals, other organisations or businesses, and regulators.



WHAT ARE THE THREE BROAD APPROACHES OF HUMAN INVOLVEMENT IN AI-AUGMENTED DECISION-MAKING?

3.14 Based on the risk management approach described above, the Model Framework identifies three broad approaches to classify the various degrees of human oversight in the decision-making process:

a. **Human-in-the-loop** suggests that human oversight is active and involved, with the human retaining full control and the AI only providing recommendations or input. Decisions cannot be exercised without affirmative actions by the human, such as a human command to proceed with a given decision.

   For example, a doctor may use AI to identify possible diagnoses of and treatments for an unfamiliar medical condition. However, the doctor will make the final decision on the diagnosis and the corresponding treatment. This model requires AI to provide enough information for the human to make an informed decision (e.g. factors that are used in the decision, their value and weighting, correlations).

b. **Human-out-of-the-loop** suggests that there is no human oversight over the execution of decisions. The AI system has full control without the option of human override.

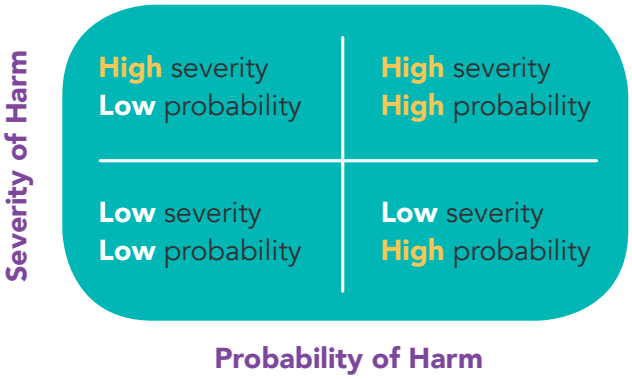
   For example, a product recommendation solution may automatically suggest products and services to individuals based on pre-determined demographic and behavioural profiles. AI can also dynamically create new profiles, then make product and service suggestions rather than relying on predetermined categories.

   A machine learning model might also be used by an airline to forecast demand or likely disruptions, and the outputs of this model are used by a solver module to optimise the airline's scheduling, without a human in the loop.

c. **Human-over-the-loop** (or human-on-the-loop) suggests that human oversight is involved to the extent that the human is in a monitoring or supervisory role, with the ability to take over control when the AI model encounters unexpected or undesirable events (such as model failure). This approach allows humans to adjust parameters during the operation of the algorithm. For example, a GPS navigation system plans the route from Point A to Point B, offering several possible routes for the driver to pick. The driver can alter parameters (e.g. due to unforeseen road congestions) during the trip without having to re-programme the route.

3.15 The Model Framework also proposes a design framework (structured as a matrix) to help organisations determine the level of human involvement required in AI-augmented decision-making. This design framework is structured along two axes: the (a) probability; and (b) severity of harm to an individual (or organisation) as a result of the decision made by an organisation about that individual (or organisation).

3.16 The definition of "harm" and the computation of probability and severity will depend on the context and vary from sector to sector. For example, the considerations of a hospital regarding the harm associated with a wrong diagnosis of a patient's medical condition will differ from the considerations of a clothing store's regarding the harm associated with a wrong product recommendation for apparels.



3.17 The matrix, however, should not be taken to imply that the probability of harm and severity of harm are the only factors to be considered in determining the level of human oversight in an organisation's decision-making process involving AI (although they are generally two of the more important factors).[4]

3.18 For safety-critical systems, it would be prudent for organisations to ensure that a person be allowed to assume control, with the AI system providing sufficient information for that person to make meaningful decisions or to safely shut down the system where human control is not possible.

[4] Other factors that organisations in various contexts may consider relevant, could also include: (a) the nature of harm (i.e. whether the harm is physical or intangible in nature); (b) the reversibility of harm, and as a corollary to this, the ability for individuals to obtain recourse; and (c) whether it is operationally feasible or meaningful for a human to be involved in a decision-making process (e.g. having a human-in-the-loop would be unfeasible in high-speed financial trading, and be impractical in the case of driverless vehicles).

USING THE PROBABILITY-SEVERITY OF HARM MATRIX

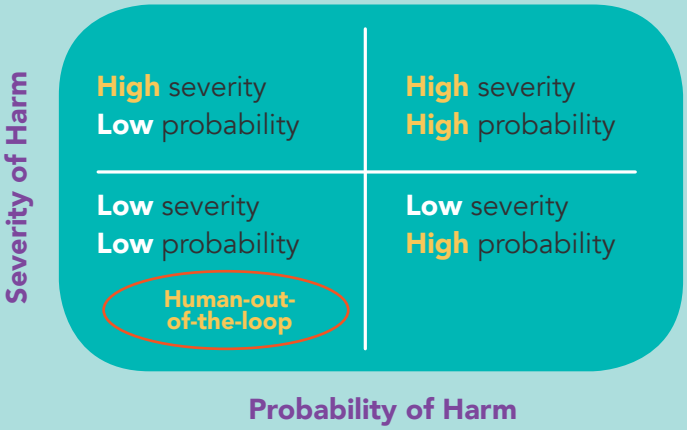
An online retail store wishes to use AI to fully automate the recommendation of food products to individuals based on their browsing behaviours and purchase histories. The automation will meet the organisation's commercial objective of operational efficiency.

Probability-severity assessment

The definition of ***harm*** can be the impact of making product recommendations that do not address the perceived needs of the individuals. The ***severity of harm*** in making the wrong product recommendations to individuals may be low since individuals ultimately decide whether to make the purchase. The ***probability of harm*** may be high or low depending on the efficiency and efficacy of the AI solution.

Degree of human intervention in decision-making process

Given the low severity of harm, the assessment points to an approach that requires no human intervention (i.e. human-out-of-the-loop).



Regular review

The organisation regularly reviews its approach (i.e. human-out-of-the-loop) to re-assess the ***severity*** and ***probability of harm***, and as societal norms and values evolve.

Note: *This is a simple illustration using bright-line norms and values. Organisations can consider testing this method of determining the AI decision-making model against cases with more challenging and complex ethical dilemmas.*