Annex 2 – Responsible Innovation and Responsible Artificial Intelligence

Applicants are suggested to consider the following requirements, as they develop the proposal, and provide justification:

- **Ethical Purpose and Societal Benefit**: the innovation is consistent with the overall ethical purposes of beneficence and non-maleficence. Civilian innovation systems must not be designed to harm or deceive people and should be implemented in ways that minimize any negative outcomes.
- **Privacy protection**: People's private data must be protected and kept confidential plus prevent data breaches that could cause reputational, psychological, financial, professional, or other types of harm.
- **Transparency & Explainability**: Communication with stakeholders must be maintained. Information about the capabilities and limitations of innovation/AI systems to support stakeholders in making informed choices about those systems. People must be informed when/if an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions, and that the decision outcomes of the AI system are explainable.
- **Accountability**: People and organizations responsible for the creation and implementation of the innovation ideas or AI algorithms should be identifiable and accountable for the impacts of that algorithm, even if the impacts are unintended.
- **Fairness**: The development or use of the innovation idea and or AI system must not result in unfair discrimination against individuals, communities, or groups.
- **Reliability & Safety**: the innovation (including AI) initiatives are expected to perform reliably and safely, remediate issues, and provide related information to customers. Monitoring, feedback, and evaluation must be ensured to identify and review new uses, identify and troubleshoot issues, manage and maintain the systems, and improve them over time.
- **Inclusiveness:** Inclusive design practices can help developers understand and address potential barriers that could unintentionally exclude people. Inclusiveness mandates that innovation (including AI) should consider all human races and experiences.
- **Conflict of Interest:** the proposed innovation and practices must not conflict with the interests of the agency funding body. Any interests which conflict, has the potential to conflict, or could reasonably be thought to conflict, must be declared and assessed.

For initiatives that employ Artificial Intelligence, applicants need to align with the Decision 1290/QD-BKHCN on Principles for Researching and Developing Responsible Artificial Intelligence Systems Responsible Artificial Intelligence of Vietnam. Unofficial English translation is as below:

MINISTRY OF SCIENCE AND TECHNOLOGY

SOCIALIST REPUBLIC OF VIET NAM

Independence - Freedom - Happiness

GUIDELINES

ON PRINCIPLES FOR RESEARCHING AND DEVELOPING RESPONSIBLE AI SYSTEMS

(Version 1.0)

1. Overview

Artificial intelligence (AI) systems are recognized for their potential to significantly benefit Vietnam's individuals, society and economy by addressing the challenges it faces. However, AI development and deployment entails inherent risks that necessitate mitigation measures to balance economic benefits of AI with ethical considerations and legal compliance. Therefore, specialised AI institutions should develop standards and guidelines, albeit soft and non-binding, to guide this process. Additionally, fostering the exchange of information on processes and best practices among stakeholders (including developers, service providers, and users) will build consensus to maximize AI benefits while managing associated risks.

In this spirit, AI systems' research and development in Vietnam should adhere to the following core principles:

- *First*, striving for a human-centered society where everyone benefits from both life and AI systems.
- *Second*, ensuring a fair balance between the benefits and risks of AI systems, specifically by: (1) promoting the benefits of AI through research, development, and innovation; and (2) minimizing the risks posed by AI systems to the rights and legitimate interests of organizations and individuals.
- *Third*, ensuring that the research and development of AI systems based on specific technologies or techniques maintain neutrality towards technology and that developers are not unduly influenced by the rapid evolvement of AI-related technologies in the future.
- *Fourth*, recognizing that current documents may serve as guidelines rather than mandates, and should be developed based on international standards and best practices to promote the research, development, and deployment of AI systems.
- *Fifth,* encouraging dialogue and information exchange among stakeholders involved with AI systems, recognizing that AI research and development across different fields vary in characteristics, applications, benefits, and risks.
- *Sixth,* committing to continuous research and updates of principles and guidelines to suit practical conditions.

2. Objectives

This guideline aims to:

- Encourage stakeholders in Vietnam to responsibly research, develop and deploy AI systems/applications.
- Foster safe and responsible research, development and deployment of AI systems/applications to minimize negative impacts on humans and society.
- Promote experience sharing in the research, development and deployment of AI systems/applications to build trust among users and society towards AI, thereby creating conducive conditions for AI research and development in Vietnam.

3. Scope

This guideline outlines several general principles for the responsible research and development of AI systems, intended for voluntary reference and application during the research, design, development, and provision of AI systems.

4. Subjects

Agencies, scientific and technological organizations, businesses and individuals engaged in the research, design, development and provision of AI systems are encouraged to adhere to the principles outlined in this guideline.

5. Concepts and Terminology

5.1. Artificial Intelligence (AI): Technology aiming to simulate human intelligence using machines, particularly computer systems.

5.2. AI systems: Machine-based systems that, for a given set of human-defined objectives, generates outputs such as content, predictions, recommendations, or decisions. (AI systems encompass models, data representations, knowledge, and processes used to perform tasks. They are developed using various AI techniques and approaches and can be designed to operate with varying levels of autonomy).

5.3. Model: A physical, mathematical, or other logical representation of a system, entity, phenomenon, process, or data.

5.4. Accountability: The state of being answerable for actions, decisions, and performance (of an AI system).

5.5. Transparency: The property of a system that provides relevant information about its operations to stakeholders. This includes aspects such as features, performance, limitations, components, procedures, metrics, design objectives, choices, assumptions, data sources, and labelling protocols. However, irrelevant disclosure of details about certain aspects of the system could potentially breach security, privacy or confidentiality requirements.

5.6. Risk: The effect of uncertainty on objectives.

5.7. Bias: Systematic differences in treating certain individuals or groups compared to others. (This treatment includes perception, observation, prediction, or decision-making).

5.8. Developers: Entities engaged in the research, development and provision of AI systems.

5.9. Users: Individuals using AI systems, including end-users or third-party providers of AI services/systems.

5.10. Stakeholders: Any individual, group, or organization that affects, is affected by or perceives themselves to be affected by a decision or action.

6. Principles for Responsible Research and Development of AI Systems, and instructions for implementations

6.1. Cooperation and Innovation

Developers should pay attention to the interoperability of AI systems, focusing on the capability of their AI systems to connect and interact effectively with other AI systems. This includes assessing the diversity of AI systems to: (1) maximize the benefits of AI systems through seamless integration, and (2) enhance coordination to mitigate potential risks.

To achieve this, developers should:

- Enhance collaboration to facilitate information sharing and ensure system interoperability.
- Prioritize the development of AI systems that conform to national or international technical standards, where applicable.
- Strengthen data format standardization and the openness of interfaces and protocols, including application programming interfaces (APIs).
- Anticipate unintended risks/events arising from the integration and interaction between AI systems.
- Promote the exchange of information on licenses and intellectual property rights such as patents (excluding trade secrets) to enhance interoperability.
- Contribute to sustainable economic development and help address economic and social challenges.
- Foster cooperation across industries, sectors and among stakeholders to expand the AI community in Vietnam.

6.2. Transparency

Developers should prioritize controlling the inputs and outputs of AI systems and their ability to clearly explain the related analyses. This principle is particularly crucial for AI systems that can impact the lives, health, privacy, or property of users, including related third parties. In such cases, developers need to focus on clearly identifying the inputs and outputs of AI systems and explaining them based on the characteristics of the applied technology and its usage. This transparency is essential for gaining societal acceptance and users' trust in AI systems.

6.3. Controllability

Developers should ensure AI systems are controllable. To assess risks related to controllability, developers need to conduct pre-assessments (which involve evaluating whether the system meets corresponding technical requirements and standards). One method of risk assessment is conducting tests in controlled environments such as laboratories or testing facilities where safety protocols are implemented before practical deployment.

In addition, to ensure controllability, developers should emphasize system monitoring (using evaluation/monitoring tools or adjustments/updates based on user feedback) and response measures (such as system shutdown or network disconnection), either through human intervention or reliable AI systems.

6.4. Safety

Developers must ensure that AI systems do not cause harm to the lives, health, or property of users or third parties, whether directly or indirectly. Essentially, developers are encouraged to refer to relevant international standards, with particular emphasis on potential changes in outputs or processes resulting from the training of AI systems.

- Conduct pre-assessments to identify and mitigate risks associated with AI systems.
- Implement safety measures throughout the development stages to ensure intrinsic safety (reducing risk factors such as energy levels of devices that could cause incidents) and functional safety (minimizing risks using additional control devices such as automatic shutdown mechanisms in case of accidents).
- Clearly articulate the designer's ideas and intentions behind AI systems and explain their relevance to stakeholders, as well as their potential impacts on the lives, health, or property of users and third parties. (This includes strategies to protect human life, health, and property in the event of accidents involving AI-enabled robots).

These measures are crucial to mitigating risks and ensuring the safe deployment and operation of AI systems in various contexts.

6.5. Security

Developers should prioritize the AI system’s security. In addition to adhering to prescribed documents and instructions and implementing necessary information security measures (as mandated by relevant authorities), developers should focus on the following aspects, with particular emphasis on potential changes in outputs or processes resulting from the training of AI systems.

- Emphasize reliability (ensuring that activities are executed as planned and not unlawfully influenced by third parties) and the system’s ability to withstand various forms of attacks or physical accidents. Simultaneously, ensure: (1) security; (2) integrity; and (3) availability of essential information related to the information security of the AI system.

- Implement necessary measures to uphold security throughout the AI system development process, based on the characteristics of the applied technologies (security by design).

6.6. Privacy

Developers must ensure that AI systems do not compromise the privacy of users or third parties. Privacy in this context includes spatial privacy (personal tranquillity), information privacy (personal data), and the confidentiality of communication information. It is essential for developers to comply with existing regulations and guidelines set by relevant authorities, consult international standards and guidelines concerning privacy rights, and adopt the following additional guidelines, with particular emphasis on potential changes in outputs or processes resulting from the training of AI systems.

- Conduct pre-assessments of privacy infringement risks and evaluate privacy impacts (from the design phase).
- Implement measures appropriate to the characteristics of the technology applied throughout the AI system development process (from the design phase) to minimize privacy infringement upon deployment.

6.7. Respect for Human Rights and Dignity

In developing AI systems involving humans, developers must pay special attention to respecting the rights and dignity of individuals involved. As far as practicable and depending on the characteristics of the technology applied, developers should take measures to ensure that there is no discrimination or unfairness due to bias (prejudice) in the data when training AI systems.

Developers need to implement preventive measures to ensure that AI systems uphold human values and social ethics in accordance with Vietnam's fundamental principles (such as patriotism, solidarity, self-reliance, humanity, honesty, responsibility, discipline, and creativity).

6.8. User support

Developers must ensure that AI systems support users and empower them to make informed choices. To achieve this, AI system developers should focus on the following aspects:

- Design interfaces that promptly deliver relevant information to assist users in decision-making and using the systems conveniently.
- Consider providing users with features that allow timely and appropriate choices (e.g., default configuration, easy-to-understand settings, feedback, alert notifications, error fixing).
- Implement measures to enhance the user-friendliness of AI systems for vulnerable groups in society (such as the elderly and people with disabilities).

Additionally, developers should provide users with necessary information, including explanations of potential changes in outputs or processes resulting from the training of AI

systems, and clear instructions on how to use AI systems to mitigate unforeseen risks (such as through user manuals or risk mitigation measures).

6.9. Accountability

Developers must uphold their accountability to stakeholders, including users, to build trust in the AI systems they develop. Specifically, developers should provide users with necessary information to enable informed decisions when selecting and utilizing AI systems. Moreover, to enhance societal acceptance of AI systems, developers should: (1) provide users with technical descriptions, algorithms, and safety mechanisms of their AI systems; and (2) seek feedback from and engage in dialogue with stakeholders.

Developers also should share information and collaborate closely with suppliers to ensure updates and manage issues related to the provision and use of AI systems.