KỶ YẾU HỘI THẢO
CONFERENCE PROCEEDINGS

Đạo đức
trong thời đại Trí tuệ Nhân tạo
Bioethics
in the era of Artificial Intelligence (AI)

TIỂU BAN KHOA HỌC XÃ HỘI
ỦY BAN QUỐC GIA UNESCO VIỆT NAM

NHÀ XUẤT BẢN HỒNG ĐỨC

MỤC LỤC - TABLE OF CONTENTS

Lời nói đầu
Introduction

Ông Christian Manhart
Trưởng Đại diện Văn phòng UNESCO tại Việt Nam

Mr. Christian Manhart
Head of Office and UNESCO Representative to Vietnam

Thưa Quý độc giả,

Quý độc giả đã hiểu được bao nhiêu về Trí tuệ Nhân tạo (AI) rồi?

Cho phép tôi lấy một ví dụ để chia sẻ với quý độc giả: một chiếc ô tô tự lái bị mất phanh đang lao với vận tốc tối đa hướng về phía một người bà và một em bé. Nếu có thể chệch hướng một chút thì một trong hai người sẽ có thể tránh được vụ tai nạn. Vậy bạn sẽ chọn ai? Người bà hay em bé?

Trên đây là một ví dụ điển hình về một tình huống lựa chọn theo đạo đức mà lại tiến thoái lưỡng nan để chúng ta thấy được rằng việc phát triển công nghệ có đạo đức quan trọng đến như thế nào. Ở đây, đạo đức giúp chúng ta xem xét xem chúng ta có đi đúng hướng để phát triển các công nghệ cho một thế giới mà chúng ta muốn sống trong đó hay không. Vậy ai sẽ là người quyết định diện mạo xã hội của chúng ta? Liệu chúng ta có để cho một rô-bốt đưa ra quyết định khi mà chúng ta còn chưa biết rô-bốt này làm thế nào để đưa ra kết luận của mình?

Đây chính là lý do mà lần đầu tiên trong lịch sử UNESCO đứng ra xây dựng một công cụ toàn cầu Khuyến nghị về Đạo đức trong Trí tuệ Nhân tạo. Và bây giờ tại Việt Nam, chúng ta cần có những quyết định về việc phát triển và ứng dụng những khía cạnh đạo đức như vậy. UNESCO và cá nhân tôi xin được chúc mừng cho sự thành công của Hội thảo *"Đạo đức trong thời đại Trí tuệ Nhân tạo"* được tổ chức hồi tháng 11 vừa qua. Thưa Quý độc giả, tôi hoàn toàn tin tưởng rằng, với sự nỗ lực to lớn của Quý vị, Trí tuệ Nhân tạo sẽ tiếp tục đóng góp vào quá trình giảm thiểu những thách thức to lớn của thế giới chúng ta hiện nay như dịch Covid-19, biến đổi khí hậu, vấn đề về bảo vệ môi trường hay phát triển kinh tế bền vững, v.v.

Nhân đây, tôi xin gửi lời cảm ơn Viện Hàn lâm Khoa học Xã hội Việt Nam, Ủy ban Quốc gia UNESCO Việt Nam cùng các bên liên quan khác đã khởi xướng sáng kiến tổ chức hội thảo này để chúng ta có thể cùng nhau chia sẻ ý tưởng, khích lệ động viên và cùng nhau xây dựng một nền Trí tuệ Nhân tạo có nền tảng đạo đức vững chắc.

Tôi cũng trân trọng chia sẻ với độc giả tài liệu Khuyến nghị của UNESCO về Đạo đức trong AI mà UNESCO đã công bố tại Paris hồi tháng 11 vừa qua để gợi ý xây dựng những công cụ pháp lý cần thiết nhằm đảm bảo sự phát triển Trí tuệ nhân tạo mà trog đó xem trọng sự phát triển của con người.

Cũng xin nói thêm rằng UNESCO sẵn sàng hỗ trợ Việt Nam trong lĩnh vực này bởi chúng tôi mong muốn Trí tuệ Nhân tạo sẽ trở thành mối quan tâm chung của tất cả chúng ta. Vì vậy, văn phòng đại diện UNESCO tại Hà Nội cũng rất mong có cơ hội được hợp tác với các đối tác trong lĩnh vực này.

Xin chúc Quý độc giả tìm thấy những thông tin hữu ích trong cuốn tài liệu này!

Trân trọng,

Christian Manhart.

Dear Readers,

How much do we know about Artificial Intelligence (AI)? Let's take one example: a self-driving car with broken brakes going at high speed towards a grandmother and a child. If it deviates a little bit, one of them could be saved. Whom will it choose? The grandmother or the child?

This is a typical example of an ethical dilemma to show how important it is to develop ethics in technologies. Ethics will help us question if we are in the right direction to develop technologies for a world we want to live in or not. Who will decide how the society is going to look like? Can we bring a robot in to make decisions when we are not sure how it reaches its conclusions?

That is why, for the first time, UNESCO has embarked developing a global Recommendation on Ethics of Artificial Intelligence. And here in Viet Nam, you have to decide on the development and applications of these ethics in your country. I wish to congratulate you for the success of the conference on *"Bioethics in the era of Artificial Intelligence"* last November. I strongly believe that, with your great efforts, AI will continue to prove its value in contributing towards the mitigation of the enormous challenges of our world, such as the COVID-19 pandemic, climate change, protection of environment and also sustainable economic growth.

Taking this opportunity, I would like to thank the Viet Nam Academy of Social Sciences, Viet Nam National Commission for UNESCO and other stakeholders for taking this initiative, where we could share ideas, get inspirations and formulate recommendations to place AI on a strong ethical basis. I am also happy to introduce to you the Recommendation on Ethics of AI, announced at UNESCO

HQ during the General Conference last November, to provide guidance to the creation of the necessary legal frameworks, to ensure a human development of AI.

Please be assured, that UNESCO will provide any support and technical assistance to Viet Nam in this area, to ensure that AI will be in the interest of the people. Therefore, the UNESCO Ha Noi Office looks forward to further opportunities of cooperation.

I wish you all find useful information in this publication.

Thank you,

Christian Manhart.

Báo cáo khai mạc
Opening Speeches

GS. TS. Đặng Nguyên Anh

Phó Chủ tịch Viện Hàn lâm Khoa học xã hội Việt Nam

Phó Chủ tịch Ủy ban Quốc gia UNESCO Việt Nam

Trưởng Tiểu ban Khoa học xã hội, UBQG UNESCO Việt Nam

Prof. Dr. Dang Nguyen Anh

Vice President, Vietnam Academy of Social Sciences

Vice President, Vietnam National Commission for UNESCO

Chairman, Sub-commission for Social Sciences,

Vietnam National Commission for UNESCO

Kính thưa Ông Phạm Văn Linh, Phó Chủ tịch Chuyên trách Hội đồng Lý luận Trung ương,

Kính thưa Ông Mai Phan Dũng, Vụ trưởng và Thư ký Ủy ban Quốc gia UNESCO Việt Nam,

Thưa các đồng chí đại diện lãnh đạo các Bộ, Ban, Ngành địa phương, các chuyên gia, các nhà khoa học trong nước và quốc tế,

Hôm nay Viện Hàn lâm KHXH VN phối hợp với UBQG UNESCO VN tổ chức hội thảo quốc tế với chủ đề *Đạo đức trong Thời đại Trí tuệ Nhân tạo*. Đây là hoạt động trong khuôn khổ chương trình nghiên cứu toàn cầu năm 2021 của Tổ chức Giáo dục, Khoa học và Văn hóa Liên Hợp Quốc UNESCO.

Lời đầu tiên, xin được thay mặt Viện Hàn lâm KHXH Việt Nam và Ban Tổ chức Hội thảo cho phép tôi được gửi đến các quý vị lời chào trân trọng và lời chúc tốt đẹp nhất, kính chúc sức khỏe toàn thể quý vị đại biểu, chúc hội thảo thành công rực rỡ.

Trong nhiều năm qua Ủy ban Quốc gia UNESCO Việt Nam đã phối hợp chặt chẽ với các Bộ, Ban, Ngành liên quan, trong đó có Viện Hàn lâm KHXH Việt Nam và Ủy ban UNESCO về khoa học xã hội với mục đích tăng cường vai trò tư vấn thực hiện các mục tiêu phát triển bền vững quốc gia. Thông qua các hội nghị, hội thảo, kết quả nghiên cứu đã được chuyển tới các nhà hoạch định chính sách, chia sẻ với công chúng và tư vấn chính sách cho Đảng và Nhà nước. Đây là một trong những chức năng quan trọng gắn với nhiệm vụ mà Viện Hàn lâm KHXH Việt Nam và Tiểu ban khoa học xã hội của Ủy ban Quốc gia UNESCO Việt Nam về đã và đang thực hiện.

Như chúng ta đều biết, sự phát triển bùng nổ của mạng Internet và Cách mạng Công nghiệp 4.0 đã đem lại những phát minh làm thay đổi hoàn toàn cách thức mà xã hội và con người vận hành dựa trên nền tảng trí tuệ nhân tạo, siêu tự động hóa và siêu kết nối. Hiện nay, thế giới đã sản xuất ra các robot có thể suy nghĩ như con người, với hệ thần kinh như một phần của bộ não, có thể nghe, nhìn, chuyển động và sử dụng ngôn ngữ của chúng ta. Trí tuệ nhân tạo còn giúp nhận diện giọng nói, khuôn mặt, có khả năng lập luận, và thậm chí có khả năng tự sửa lỗi. Trí tuệ nhân tạo đang xoá mờ ranh giới giữa các yếu tố vật chất, kĩ thuật số và sinh học. Sự phát triển của nó diễn ra rất nhanh, từng giờ, từng ngày và mang lại những lợi ích vô cùng to lớn cho tất cả các lĩnh vực của đời sống xã hội, như kinh tế, sức khỏe, tài chính ngân hàng cho tới khoa học, giáo dục và cuộc sống sinh hoạt hàng ngày của chúng ta. Công nghệ trí tuệ nhân tạo đứng sau những chiếc xe oto tự lái mà cách đây 10 năm có lẽ không ai trong chúng ta có thể hình dung được. Nhiều quốc gia đang đặt ra kế hoạch đầy tham vọng là trở thành nhà lãnh đạo toàn cầu về trí tuệ nhân tạo với những đột phá trong lĩnh vực này.

Tốc độ phát triển và sự tiến bộ nhanh chóng của trí tuệ nhân tạo trong 5-10 năm qua đã thúc đẩy việc ban hành các biện pháp để bảo vệ con người khi công nghệ trí tuệ nhân tạo trở nên phổ biến, và là một phần không thể thiếu trong các sản phẩm, dịch vụ. Xe ô tô tự hành là phương tiện có khả năng cảm nhận môi trường và di chuyển mà không có sự tham gia của con người. Để chiếc xe di chuyển an toàn, một lượng lớn dữ liệu được thu thập bằng vô số cảm biến trên xe, và được xử lý bởi hệ thống máy tính tự động. Tuy nhiên, xe tự hành luôn phải đưa ra quyết định đúng đắn trong các tình huống giao thông. Khi chọn cách phanh gấp để tránh đâm va vào người đi bộ thì đó cũng là một quyết định đạo đức nhằm chuyển rủi ro từ người đi bộ sang những người trong xe, cho thấy tầm quan trọng của đạo đức trong sự phát triển của công nghệ trí tuệ nhân tạo.

Trí tuệ nhân tạo có tác động rất lớn trong nhiều lĩnh vực, bao gồm cả các ngành luật và tư pháp cho đến lĩnh vực hành chính, lập pháp. Trí tuệ nhân tạo có thể giúp tạo ra một hệ thống tư pháp hình sự công bằng, trong đó máy móc giúp đánh giá, cân nhắc các yếu tố hiệu quả và chính xác hơn trong tư vấn và xét xử tranh tụng bằng cách tận dụng tốc độ và khả năng nhập dữ liệu lớn. Trong các trường hợp đó, trí tuệ nhân tạo sẽ đưa ra những quyết định sáng suốt, không thiên vị và chủ quan. Các hệ thống phần mềm dành cho thẩm phán có thể ứng dụng công cụ trí tuệ nhân tạo để hỗ trợ trong việc ban hành quyết định, thực thi công lý.

Tuy nhiên vẫn còn nhiều vấn đề về trí tuệ nhân tạo mà chúng ta chưa hiểu rõ, chưa thống nhất. Khó có thể khẳng định hay lường trước được những tác động xã hội cũng như những tác động tới phạm trù đạo đức trong lĩnh vực trí tuệ nhân tạo. Trên thực tế, trí tuệ nhân tạo có thể dẫn đến gia tăng bất bình đẳng trong xã hội, tạo nên một thị trường lao động phân tách, chia ngạch với mức lương quá chênh lệch, làm trầm trọng thêm phân hóa, phân tầng xã hội. Điều gì sẽ xảy ra khi máy móc trở nên thông minh hơn và chúng ta không điều khiển đc chúng? Điều gì sẽ xảy ra khi máy móc có trí thông minh vượt con người, khi tạo ra siêu trí tuệ là điều hoàn toàn có thể trong điều kiện hiện nay. Trí tuệ nhân tạo không biết làm thế nào để chấp nhận cái tốt và cái xấu, không phân biệt được những giá trị đạo đức khi đưa ra quyết định giúp cho con người. Sự thiếu minh bạch của các công cụ và dữ liệu lớn được sử dụng, dẫn đến các quyết định không phải lúc nào cũng dễ hiểu và phù hợp đối với con người. Trí tuệ nhân tạo có thể thiếu trung lập dẫn đến các quyết định thiếu chính xác, thậm chí phân biệt đối xử hoặc vi phạm nhân quyền, đời tư của con người. Đó là lí do trí tuệ nhân tạo cần được thiết kế và ứng dụng với những quy tắc đạo đức, ở đây câu trả lời chinh là con người với khả năng vượt trội và trách nhiệm giải trình, bởi lẽ con người tạo ra máy móc và cũng chính là chủ thể để quy định các nguyên tắc mà máy móc phải tuân thủ.

Với tiêu đề *Đạo đức trong Thời đại Trí tuệ Nhân tạo*, cuộc hội thảo hôm nay nhằm mục đích hình thành một cơ sở lý luận, chia sẻ kinh nghiệm quốc tế và phát triển trí tuệ nhân tạo, xem xét chiều cạnh đạo đức và nhân văn của sự phát triển. Có thể nói, đây là vấn đề rất mới, cần được nghiên cứu thấu đáo, đánh giá một cách khoa học. Đạo đức trí tuệ nhân tạo bao gồm một hệ thống các nguyên tắc nhằm quy định sự phát triển và sử dụng có trách nhiệm của công nghệ trí tuệ nhân tạo. Bộ quy tắc đạo đức, hay còn gọi là nền tảng giá trị trí tuệ nhân tạo, được ban hành nhằm mục đích hướng dẫn cho các bên khi đối mặt trước một quyết định đạo đức liên quan đến việc sử dụng trí tuệ nhân tạo. Tôi tin tưởng rằng, thông qua các bài trình bày của các chuyên gia, cùng các tranh luận, thảo luận thẳng thắn, chúng ta sẽ nhận diện được nội hàm và bản chất của các khái niệm, xem xét trí tuệ nhân tạo dưới góc nhìn đạo đức, đồng thời gợi mở và đề xuất được những định hướng, chính sách pháp luật phù hợp với các quy chuẩn đạo đức trong trí tuệ nhân tạo ở Việt Nam.

Nhân dịp này tôi xin chân thành cám ơn Ủy ban Quốc gia UNESCO Việt Nam và cá nhân ông Mai Phan Dũng và toàn thể các quý vị đại biểu đã tham gia, đồng hành cùng Viện Hàn lâm KHXH Việt Nam, tiểu ban UNESCO về KHXH, xin cám ơn các vị khách quốc tế đã tham dự Hội thảo này.

Thay mặt Ban Tổ chức Hội thảo tôi xin cám ơn các quý vị đã tham dự trực tiếp cũng như các đại biểu tham dự trực tuyến.

Thay mặt Viện Hàn lâm KHXH Việt Nam tôi xin chính thức khai mạc Hội thảo *Đạo đức trong Thời đại Trí tuệ Nhân tạo*, xin kính chúc các quý vị đại biểu hạnh phúc và thành công, xin trân trọng cám ơn các quý vị!

Dear Assoc. Prof. Dr. Pham Van Linh, Vice Chairman of the Central Theoretical Council of the Communist Party of Vietnam,

Dear Mr. Mai Phan Dung, Director-General of Department of Cultural Affairs and UNESCO (MOFA), Secretary-General of Vietnam National Commission for UNESCO,

Leaders representing Ministries and Agencies,

Vietnamese and International scholars

Today, the Vietnam Academy of Social Sciences in collaboration with the Vietnam National Commission for UNESCO held an international conference with the theme ***Bioethics in the Era of Artificial Intelligence (AI)***. This is an activity within the framework of a global research program in 2021 of the United Nations Educational, Scientific and Cultural Organization (UNESCO).

First of all, on behalf of the Vietnam Academy of Social Sciences and the Conference's Organizing Committee, allow me to send the respectful greetings and best wishes of good health to all the delegates, and may our Conference be a great success.

Over the past many years, Vietnam National Commission for UNESCO has closely coordinated with relevant Ministries, Departments and Agencies, including the Vietnam Academy of Social Sciences and the UNESCO Sub-Commission for Social Sciences with the aim of strengthening the advisory role in implementing national sustainable development goals. Through conferences and seminars, research outcomes have been presented to policy makers, shared with the public, and policy consultation has been provided for the Party and State. This is one of the important functions associated

with other tasks that the Vietnam Academy of Social Sciences and the UNESCO Sub-Commission for Social Sciences have been performing.

As we all know, the explosive development of the Internet and the Industrial Revolution 4.0 have brought inventions that completely change the way that society and people operate basing on artificial intelligence, super automation and hyper-connectivity. Currently, the world has produced robots that can think like humans, with the nervous system as part of the brain, able to hear, see, move and use our language. Artificial intelligence also helps recognize voices and faces, has the ability to reason, and even has the ability to self-correct. Artificial intelligence is blurring the lines between physical, digital and biological elements. Development of AI is happening very fast, every hour, every day and brings enormous benefits to all areas of social life, such as from economy, healthcare, finance and banking to science, education and our daily lives. Artificial intelligence technology behind self-driving cars is perhaps the thing that none of us could have imagined 10 years ago. Many countries are setting out ambitious plans to become a global leader in artificial intelligence with breakthroughs in the field.

The development speed and rapid advancement of artificial intelligence in the past 5-10 years have prompted the enactment of measures to protect people as AI technology becomes widespread, and is a integral part of products and services. Self-driving cars are vehicles capable of sensing the environment and moving without human involvement. In order for the vehicle to move safely, an enormous amount of data is collected by numerous sensors on the vehicle, and processed by an automated computer system. However, such autonomous vehicles always have to make the right decisions in traffic situations. Choosing to brake hard to avoid hitting a pedestrian is also an ethical decision to transfer risk from the pedestrian to the

people inside the vehicle, demonstrating the importance of ethics in the development of artificial intelligence technology.

Artificial intelligence produces huge impacts in many fields, including the legal and judicial sectors to the administrative and legislative sectors. Artificial intelligence can help create a fair criminal justice system, in which machines help evaluate and weigh factors more effectively and accurately in litigation and adjudication by taking advantages of speed and ability to import large data. In such cases, artificial intelligence will make informed, unbiased and subjective decisions. Software systems for judges can apply artificial intelligence tools to assist in decision-making and justice enforcement.

However, there are still many issues about artificial intelligence that we do not fully understand and agree on. It is difficult to confirm or predict the social impacts as well as the ethical implications in the field of artificial intelligence. In fact, artificial intelligence can lead to increased inequality in society, creating a segregated labor market, setting ranks with excessively different wages, exacerbating social divisions and stratification. What will happen when machines get smarter and we can't control them? What will happen when machines have intelligence surpassing humans, when creating super intelligence is completely possible in today's conditions. Artificial intelligence does not know how to accept good from bad, does not distinguish moral values when making decisions to help people. The lack of transparency of the tools and big data used leads to decisions that are not always understandable and relevant to humans. Artificial intelligence can lack neutrality, leading to inaccurate decisions, even discriminatory or violating human rights and privacy. That is why artificial intelligence needs to be designed and applied along with ethical principles, here the answer is on humans with superior ability and accountability, because humans create machines and are also the subject to stipulate the principles that the machine must follow.

With the theme *Bioethics in the Era of Artificial Intelligence (AI)*, today's conference aims to form a theoretical base, share international experiences on artificial intelligence development, considering the ethical and humanity aspects of development. It can be said that this is a very new issue that needs to be thoroughly researched and evaluated scientifically. Ethics of artificial intelligence includes a system of principles that govern the responsible development and usage of artificial intelligence technology. The Code of Ethics, or AI Value Platform, is intended to guide parties when faced with an ethical decision regarding the use of artificial intelligence. I believe that, through the presentations of experts, together with open and frank discussions, we will be able to identify the contents and nature of the concepts, and look at artificial intelligence from an ethical perspective, and at the same time point out and suggest orientations, policies and laws in accordance with ethical standards of artificial intelligence in Vietnam.

On this important occasion, I would like to sincerely thank the Vietnam National Commission for UNESCO and especially Mr. Mai Phan Dung personally as well as all the delegates who participated and accompanied the Vietnam Academy of Sciences and the UNESCO Sub-commission for social sciences. Special thanks to international guests attending our conference today.

On behalf of the Conference's Organizing Committee, I would like to thank all of you who are attending in person as well as those who are attending online.

On behalf of the Vietnam Academy of Social Sciences, I would like to officially announce the start of the Conference on Bioethics in the Era of Artificial Intelligence (AI).

Finally, I would like to wish you all happiness and success.

Thank you very much!

Ông Mai Phan Dũng

Vụ trưởng Vụ Ngoại giao Văn hóa và UNESCO

Tổng thư ký Ủy ban Quốc gia UNESCO Việt Nam

Mr. Mai Phan Dung

Director General, Dept. of Cultural Diplomacy & UNESCO, MOFA

Secretary-General, Vietnam National Commission for UNESCO

Kính thưa GS.TS. Đặng Nguyên Anh, Phó Chủ tịch Viện Hàn lâm Khoa học xã hội Việt Nam, Phó Chủ tịch Ủy ban Quốc gia UNESCO Việt Nam.

Thưa Quý vị Đại biểu,

Trước hết cho phép tôi thay mặt Ủy ban Quốc gia UNESCO Việt Nam nhiệt liệt chào đón các quí vị đại biểu tham dự Hội thảo ngày hôm nay, đồng thời đánh giá cao sáng kiến, nỗ lực của Viện Hàn lâm Khoa học xã hội Việt Nam trong việc tổ chức Hội thảo này.

Kính thưa quí vị đại biểu,

Chúng ta đang sống trong một thế giới mà sự đi lên của công nghệ là những xu thế tất yếu. Trí tuệ nhân tạo sẽ chính là một xu thế công nghệ đánh dấu bước ngoặt chuyển mình của cả thế giới. Trí tuệ nhân tạo được thừa nhận là đã có đóng góp to lớn cho sự phát triển và thịnh vượng của các quốc gia. AI là công cụ sản xuất rất quan trọng, làm thay đổi phương thức sản xuất, có thể dẫn đến thay đổi quan hệ sản xuất, trình độ phát triển của xã hội. Trí tuệ nhân tạo trong tương lai có thể trực tiếp trở thành lực lượng sản xuất, thay thế và giải phóng sức lao động của con người.

Tuy nhiên, AI cũng đặt ra rất nhiều vấn đề và thách thức lớn, như là: khoét sâu hơn sự chênh lệch về công nghệ giữa các quốc gia; làm sâu sắc hơn sự bất bình đẳng trên thế giới, giữa các quốc gia và trong từng quốc gia; đe dọa đa dạng văn hóa; đe dọa an ninh quốc gia; tác động tiêu cực đến môi trường và hệ sinh thái… Một vấn đề rất đáng lo ngại và là chủ đề của Hội thảo hôm nay, đó là "Đạo đức trí tuệ nhân tạo", các tác động tiêu cực của AI đến quyền và phẩm giá con người, mối quan hệ giữa con người và máy móc; xâm phạm quyền tự do cá nhân và các quyền cơ bản khác…

Với sứ mệnh "phòng thí nghiệm của các ý tưởng", UNESCO đã sớm nhận thấy cần phải xây dựng các tiêu chuẩn quốc tế về đạo đức trong thời đại AI, hỗ trợ các quốc gia xây dựng chính sách để tận dụng lợi thế của AI nhưng phải hạn chế các mặt tiêu cực. Vì vậy, UNESCO đang xây dựng Dự thảo Khuyến nghị về Đạo đức trong Trí tuệ Nhân tạo, xác định và nhấn mạnh các giá trị tích cực, các mặt tiêu cực, các nguyên tắc cơ bản của AI, hướng dẫn cho việc phát triển và sử dụng có trách nhiệm các công nghệ AI, nguyên tắc về quản trị, cộng tác, thích ứng, nghĩa vụ của khu vực công và tư. Dự thảo đã được các quốc gia thành viên đóng góp ý kiến và dự kiến sẽ thông qua tại Kỳ họp Đại Hội đồng lần thứ 41 (diễn ra từ ngày 09 - 24/11/2021 tại Paris).

Kính thưa quí vị đại biểu,

Tôi hoan nghênh những đóng góp tích cực của Viện Hàn lâm Khoa học xã hội Việt Nam, cơ quan đầu mối của Việt Nam trong các hoạt động của UNESCO liên quan đến lĩnh vực khoa học xã hội nói chung và trong sáng kiến tổ chức Hội thảo "Đạo đức trong thời đại Trí tuệ Nhân tạo (A.I)". Chúng tôi đánh giá cao việc Viện Hàn lâm Khoa học xã hội Việt Nam đã có đóng góp cho dự thảo Khuyến nghị về Đạo đức trong Trí tuệ nhân tạo của UNESCO. Điều này vừa thể hiện sự trách nhiệm của Việt Nam đối với một vấn đề quan trọng của thế giới, vừa lồng ghép được các lợi ích của Việt Nam trong bối cảnh Việt Nam sẽ cần nâng cao năng lực Trí tuệ nhân tạo và năng lực sản xuất của quốc gia để bắt kịp sự phát triển của thế giới.

Tôi tin tưởng rằng Hội thảo sẽ là dịp để nhìn lại thực trạng những thay đổi lớn trong xã hội mà cuộc Cách mạng Công nghiệp 4.0 và sự phát triển của Trí tuệ nhân tạo mang lại, nhận diện rõ hơn những thách thức về phương diện đạo đức của Trí tuệ Nhân tạo, học hỏi kinh nghiệm quốc tế, đề ra giải pháp để vừa nâng cao năng trình độ kỹ thuật, đồng thời có sự chuẩn bị cho các giai đoạn phát triển cao hơn của Trí tuệ nhân tạo tại Việt Nam.

Tôi cũng khuyến nghị trong thời gian tới, Việt Nam tăng cường hợp tác với UNESCO trong lĩnh vực này, thúc đẩy nhận thức, giáo dục về Trí tuệ nhân tạo, xây dựng chính sách phù hợp, xử lý những thách thức do trí tuệ nhân tạo mang lại, đặc biệt từ góc nhìn đạo đức, nhân văn.

Xin chúc Hội thảo thành công. Chúc các Quý vị đại biểu sức khỏe.

Xin trân trọng cảm ơn./.

Dear Prof.Dr. Dang Nguyen Anh, Vice President of Vietnam Academy of Social Sciences, Vice Chairman of the Vietnam National Commission for UNESCO

Distinguished Delegates,

First of all, allow me on behalf of the Vietnam National Commission for UNESCO to warmly welcome all the distinguished delegates to the Conference today, and also, I would like to express the high appreciation to the initiative and efforts of the Vietnam Academy of Social Sciences in organizing this important Conference.

Ladies and Gentlemen,

We are living in the world where the advancement of technology is inevitable. Artificial intelligence will be a technology trend that marks a turning point in the world's transformation. Artificial intelligence is recognized as having made a significant contributions to the development and prosperity of all nations. AI is a very important production tool, which can change production methods and lead to changes in production relations and the development level of society. Artificial intelligence in the future can directly become a productive force that may replace and liberate human labor.

However, AI is also posing various big problems and challenges, such as: widening technological gaps between countries; deepening inequalities around the world, between countries and within each country; threatening cultural diversity; threatening national security; causing negative impacts on the environment and ecosystems… An issue of great concern and topic of today's Conference is “ethics in artificial intelligence”, the negative impacts of AI on rights and human

dignity, the relationship between man and machine; infringing upon individual liberties and other fundamental rights…

With the mission of "the laboratory of ideas", UNESCO recognized early on the need to develop international ethical standards in the age of AI, assisting countries in formulating policies to take advantage of AI at the same time limiting the negative aspects. Therefore, UNESCO is developing a Draft Recommendation on Ethics of Artificial Intelligence, identifying and emphasizing the positives, negatives, fundamentals of AI, guidelines for the development and responsible use of AI technologies, principles of governance, collaboration, adaptation, public and private sector's obligations. The draft has been commented on by member states and is expected to be approved at the 41st General Assembly session (taking place from November 9 to 24, 2021 in Paris).

Distinguished delegates,

I would like to acclaim the active contributions of Viet Nam Academy of Social Sciences, Vietnam's focal point in UNESCO's activities related to the field of social sciences in general and especially in the initiative to organize the today Conference with the theme on "Bioethics in the era of Artificial Intelligence (AI)". We highly appreciate the contributions of the Vietnam Academy of Social Sciences to the UNESCO's Draft Recommendation on Ethics of Artificial Intelligence. This not only shows Vietnam's responsibility for an important world issue, but also integrates Vietnam's interests in the context that Vietnam will need to improve her AI capabilities and national production to catch up with the world's development.

I believe that the Conference will be an opportunity to look back at the great changes in society brought about by the Industrial Revolution 4.0 and the development of Artificial Intelligence, to better identify challenges in terms of ethics of Artificial Intelligence, to learn

from international experiences, then to propose solutions to both improve technical skills and prepare for higher level of AI development in Vietnam.

I would also like to recommend that in the coming time, Vietnam will further strengthen the cooperation with UNESCO in this field in order to promote awareness and education on Artificial Intelligence, to develop appropriate policies, and handle challenges posed by artificial intelligence, especially from an ethical and humanistic point of view.

I wish the Conference the great success. May all the distinguished delegates have good health.

Thank you very much./.

Phần 1: Trí tuệ Nhân tạo
và các khía cạnh đạo đức

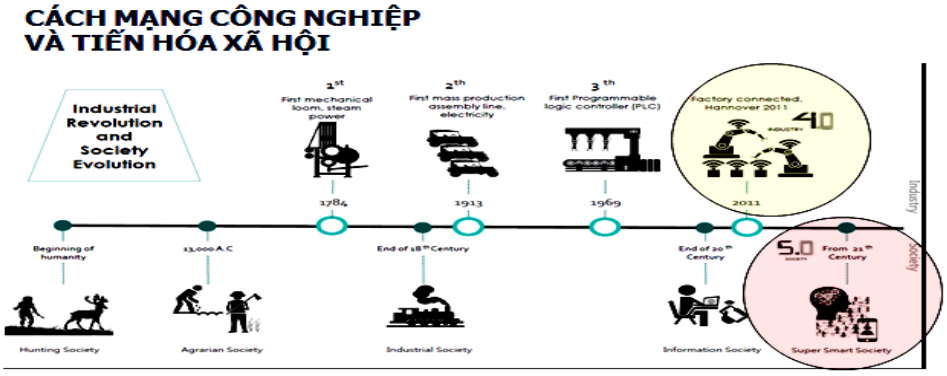
Part 1: AI and Ethical Aspects

ĐẠO ĐỨC TRONG QUẢN TRỊ, SỬ DỤNG TRÍ TUỆ NHÂN TẠO AI (*ARTIFICIAL INTERLIGENT*) TRƯỚC TÁC ĐỘNG CỦA CUỘC CÁCH MẠNG CÔNG NGHIỆP 4.0

Thượng tướng PGS.TS. Nguyễn Văn Thành
Phó Chủ tịch chuyên trách Hội đồng Lý luận Trung ương
ThS. Cao Thị Phương

1. Lịch sử loài người đã và đang trải qua các cuộc cách mạng công nghiệp: ***CMCN lần thứ nhất vào năm 1784*** khởi nguồn từ nước Scotland, đặc trưng là cơ khí hóa với máy chạy bằng thủy lực và phát minh ra động cơ hơi nước (*phát minh này của James Watt công bố năm 1775*) - Kỷ nguyên sản xuất cơ khí. Loài người chuyển từ văn minh nông nghiệp sang văn minh công nghiệp.

CMCN lần thứ hai: Từ năm 1871 - 1914 đặc trưng là động cơ điện, vận tải, hóa học, sản xuất thép và sản xuất trên cơ sở điện cơ khí và sang giai đoạn tự động hóa và khởi nguồn từ Mỹ. ***CMCN lần thứ ba*: *Từ năm 1969,*** với sự ra đời của công nghệ thông tin (CNTT), sử dụng điện tử và CNTT để tự động hóa sản xuất. Được xúc tác bởi chất bán dẫn, siêu máy tính, laptop (1970 và 1980), Internet (thập niên 1990) trung tâm và khởi nguồn từ Mỹ.



Cuộc cách mạng công nghiệp lần thứ tư **(*The Fourth Industrial Revolution - FIR***) đã và đang diễn ra một cách nhanh chóng, ***điểm "đòn bẩy" là***: ***AI*** (trí tuệ nhân tạo); công nghệ in ***3D***, công nghệ sinh học, CN vật liệu mới, CN tự động hóa, Robot, công nghệ kết nối vạn vật (***IoT***) và Internet các dịch vụ (IoS). *Đặc trưng của Cách mạng Công nghiệp 4.0 là* các hệ thống sản xuất thực - ảo (*Cyber-Physical Systems*-CPS), lần đầu tiên được *Dr.Jame Truchat,* Giám đốc điều hành của National Instrument đưa ra vào năm 2006, trong đó thiết bị thông minh làm việc với nhau qua mạng không dây hoặc qua "*đám mây*". Thuật ngữ "*Industrie 4.0*", bắt đầu từ dự án trong chiến lược CNC của chính phủ Đức, thuật ngữ được sử dụng lần đầu vào năm 2011 tại Hội chợ Hannover; chính thức nhận diện khái niệm, nội hàm tại Diễn đàn Kinh tế Thế giới (WEF) lần thứ 46, ngày 20/1/2016.

Quy mô tốc độ phát triển chưa từng có trong lịch sử, tốc độ phát triển cấp số nhân, tác động to lớn về kinh tế và môi trường sinh thái. Tiêu chí tốc độ lan truyền của công nghệ được sử dụng đạt ngưỡng 50 triệu người (***điện thoại cần 75 năm, radio cần 38 năm, Tivi cần 13 năm, internet chỉ cần 4 năm, Facebook cần 3,5 năm***). Về kinh tế tác động đến tiêu dùng, sản xuất, năng xuất và giá cả. Bản đồ kinh tế thế giới, bản đồ sức mạnh của các doanh nghiệp cũng đang được vẽ lại. ***Bốn Siêu cường công nghệ đinh hình tương lai phát triển thế giới****:*

(1) Công nghệ di động: Thiết bị 4 rộng khắp: (i) cảm biến mọi nơi, (ii) kết nối mọi nơi; (iii) dữ liệu mọi nơi, (iv) dịch vụ mọi nơi. đạt lượng thuê bao sử dụng chưa từng có: Cung cấp khả năng tiếp cận chưa từng thấy, kết nối mọi người di chuyển bất kể họ đang ở đâu trên thế giới.

(2) Điện toán đám mây: đạt quy mô không thể tưởng tượng so với trước đây. Cung cấp công suất ở quy mô chưa từng có trước đây,

cho phép các tổ chức thêm hoặc loại bỏ các thành phần khác nhau vào cơ sở hạ tầng của họ một cách nhanh chóng và khi cần thiết.

(3) Internet vạn vật (IoT): thông qua hệ thống các công nghệ kết nối và các nền tảng khác nhau. Là một phần tích hợp của Internet tương lai bao gồm các phát triển của Internet và mạng hiện tại và tiến hóa với cơ sở hạ tầng mạng động toàn cầu dựa trên giao thức liên kết và tương tác "***vạn vật***" hữu hình và ảo sử dụng các giao diện thông minh Smart được tích hợp vào mạng thông tin một cách thông suốt Kết nối các thế giới kỹ thuật số và thế giới vật lý: Kết nối các thế giới vật lý và kỹ thuật số, đưa công nghệ vào mọi khía cạnh của sự tiến bộ của con người.

Các siêu kết nối thông qua ***IoT*** và điện toán đám mây sẽ cho phép truyền thông tin và giao tiếp phổ quát toàn cầu và gần như tức thời, xuất hiện những cách thức cung cấp hàng hóa dịch vụ trước đây là điều không tưởng. Viễn cảnh ***IoT***: *(i)* ***IoE****: Internet năng lượng; (ii)* ***IoS****: Internet dịch vụ; (iii)* ***IoM****: Internet Truyền thông; (iv)* ***IoP****: Internet con người, (v)* ***IoT****: Internet vạn vật.* ***IoT*** sẽ thay đổi cả phương thức hoạt động của một nền kinh tế tạo mô hình kinh doanh mới .

(4) **Trí thuệ nhân tạo AI (Artificial Intelligence):**

AI Là hệ thống máy tính thể hiện hành vi với các chỉ số trí thông minh (***Smart***). Phân loại: *(i)* Hệ thống có tư duy như con người (*mạng lưới thần kinh và cấu trúc nhận thức); (ii)* Hệ thống hành động như con người (*suy luận tự động); (iii)* Hệ thống tư duy hợp lý (*suy luận, tối ưu hóa*); *(iv)* Hệ thống hành động hợp lý *(phần mềm thông minh, Robot đạt mục tiêu thông qua nhận thức xây dựng kế hoạch; giao tiếp, quyết định và hành động).* Trí tuệ nhân tạo ở khắp mọi nơi: Học hỏi sâu giúp chúng ta khai thác một lượng lớn dữ liệu trong thời gian thực và sử dụng những hiểu biết đó để đẩy mạnh việc khám phá học thuật và tạo ra các mô hình kinh doanh hoàn toàn mới.

Theo dòng chảy của cuộc cách mạng 4.0, trí tuệ nhân tạo ngày càng được phổ biến và ứng dụng rộng rãi trong mọi lĩnh vực của đời sống xã hội, mặc dù được John McCarthy - nhà khoa học máy tính người Mỹ đề cập lần đầu tiên vào những năm 1956 nhưng đến ngày nay thuật ngữ trí tuệ nhân tạo mới thực sự được biết đến rộng rãi và được các "***ông lớn***" của làng công nghệ chạy đua phát triển.

2. Những tác động của cuộc CMCN lần thứ tư (4.0):

Cạnh tranh việc làm, *Theo Diễn đàn Kinh tế thế giới*: Đến năm 2025, người sử dụng lao động sẽ phân chia công việc giữa con người và máy móc một cách gần như bình đẳng 50-50. Tự động hóa nhanh hơn dự kiến, sẽ thay thế 85 triệu việc làm trong năm 2021. Công nghệ sẽ thay đổi nhiệm vụ, công việc và kỹ năng vào năm 2025. Khoảng 43% doanh nghiệp được khảo sát cho biết rằng họ sẽ giảm lực lượng lao động vì áp dụng công nghệ, 34% có kế hoạch tuyển dụng lực lượng lao động mới do tích hợp công nghệ.

Các ngành nghề mới nổi phản ánh nhu cầu lớn hơn về việc làm trong nền kinh tế xanh; vai trò đi đầu trong nền kinh tế dữ liệu và AI; và các vai trò mới trong kỹ thuật, điện toán đám mây và phát triển sản phẩm. Các công việc đang phát triển nêu bật tầm quan trọng liên tục của sự tương tác giữa con người trong nền kinh tế mới thông qua các vai trò trong nền kinh tế chăm sóc; tiếp thị, bán hàng và sản xuất nội dung.

Đến năm 2025, tư duy phân tích, sáng tạo và linh hoạt sẽ là một trong những kỹ năng được tìm kiếm nhiều nhất. Các nhà tuyển dụng coi tư duy phản biện, phân tích và giải quyết vấn đề ngày càng có tầm quan trọng trong những năm tới. Điểm mới nổi bật trong năm nay là các kỹ năng quản lý bản thân, chẳng hạn như học tập tích cực, khả năng phục hồi, chịu đựng căng thẳng và linh hoạt.

Tác động đối với doanh nghiệp: có bốn tác động chính:*(i)* về kỳ vọng của khách hàng, *(ii)* về nâng cao chất lượng sản phẩm, dịch

vụ, *(iii)* về đổi mới hợp tác và *(iv)* về các hình thức tổ chức sản xuất và sáng tạo. *Tăng sức ép gia tăng dòng nhập cư*: *Các doanh nghiệp cạnh tranh* nhất sẽ tập trung vào việc nâng cấp kỹ năng cho công nhân của họ. *Làm việc từ xa*: Khoảng 84% người sử dụng lao động đang nhanh chóng số hóa các quy trình làm việc, mở rộng đáng kể hình thức làm việc từ xa. Có khả năng 44% lực lượng lao động sẽ di chuyển sang làm việc từ xa. Tuy nhiên, 78% các nhà doanh nghiệp cho rằng sẽ có một số tác động tiêu cực đến năng suất của người lao động và nhiều doanh nghiệp đang thực hiện các bước giúp nhân viên của họ thích nghi dần với làm việc từ xa.

Tác động lên chính quyền: xuất hiện mô hình chính phủ trí tuệ nhân tạo (AI-Government); Tiến đến 100 năm Liên hợp quốc (2045) , Bàn tròn UN 2045 do Ramu Damodaran Tổng giám đốc Academic Impact của Liên Hợp Quốc, Tổng Biên tập Tạp chí Biên Niên sử Liên hợp quốc và cha đẻ Internet Vint Cerf thảo luận về mô hình mới con người hướng tới: "*kinh tế trọng tâm vì con người, hệ sinh thái internet và trí tuệ nhân tạo mới cho công việc và cuộc sống*". Nội hàm thành phố Xã hội Trí tuệ nhân tạo (AIWS City) trong mô hình *Thành phố thông minh (Smart city).* Khả năng thích ứng với công nghệ mới của các cơ quan công quyền sẽ quyết định năng lực quản lý của họ. Chính quyền và các cơ quan quản lý của mình sẽ cần phối hợp chặt chẽ với doanh nghiệp và các hội, hiệp hội và người dân.

Tác động đối với con người *(tồn tại con người):* về sinh học, xã hội, và tinh thần đặt ra những yêu cầu mới.

Tác động lên báo chí, truyền thông: Công nghệ viết tin tự động: Phần mềm tự động viết tin tức tài chính đã được hãng thông tấn AP đưa vào sử dụng từ năm 2014 với tốc độ lên tới 2.000 bản tin/giây và tiếp tục được mở rộng sang lĩnh vực khác; trí tuệ nhân tạo AI tự sáng tạo nội dung truyền thông.

Tác động đến an ninh, an toàn*:* ngày **01/11/2002,** cụm từ *"An ninh phi truyền thống"* chính thức xuất hiện trong "*Tuyên bố chung*

ASEAN - Trung Quốc về hợp tác trên lĩnh vực an ninh phi truyền thống". Theo quan điểm của UN (Liên hợp quốc) gồm ***10*** *mối đe dọa* (khủng bố, ma túy, hải tặc, rửa tiền, tin tặc, thảm họa môi trường, dịch bệnh, buôn bán người, di cư bất hợp pháp và cực đoan dân tộc, tôn giáo).

Lợi dụng kết nối internet để thực hiện tội phạm, thách thức trong quản trị không gian mạng và không gian (ảo): Drone, Flycam (vật thể không người lái). Xuất hiện những đột biến trên không gian ảo, kinh doanh đa cấp ảo, bong bóng đầu tư vào các loại tiền ảo, bitcoin gây tiềm ẩn khủng hoảng lan truyền tới thị trường tài chính toàn cầu mà chưa có các biện pháp kiểm soát hiệu quả.

3. Sử dụng công cụ SWOT (các điểm mạnh, yếu, cơ hội và thách thức) về trí tuệ nhân tạo AI hiện nay

(1) Xu hướng về phát triển AI hiện nay và các năm tiếp theo

Trợ lý ảo Chatbot - phần mềm AI giúp tương tác/nói chuyện tự động với từng khách hàng 24 giờ/7 ngày (64% người dùng lựa chọn nhắn tin hơn là gọi điện hoặc email). Trợ lý ảo tương tác giúp tiết kiệm 8 tỷ USD cho các DN vào năm 2022. 25% công ty sẽ sử dụng trợ lý ảo chăm sóc dịch vụ khách hàng trong tương lai gần.

Thời gian đào tạo trí tuệ nhân tạo AI giảm xuống, máy học sâu hơn

Tăng tốc độ xe ô tô không người lái (đến 2030, sẽ đạt 40%).

Máy học (*automated machine learning - AutoML*) và trí tuệ nhân tạo đưa vào sản xuất đại trà (khoa học dữ liệu trở thành nghề hot nhất), việc áp dụng các công cụ AI là xu hướng quan trọng nhất trong các năm tiếp theo.

Trí tuệ nhân tạo tự lập trình các ứng dụng, Sự kết hợp của các mạng thần kinh nhân tạo rộng khắp; Phát triển các hệ thống AI chuyên dùng

AI thúc đẩy chuyển đổi sang nền kinh tế số; thách thức xuất hiện do thiếu hụt nguồn nhân lực có kỹ năng AI.

AI nhận dạng khuôn mặt. Chíp AI tăng cường trí tuệ nhân tạo. AI nhận dạng ngữ cảnh, nhận dạng các khái niệm trừu tượng. AI học nhận thức, tình cảm.

(2) Phân tích SWOT

❖ **Điểm mạnh:**

AI (Trí tuệ nhân tạo) có khả năng tự thích nghi, tự học và tự phát triển, tự đưa ra các lập luận để giải quyết vấn đề, có thể giao tiếp như người… tất cả là do AI được cài một cơ sở dữ liệu lớn, được lập trình trên cơ sở dữ liệu đó và tái lập trình trên cơ sở dữ liệu mới sinh ra. Cứ như vậy cấu trúc của AI luôn luôn thay đổi và thích nghi trong điều kiện và hoàn cảnh mới.

Đặc trưng của công nghệ AI là năng lực "tự học" của máy tính, do đó có thể tự phán đoán, phân tích trước các dữ liệu mới mà không cần sự hỗ trợ của con người, đồng thời có khả năng xử lý dữ liệu với số lượng rất lớn và tốc độ cao.

AI mang lại Chất lượng sống tốt hơn với các chỉ số hài lòng của công dân; *(i)* Tốc độ xử lý, ra quyết định nhanh; *(ii)* Phát triển quy mô lớn, mang lại hiệu quả rõ rệt; *(iii)* Độ chính xác cao.

AI đã xuất hiện trong nhiều ngành, từ cung cấp dịch vụ mua sắm ảo và ngân hàng trực tuyến đến giảm chi phí đầu tư trong sản xuất và hợp lý hóa chẩn đoán trong chăm sóc sức khỏe. AI đã thúc đẩy hầu hết các ngành công nghiệp tiến lên và thay đổi cuộc sống của nhiều người.

❖ **Nguy cơ thách thức**

Bản chất của trí tuệ nhân tạo vẫn do con người làm ra, họ xây dựng các thuật toán, lập trình bằng các công cụ phần mềm công

nghệ thông tin, giúp các máy tính có thể tự động xử lý các hành vi thông minh như con người do đó vấn đề đạo đức trong trong quản trị AI ứng phó với các thách thức:.

(1) Mất việc làm; *(2)* Nhận thức về AI còn hạn chế; *(3)* Chưa có hành lang pháp lý hoặc công ước quốc tế; *(4)* Nguy cơ về đạo đức; quan ngại về tính riêng tư, các vấn đề nhạy cảm; *(5)* Mang tính chủ quan của người thiết kế lập trình. *(6)* An ninh mạng, nguy cơ bị hack; *(7)* Máy móc, AI vi phạm thì xử lý ai?; *(8)* Khoảng cách về công nghệ đối với một số nước; *(9)* Mất kiểm soát với AI: AI kiểm soát thế giới?

AI là công nghệ sử dụng đến kỹ thuật số có khả năng thực hiện những nhiệm vụ mà bình thường phải cần tới trí thông minh của con người, được xem là phổ biến nhất: *(1)* Phức tạp, cần nhân lực có kỹ năng cao; *(2)* Vẫn là máy, vô thức với con người; *(3)* Tính thiên vị hoặc sai lệch (AI học từ các mẫu trong dữ liệu, và có thể sử dụng dữ liệu sai do đó đưa ra các giả định sai, không đại diện); *(4)* Có tiềm năng thông minh vượt khả năng con người; *(5)* Tính trách nhiệm xã hội; *(6)* Chính quyền thường chậm tiếp thu tiếp cận với AI; *(7)* Yêu cầu ứng dụng tư duy hệ thống (để tác động quyết điểm đòn bẩy); *(8)* Yêu cầu đầu tư vốn lớn; *(9)* Cần có cơ sở dữ liệu đủ lớn cho máy học; *(10)* Cần có cơ sở hạ tầng thiết bị thông tin, hệ thống cảm biến, hạ tầng kết nối băng rộng; *(11)* Vấn đề máy học cảm xúc?

❖ **Cơ hội:**

Trí tuệ nhân tạo hiện nay tiếp cận hầu hết mọi ngóc ngách của cuộc sống, từ thể chất, tinh thần đến trạng thái cảm xúc. Các hệ thống AI được sử dụng để xác định ai nhận được các dịch vụ công, quyết định AI có cơ hội được tuyển dụng cho một công việc và tất nhiên chúng ảnh hưởng đến những thông tin mà mọi người nhìn thấy và có thể chia sẻ trực tuyến.

Xây dựng và phát triển chính phủ trí tuệ nhân tạo quản trị trong xã hội trí tuệ nhân tạo.

Cuộc đua toàn cầu về phát triển AI bắt đầu từ 2017; tiềm năng ứng dụng AI rất nhanh và rộng rãi;

Nhu cầu rất lớn trên mọi lĩnh vực: truyền thông, quốc phòng, an ninh, kinh tế, quản lý xã hội, khoa học, y tế, bảo vệ môi trường, năng lượng, tài chính, thương mại, giám sát quản lý cơ sở hạ tầng, dự báo, xe tự lái thông minh cho người khuyết tật…

Giảm căng thẳng cho nhân viên vì AI sẽ thực hiện các nhiệm vụ tẻ nhạt, lặp lại thay con người;

Sự kết hợp giữa AI và các dạng công nghệ mới; thị trường AI tăng trưởng rất nhanh; thay thế con người trong quản lý nhiều lĩnh vực, giảm nhu cầu sử dụng nhân lực, nâng cao năng suất lao động.

4. Đạo đức trong quản trị, sử dụng trí tuệ nhân tạo

Đạo đức AI: Tính trách nhiệm và minh bạch của AI. Trong tương lai gần, AI sẽ: *(1)* Thiết kế phác đồ điều trị y tế cho bệnh nhân; *(2)* Tự động tìm kiếm, lựa chọn sơ yếu lý lịch các ứng viên tìm việc và ra các quyết định tuyển dụng nhân sự cho các vị trí việc làm. *(3)* Tự động lái xe ô tô; *(4)* AI sẽ tham gia vào rất nhiều các hoạt động mang tính rủi ro cao khác.

Vấn đề đặt ra: Trong khuôn khổ khóa họp lần thứ 48 Hội đồng Nhân quyền Liên hợp quốc (LHQ) kéo dài 4 tuần (13/9 - 11/10/2021), Cao ủy Nhân quyền LHQ Michelle Bachelet ngày 15/9 nhấn mạnh nhu cầu khẩn cấp về việc cấm mua bán và sử dụng các hệ thống trí tuệ nhân tạo (AI) có nguy cơ nghiêm trọng đối với nhân quyền cho đến khi các biện pháp bảo vệ thích hợp được thực thi. Bà Bachelet cũng kêu gọi cấm các ứng dụng AI không được sử dụng nếu không tuân thủ luật nhân quyền quốc tế.

“Trí tuệ nhân tạo có thể là một động lực tốt, giúp xã hội vượt qua một số thách thức lớn của thời đại chúng ta. Nhưng các công nghệ AI có thể có những tác động tiêu cực, thậm chí là thảm khốc

nếu chúng được sử dụng mà không quan tâm đầy đủ đến việc chúng ảnh hưởng đến nhân quyền của con người như thế nào".

Ai là chủ nhân phải chịu trách nhiệm khi tai nạn do các thiết kế, các quyết định do AI (Trí tuệ nhân tạo), các thiết bị AI (Trí tuệ nhân tạo) gây ra? Vấn đề sử dụng nguồn dữ liệu không minh bạch, xâm phạm quyền riêng tư, việc bán các dữ liệu cá nhân người dùng khi không được phép?

Rủi ro đối với quyền con người càng cao, thì các yêu cầu pháp lý đối với việc sử dụng công nghệ AI càng phải chặt chẽ hơn. Nhưng vì có thể mất thời gian trước khi các rủi ro có thể được đánh giá và giải quyết, các quốc gia cần đặt ra chính sách về việc sử dụng công nghệ tiềm ẩn rủi ro cao.

Với tốc độ phát triển nhanh chóng và liên tục của AI (Trí tuệ nhân tạo), việc đưa ra quy định về trách nhiệm giải trình cách dữ liệu được thu thập, lưu trữ, chia sẻ và sử dụng là một trong những câu hỏi cấp bách nhất về nhân quyền mà thế giới phải đối mặt. Nguy cơ phân biệt đối xử liên quan đến các quyết định do AI điều khiển - những quyết định có thể thay đổi, xác định hoặc gây thiệt hại cho cuộc sống con người - đều có thể trở nên hiện thực.

5. Sử dụng AI (trí tuệ nhân tạo) trong quy hoạch, thiết kế và quản trị thành phố thông minh theo bộ chỉ số quốc tế ISO 37122:2019 (E)

AI (Trí tuệ nhân tạo), điều khiển mô hình hệ tích hợp thực, ảo trong thành phố thông minh (***Smart City***).

AI đưa ra các cơ chế, chính sách hỗ trợ các doanh nghiệp (sức sống của nền kinh tế) trong Smart City.

AI ứng dụng trong mô hình chính quyền thông minh, Chính phủ trí tuệ nhân tạo AI Government

AI tích hợp điều hành Mô hình di chuyển thông minh

AI tích hợp Mô hình quản lý cảng thông minh

AI quản lý và giám sát môi trường thông minh (Môi trường đất, nước, khong khí, cảng biển, dải vùng bờ và các lưu vực sông)

AI giám sát Cư dân thông minh

AI kiểm soát các mối nguy cơ an ninh phi truyền thống

AI hỗ trợ đào tạo nguồn nhân lực mới thích ứng và chiến lược tạo nguồn nhân lực thích ứng, bồi dưỡng đào tạo, đào tạo lại tạo nguồn nhân lực đáp ứng yêu cầu.

AI giúp kiểm soát thành phố trở lên an toàn hơn, lành mạnh hơn, hiệu quả hơn ➔ AI hỗ trợ xây dựng mô hình Smart City đạt các chỉ tiêu An ninh, An sinh, An toàn.

AI quản lý quy hoạch và xây dựng thành phố trên quan điểm coi đây là một hệ sinh thái (***digital ecosystem***) phức hợp về kinhh tế - xã hội - công nghệ, sử dụng Big data và IoT trong điều hành.

AI và robot thúc đẩy tổ chức sản xuất năng xuất, chất lượng hiệu quả

6. Một số yêu cầu để xây dựng và quản trị thành công Thành phố thông minh gắn kết nội hàm Thành phố xã hội trí tuệ nhân tạo (AIWS City):

AI (Trí tuệ nhân tạo): nền tảng của thành phố thông minh và mô thình Thành phố xã hội trí tuệ nhân tạo (***AIWS***); Xây dựng mạng lưới trí tuệ nhân tạo AI rộng khắp: AI of Things (***AIOT***), mọi lĩnh vực quản lý đô thị phục vụ cư dân đều dựa trên nền tảng trí tuệ nhân tạo liên kết mọi lúc, mọi nơi theo thời gian thực.

AIWS (*Thành phố xã hội trí tuệ nhân tạo*) phải được thiết kế, quản trị đảm bảo các chỉ số An ninh - An sinh - An toàn, cung cấp các giải pháp bằng AI và xác nhận dấu ấn điện tử bằng ***block chain***.

AIWS (*Thành phố xã hội trí tuệ nhân tạo*) phải xây dựng bộ tiêu chuẩn ISO về đạo đức và văn minh đô thị. Đổi mới sáng tạo của dân

cư AIWS được tích hợp, liên kết và chia sẻ phục vụ doanh nghiệp và nâng cao chất lượng dịch vụ đô thị.

AIWS (*Thành phố xã hội trí tuệ nhân tạo*) gắn liền với Chính quyền số, Chính phủ Trí tuệ nhân tạo (AI Government). Không gian của AIWS không bị giới hạn bởi không gian địa lý và có thể chia sẻ, kết nối các nền văn hóa để xây dựng xã hội hài hòa.

AIWS (*Thành phố xã hội trí tuệ nhân tạo*) phải tăng cường kinh tế chia sẻ, nâng cao khả năng thanh toán điện tử, sử dụng tiền kỹ thuật số, block chain giúp tăng trưởng GDP, giám sát thu ngân sách, chống rửa tiền, giảm tội phạm.

AIWS (*Thành phố xã hội trí tuệ nhân tạo*) đưa ra các giải pháp kiểm soát khủng hoảng thế giới ảo và hạn chế những tác động tiêu cực lên thế giới thực.

AIWS (*Thành phố xã hội trí tuệ nhân tạo*) tăng cường nền kinh tế chăm sóc, tiếp thị, bán hàng và sản xuất nội dung;

AIWS (*Thành phố xã hội trí tuệ nhân tạo*) tăng tính minh bạch, dễ kiểm tra, giám sát mọi quyết định, mọi hoạt động của Chính quyền thành phố, các cơ quan chức năng bảo đảm tính hiệu quả.

AIWS (*Thành phố xã hội trí tuệ nhân tạo*) thúc đẩy phát triển các kỹ năng quản lý bản thân, chẳng hạn như học tập tích cực, khả năng phục hồi, chịu đựng căng thẳng và linh hoạt.

7. Gợi ý quy trình quy hoạch, xây dựng mô hình Thành phố thông minh gắn kết với nội hàm Thành phố xã hội trí tuệ nhân tạo (AIWS)

Bước 1. Quy hoạch không gian ảo, không gian kết nối thực - ảo

Bước 2. Xây dựng bộ chỉ số đạo đức và văn minh đô thị AIWS trên cơ sở kết nối các chỉ số ISO 37120:2018 và ISO 37122:2019 Smart City.

Bước 3. Thỏa thuận Khế ước về xã hội Trí tuệ nhân tạo (AIWS).

Bước 4. Xây dựng hành lang pháp lý để công nhận mô hình AIWS hợp pháp rộng rãi.

Bước 5. Khai thác nền tảng block chain và kết nối trí tuệ nhân tạo hiện có và cung cấp các truy cập, khai thác tới các cấp chính quyền đô thị - từng hộ dân, người dân.

Bước 6. Kết nối các Nhà lãnh đạo thế giới, các Thị trưởng, các Chuyên gia đô thị và các Nhà khoa học đổi mới, sáng tạo để tạo ngân hàng dữ liệu cho AIWS và máy học **learning machine.**

Bước 7. Nâng cao năng lực cạnh tranh của thành phố, khả năng tự phục hồi trước thiên tai, thảm họa và khủng hoảng kinh tế.

Bước 8. Vận hành thử các trung tâm Trí tuệ nhân tạo ở tất cả các cấp chính quyền đô thị - kết nối người dân. Hiệu chỉnh và đưa AIWS vào phục vụ người dân của Thành phố thông minh, hướng tới thành phố đáng sống, thành phố có giá trị.

8. Các căn cứ xác định nội hàm đạo đức trong quản trị trí tuệ nhân tạo AI trong thành phố thông minh gắn kết với nội hàm Thành phố xã hội trí tuệ nhân tạo (AIWS)

Thiết kế AI gắn với đạo đức xã hội.

Thiết kế AI bảo đảm lợi ích của các bên liên quan.

Thiết kế AI bảo đảm sự tham gia & kết nối: đây là một trụ cột của cuộc cách mạng công nghiệp 4.0 và một trong 15 sản phẩm là thành phố thông minh (Smart City).

Thiết kế AI bảo đảm liên kết khu vực với trọng tâm là cộng đồng**:** Tăng cường kết nối và hợp tác trong khu vực như một thể thống nhất, thông qua chiến lược phát triển kinh tế vùng.

Thiết kế AI tạo động lực chiến lược và các sáng kiến nền tảng**:** Chiến lược tốt cần phải xác định các sáng kiến để có “thành công sớm" qua đó tạo động lực thúc đẩy.

Thiết kế AI rõ ràng về mục đích sử dụng và lợi ích.

Thiết kế AI trên nền tảng xây dựng tư duy chiến lược**:** Tốc độ phát triển công nghệ rất nhanh, cần có khung thời gian cho các cấp độ chiến lược. Tốn nhiều thời gian vào chi tiết kỹ thuật trong Chiến lược sẽ làm chiến lược đó sớm lạc hậu.

Thiết kế AI cần rút ra từ các bài học.Thiết kế AI gắn kết quy hoạch đô thị và quy hoạch liên vùng trong Smart City***.***

Thiết kế AI phải đối chiếu với các chỉ số đánh giá kết quả (Indicators): Một số tổ chức đang xây dựng các chỉ số này như ISO (Năm 2018, PGS Nguyễn Văn Thành đã hoàn thành Dự thảo lần đầu ISO Smart City cho WCCD), ITU thông qua Liên hợp quốc. Chú ý các chỉ số này phải có ý nghĩa với dân cư và doanh nghiệp là những người sẽ lựa chọn nơi đầu tư và sinh sống.

Thiết kế AI phải gắn với tạo ra văn hóa thông minh lâu dài (Smart Culture): Sự tham gia của tất cả các bên liên quan, cùng chia sẻ những thách thức và thành công là mấu chốt để tạo ra văn hóa thông minh (Smart City Culture)

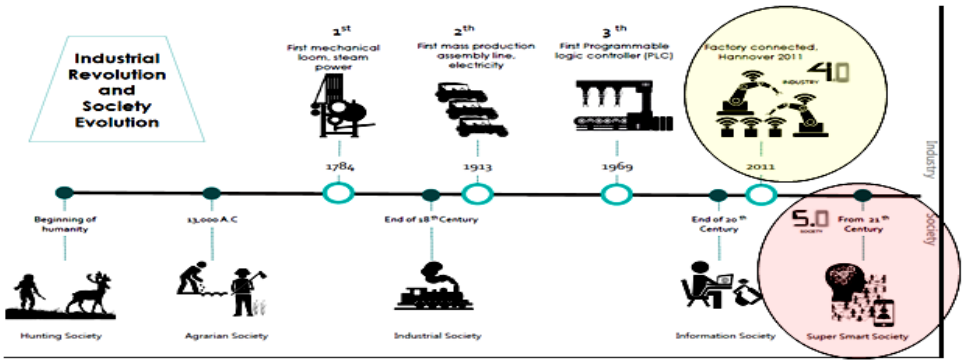
Thiết kế AI phải bảo đảm tính kiểm soát và các giới hạn cho phép.

ETHICS IN ARTIFICIAL INTELLIGENCE MANAGEMENT AND USAGE UNDER THE IMPACTS OF INDUSTRIAL REVOLUTION 4.0

Senior Lieutenant General Assoc.Prof. Dr. Nguyen Van Thanh
Vice Chairman of the Central Theoretical Council
MA. Cao Thi Phuong

1. Human history has been going through industrial revolutions: ***The First Industrial Revolution in 1784*** originated in Scotland, characterized by mechanization with hydraulic machines and the invention of the steam engine (*this invention by James Watt published in 1775*) - The era of mechanical production. Humanity changed from the agricultural civilization to the industrial civilization.

The Second Industrial Revolution: From 1871 to 1914 characterized by electric propulsion, transportation, chemistry, steel-making and electromechanical-based production, moving to the automation phase and starting in the US***. The Third Industrial Revolution: Started in 1969***, with the advent of information technology (IT), using electronics and IT to automate production. The revolution was also powered by semiconductors, supercomputers, laptops (1970s and 1980s), the Internet (1990s) that centered and originated in the US.



The fourth industrial revolution **(FIR**) has been happening rapidly, ***the "leverage" point is***: **AI** (artificial intelligence); **3D** printing technology, biotechnology, new materials technology, automation industry, robotics, technology of Internet of Things (**IoT**) and Internet of services (**IoS**). The special features of the industrial revolution 4.0 are Cyber-Physical Systems (CPS), that was first introduced by Dr. James Truchat, Executive Director of National Instrument in 2006, in which intelligent equipments working together over a wireless network or through the “cloud”. The term “Industrie 4.0” derived from the project in the German government's CNC strategy, then it was first used in 2011 at the Hannover Fair, and officially identified and connotated at the 46th World Economic Forum (WEF), January 20, 2016.

The scale of the growth rate has unprecedented in history, it is an exponential growth rate bringing about huge economic and ecological impacts. The target for the spreading speed of technology being used is to reach the threshold of 50 million people (for telephone to be 75 years, for radio to be 38 years, for television to be13 years, for internet to be 4 years, for Facebook to be 3.5 years). In economic term, there have been impacts on consumption, production, productivity and prices. The world economic map and the power map of businesses are also being redrawn. For example, 3D printing technology was worth $3.1 billion/year, 35% higher compared to 2012; in 6 years after that it increased by an average of 32% a year and reached 21 billion USD in 2020. ***The Four Technological Superpowers have been shaping the future of world development***:

(1) Mobile technology: Devices with four things everywhere: (i) sensing everywhere, (ii) connecting anywhere; (iii) data anywhere, (iv) servicing anywhere. That has reached unprecedented number of subscribers; provided unprecedented accessibility that connects people on the go no matter where they are in the world.

(2) Cloud Computing: That has reached unimaginable scale compared to before; provided deliver capacity at a scale like never before; allowed organizations to add or remove disparate components to their infrastructure quickly and as needed.

(3) Internet of Things (IoT): It is done through a system of connection technologies and different platforms. Being an integrated part of the future Internet, it includes developments of the current Internet and networks and evolves with a dynamic global network infrastructure based on a physical and virtual "***things***" interoperability and interconnection protocol using smart interfaces integrated into the information network to connect with the digital and physical worlds: Connecting the physical and digital worlds, bringing technology into every edge of human progress.

Hyperconnections through ***IoT*** and cloud computing have enabled communication and information transmission universally and almost instantaneously, introduced ways of delivering goods and services that were previously unthinkable. ***IoT*** Outlook: (i) ***IoE***: Internet of Energy; (ii) ***IoS***: Internet of Services; (iii) ***IoM***: Internet of Media; (iv) ***IoP***: Internet of People, (v) ***IoT***: Internet of Things. ***IoT*** may also change the way an economy operates to create a new business model.

(4) Artificial Intelligence (AI)

AI is a computer system that exhibits behavior with smart indicators. That may be classified as: (i) Systems think like human (*neural networks and cognitive structures*); (ii) Systems act like human (*automatic inference*); (iii) Systems think rationally (*inference, optimization*); (iv) Systems act reasonaly (*smart software, robot achieve the set goals through awareness, planning, communication, decision-making and action*). Artificial intelligence is everywhere: Deep learning enables us to explore massive amounts of data in real

time and use those insights to drive academic discovery and create completely new business models.

Following the flow of the revolution 4.0, artificial intelligence is increasingly popular and widely applied in all areas of social life. Though it was mentioned for the first time in 1956 by John McCarthy - an American computer scientist, until today the term artificial intelligence is really widely known and being competitively developed by "***big guys***" of the technology village.

2. Impacts of the Fourth Industrial Revolution (4.0):

Job competition, *according to the World Economic Forum*: By 2025, employers will divide work between humans and machines on an almost equal 50-50 basis. Faster automation than expected would replace 85 million jobs in 2021. Technology will change tasks, jobs and skills by 2025. About 43% of surveyed businesses said they would reduce their workforce because of technology adoption, 34% planned to recruit new workforce due to technology integration.

Emerging occupations reflect a greater need for jobs in the green economy, the leading role in the data economy and AI, and new roles in engineering, cloud computing, and product development.

The growing jobs highlight the continuing importance of human interaction in the new economy through roles in the healthcare economy; marketing, sales and content production.

By 2025, analytical thinking, creativity and flexibility will be among the most in-demand skills. Employers will attach greater importance to critical thinking, problem analyzing and solving skills in the coming years. The self-management skills, such as active learning, resilience, stress tolerance, and flexibility, have become significant recently.

Impacts on business: There are four main impacts on: (i) customer expectations, (ii) product and service quality improvement, (iii) collaborative innovation, and (iv) forms of production organization and innovation. *With the rising pressure to increase immigration*: *The most competitive businesses* will focus on upgrading the skills of their workers. *Remote working*: About 84% of employers are rapidly digitizing their work processes, therefore, significantly expanding remote works. It is likely that 44% of the workforce will move to remote work. However, 78% of entrepreneurs think there will be some negative impacts on worker productivity, and many businesses are taking steps to help their employees adapt to working on distance.

Impacts on governments: There appears an artificial intelligence government model (AI-Government). Also towards the 100th anniversary of the United Nations (2045), at the United Nations 2045 Roundtable moderated by Ramu Damodaran, Chief of the United Nations Academic Impact , the Editor-in-Chief of the United Nations Chronicle and Father of the Internet Vint Cerf discussed a new model that people are looking forward to: "The People Centered Economy: The New AI and Internet Ecosystem for Work and Life". There are insights of artificial intelligence world society (AIWS City) in the smart city model. The ability of public authorities to adapt to new technology will determine their management capacity. Governments and their regulatory agencies will need to work closely with businesses, associations and citizens.

Impacts on people *(human existence):* In biological, social, and mental health aspects, there poses new demands.

Impacts on press and media: Automatic news-composing technology: The software that automatically writes financial news has been put into use by the AP news agency since 2014 with a speed of

up to 2,000 news/second and continues to be expanded to other fields; Artificial intelligence is able to create media content on its own.

***Impacts on security and safety**:* **On November 1, 2002**, the phrase "*non-traditional security*" officially appeared in the "*Joint Declaration of ASEAN and China on Cooperation in the Field of Non-Traditional Security Issues*". *According to the United Nations*, there are 10 threats (terrorism, drugs, piracy, money laundering, hackers, environmental disasters, epidemics, human trafficking, illegal migration and extreme ethnicity, religion).

Taking advantage of the internet connection to commit crimes and challenges in administration of cyberspace and (virtual) space: Drone, Flycam (unmanned object). There emerges mutations in the virtual space, virtual multi-level-marketing business, investment bubbles in virtual currencies, for example bitcoin causes potential crisis spreading to the global financial market without effective control measures.

3. Using the SWOT tool (Strengths, Weaknesses, Opportunities and Threats) on AI today

(1) Current trends in AI development and in the coming years

Virtual Assistant Chatbot - AI software that helps to interact/talk automatically with each customer 24 hours/7 days (64% of users choose texting over calling or emailing). Interaction by virtual assistants will save businesses 8 billion USD by 2022. 25% of companies might use virtual assistants to take care of customer service in 2020.

AI training time is reduced, deeper learning machine.

The speed of autonomous vehicles increases (by 2030, will reach 40%).

Automated machine learning (AutoML) and artificial intelligence are put into mass production (data science becomes the hottest profession), the application of AI tools is the most important trend in the coming years.

There are self-programming artificial intelligence, combination of extensive artificial neural networks and Development of specialized AI systems.

Artificial intelligence accelerates the transformation to the digital economy, creates challenges due to the shortage of human resources with AI skills.

Face recognition AI: AI chips enhance artificial intelligence. There are AIs with context recognition, recognition of abstract concepts, or cognitive and affective learning.

(2) SWOT analysis

❖ **Strengths:**

AI (Artificial Intelligence) is capable of self-adapting, self-learning and self-developing, self-making arguments to solve problems, being able to communicate like humans… all of these are because AI is installed on a big database, programmed on top of that databese and reprogrammed on the nascent database. With that process, the structure of AI is always changing and adapting to new conditions and circumstances.

The typical feature of AI technology is the “self-learning” ability of computers, therefore, they can judge and analyze new data in advance without human assistance, and at the same time have the ability to process a large quantity of data at high speed.

AI brings about a better life quality with higher indicators of citizens’ satisfaction: (i) High speed in processing and decision

making; (ii) Large-scale development, bringing tangible results; (iii) High accuracy.

AI is already present in many industries, from providing virtual shopping and online banking to reducing investment costs in manufacturing and streamlining diagnostics in healthcare services. AI has driven most industries forward and changed many people's lives.

❖ **Weaknesses and Threats**

The essence of artificial intelligence is still made by humans, who create algorithms, programme with information technology software tools, enable computers to automatically handle intelligent behaviors like humans, therefore, ethical issues in AI governance are facing challenges:.

(1) Job loss; (*2*) Awareness of AI is still limited; (*3*) There is no legal corridor or international convention; (*4*) Moral hazard, privacy concerns, sensitive issues; (*5*) Subjectivity of the programmers; (*6*) Cybersecurity, risks of being hacked; (*7*) Who will handle violations by machines and AI? (*8*) Technology gap for some countries; (*9*) Losing control with AI: AI may control the world?

AI is most commonly known as a technology using digital capable of performing tasks that normally require human intelligence: *(1)* Complexed works requiring highly skilled human resources; (*2*) Still being a machine, unconscious with humans; (*3*) Bias or inaccuracy (AI learns from patterns in data, therefore, it may misuse data thereby make false, non-representative assumptions); (*4*) Having the potential to be intelligent beyond human ability; (*5*) Social responsibility; (*6*) Government are usually slow in approaching AI; (*7*) Requiring the application of systematically thinking (to influence deciding the leverage point); (*8*) Requiring large capital investment; (*9*) Requring to have a database large enough for machine learning; (*10*) Requiring infrastructure for information

equipment, sensor systems, and broadband connection infrastructure; (*11*) Issues of emotional machine learning.

❖ **Opportunities:**

Artificial intelligence now approaches almost every angle of life, from physical, mental to emotional state. AI systems are used to determine who receives public services, determine AI's chances of being hired for a job, and of course they affect what information people may see and share online.

It is to build and develop the AI Government governing in an artificial intelligence society.

The global race on AI development begins in 2017; the potential for AI applications increases very fast and widespread.

There are very high demands in all fields: communication, defense, security, economy, social management, science, health, environmental protection, energy, finance, commerce, infrastructure management and supervision, forecasting, smart self-driving cars for people with disabilities, etc.

AI may reduce employees' stress because it will perform tedious, repetitive tasks instead of humans.

With the combination of AI and new forms of technology, and the fast growth rate of AI market, AI will replace people to manage many fields, reduce the need for human resources, and improve labor productivity.

4. Ethics in AI administration and usage

Ethics in AI: Accountability and transparency. In the near future, AI will: (*1*) Design medical treatment regimens for patients; (*2*) Automatically search, select resumes of job seekers and make recruitment decisions for positions. (*3*) Automatically drive cars; (*4*) AI will participate in many other high-risk activities.

The problem posed: Within the framework of the 48th session of the United Nations Human Rights Council (UN) lasting for 4 weeks (September 13 - October 11, 2021), UN High Commissioner for Human Rights Michelle Bachelet on September 15 emphasized the urgent need for prohibiting the sale and use of artificial intelligence (AI) systems as it has posed a serious risk to human rights until appropriate safeguards are in place. Ms. Bachelet also called for a ban on AI applications if they do not comply with international human rights law.

"Artificial intelligence can be a force for good, helping societies to overcome some of the great challenges of our time. But AI technologies can have negative, even catastrophic, effects if they are used without sufficient regard to how they affect people's human rights."

Who will take the responsibility when accidents are caused by designs, decisions made by AI? How to deal with the problems of using non-transparent data sources, invasion of privacy, the sale of personal data of users without permission?

The higher the risk to human rights, the tighter legal requirements for the use of AI technology must be. But because it may take time before risks can be assessed and addressed, countries need to set policies on usage of the potentially high-risk technology.

Given the rapid and continuing pace of AI development, to set up accountability for how data is collected, stored, shared and used is one of the most pressing human rights questions that the world has to face. The risk of discrimination associated with AI-driven decisions - decisions that can change, define, or damage human lives - can become real.

5. Using AI in Smart City planning, design and management by ISO 37122:2019(E)

AI to control models of real and virtual integrated system in Smart City.

AI to produce mechanisms and policies to support businesses (the vitality of the economy) in Smart City.

AI to be applied in Smart Government, AI Government models.

AI to be integrated to operate smart mobility model.

AI to be integrated in Smart Port Management Model.

AI to be smart environment manager and monitor (Environment of land, water, air, seaport, coastal strip and river basins).

AI to be Smart Resident Monitor.

AI to control non-traditional security threats.

AI to support training of new and adaptive human resources and strategies for creating adaptive human resources, aditional training, retraining to build human resources to meet requirements.

AI to help controling city safer, healthier, more efficient ➔ AI to support building a Smart City model that reaches Security, Wellbeing and Safety criteria.

AI to involve in mânging the city plan and construction from the perspective of a complex digital ecosystem of economy-society-technology, using Big data and IoT in operation.

AI and robots to promote production, productivity, quality and efficiency.

6. Some requirements for successful construction and governance of a Smart City integrating the contents of an Artificial Intelligence World Society (AIWS City):

Artificial Intelligence: the foundation of the Smart City and the Artificial Intelligence World Society (AIWS); Building a wide AI network: AI of Things (AIOT), all areas of urban management serving residents are based on an artificial intelligence platform that connects anytime, anywhere in real time .

AIWS (*Artificial Intelligence World Society*) must be managed according to the artificial intelligence indicators of Security - Well-being - Safety, providing AI solutions and confirming the electronic imprint by block chain.

AIWS (*Artificial Intelligence World Society*) must develop a set of ISO for urban ethics and civilization. AIWS residential innovation is integrated, linked and shared to serve businesses and improve urban service quality.

AIWS (*Artificial Intelligence World Society*) is associated with Digital Government, AI Government. The space of AIWS is not limited by geographical space and can share and connect cultures to build a harmonious society.

AIWS (*Artificial Intelligence World Society*) must strengthen the sharing economy, improve the ability to make electronic payments, use digital currency and block chain to help improve GDP, monitor budget revenue, fight against money laundering and reduce crime.

AIWS (*Artificial Intelligence World Society*) offers solutions to control crises in the virtual world and limit negative impacts on the real world.

AIWS (*Artificial Intelligence World Society*) strengthens the healthcare economy, marketing, sales and content production;

AIWS (*Artificial Intelligence World Society*) increases transparency, makes it easy to check and supervise all decisions and activities of the City Government and functional agencies to ensure efficiency.

AIWS (*Artificial Intelligence World Society*) promotes the development of self-management skills, such as active learning, resilience, stress tolerance, and flexibility.

7. Recommendations for planning and building a Smart City model associated with the contents of AIWS

Step 1. Planning virtual space, real - virtual connection space

Step 2. Building the AIWS Urban Ethics and Civilization Index on the basis of connecting with the ISO 37120:2018 and ISO 37122:2019 Smart City

Step 3. Contract Agreement on AIWS

Step 4. Building a legal corridor to widely recognize the legal AIWS model

Step 5. Exploiting the existing block chain platform, connecting artificial intelligence and providing access and mining to all levels of urban government - for each household, each citizen

Step 6. Connecting World Leaders, Mayors, Urban Professionals, and Innovative Scientists to create data banks for AIWS and learning machine

Step 7. Enhancing the city's competitiveness, resilience to natural disasters, accidents and economic crises

Step 8. Operating on trial Artificial Intelligence centers at all levels of urban government - connecting people; Calibrating and putting AIWS to serve the citizens of Smart City towards a livable city, a city of values.

8. Grounds for determining ethical contents in AI administration in Smart City associated with the contents AIWS

AI design is associated with social ethics.

AI design secures the interests of all stakeholders.

AI Design guarantees engagement and connection: this is a pillar of the Industrial Revolution 4.0 and one of 15 products to be Smart City.

AI Design ensures regional linkages with the community focus**:** Strengthen connectivity and cooperation in the region as a whole, through a regional economic development strategy.

AI Design creates strategic impetus and foundational initiatives**:** A good strategy needs to identify initiatives for "early success" thereby creating momentum.

AI Design is clear about the purpose of usage and benefits.

AI Design on the basis of building strategic thinking**:** The pace of technology development is so fast, there needs a timeframe for strategic levels. Spending more time on technical details in a strategy will make it soon out of date.

AI Design should withdraw lessons. AI design integrates urban planning and inter-regional planning in Smart City.

AI Design must be compared with the outcome evaluation indicators: Some organizations are developing these indicators such as ISO (In 2018, Associate Professor Nguyen Van Thanh completed the First Draft ISO Smart City for WCCD), ITU through the United Nations. It should be noted that these indicators must be meaningful to residents and businesses who will choose the place to invest and live.

AI Design must be associated with creating a long-term Smart Culture: Having the involvement of all stakeholders, sharing challenges and successes are key to creating a Smart City Culture.

AI Design must ensure controllability and allowable limits.

ỨNG DỤNG TRÍ TUỆ NHÂN TẠO TRONG CHĂM SÓC SỨC KHỎE VÀ CÁC VẤN ĐỀ VỀ ĐẠO ĐỨC

TS. Nguyễn Phi Lê, TS. Đinh Viết Sang
Trường Công nghệ thông tin và Truyền thông,
Đại học Bách Khoa Hà Nội,
Trung tâm nghiên cứu quốc tế và trí tuệ nhân tạo (BK.AI)

1. Ứng dụng trí tuệ nhân tạo trong y học

Ra đời từ những năm 1950, trí tuệ nhân tạo (AI) được mô tả là *"khoa học và công nghệ tìm cách tạo ra các hệ thống tính toán thông minh như con người"*. Mục tiêu của trí tuệ nhân tạo là xây dựng các hệ thống máy tính có thể đại diện hoặc tái tạo lại quá trình suy nghĩ của con người. Trong những năm gần đây, cùng với sự dùng nổ của dữ liệu lớn và các hệ thống siêu máy tính, AI đã đạt được những bước tiến quan trọng và được ứng dụng ở nhiều lĩnh vực trong đời sống. Trong đó, chăm sóc sức khoẻ là một trong những ngành có nhu cầu ứng dụng AI cao. AI có tiềm năng mang lại lợi ích rất lớn trong y tế. Hiện nay, ở nhiều nước trên thế giới, trong đó có nước ta, hệ thống y tế luôn trong tình trạng quá tải; các bác sỹ luôn phải làm việc quá giờ. Trong bối cảnh đó, ứng dụng AI có thể thay thế một phần việc khám chữa bệnh của bác sỹ cũng như đẩy nhanh việc khám chữa bệnh bằng việc tự động hoá một số công đoạn. Một số thành tựu cụ thể ứng dụng AI trong chăm sóc sức khoẻ có thể kể đến như: IBM Watson Oncology, một hệ thống chẩn đoán bệnh tự động ứng dụng trí tuệ nhân tạo, có thể quyết định các loại thuốc điều trị cho bệnh nhân ung thư với hiệu quả tương đương hoặc tốt hơn so với các chuyên gia con người. Thuật toán đọc thông tin X-quang của Stanford cũng xác định bệnh viêm phổi tốt hơn bác sĩ X-quang con

người. Những ví dụ trên đã cho chúng ta thấy rằng AI là một ứng cử viên tuyệt vời trong việc hỗ trợ chẩn đoán bệnh. Một số nhà nghiên cứu cho rằng, một khi chi phí ban đầu trở nên thấp hơn, các dịch vụ sử dụng AI sẽ là một đối thủ đáng gờm mà con người khó có thể cạnh tranh[1]. Trong nghiên cứu[2], Sani viết rằng "Hiện nay, AI đang được sử dụng để chẩn đoán ung thư, bệnh lao, da, mắt, đột quỵ và các bệnh khác, và nó chính xác hơn, nhanh hơn và rẻ hơn". Trong một số nghiên cứu, AI đã được sử dụng để hỗ trợ phát hiện bệnh lao trên ngực thông qua chụp X quang. Trong số 150 nghiên cứu đã được thực hiện, AI *"có tỉ lệ chính xác phân loại lao khi chụp X-quang phổi với AUC là 0,99"*[3]. Điều này có thể giúp giảm thời gian chờ đợi của bệnh nhân vì bệnh nhân được chẩn đoán nhanh hơn, và do đó có thể được điều trị nhanh hơn[4]. AI cũng có thể phân tích hình ảnh chẩn đoán như MRI, X-quang và PET với độ chính xác tốt hơn và thời gian nhanh hơn so với các bác sĩ. Ứng dụng AI có thể giúp bác sỹ có thể điều trị cho bệnh nhân trước khi tình trạng của họ quá nghiêm trọng và sẽ tránh được các vấn đề có thể phát sinh do lỗi của bác sĩ[5].

AI cũng có thể giúp những người lớn tuổi, một thành phần chiếm hơn 10% dân số toàn thế giới[6]. Trong một nghiên cứu liên quan tới người cao tuổi, AI được so sánh với bốn bác sĩ X-quang trong việc phát hiện bệnh viêm phổi ở bệnh nhân. Nghiên cứu cho thấy rằng AI hoạt động tốt hơn các chuyên gia y tế[7]. Ngoài ra, AI còn được ứng dụng để cải thiện chăm sóc sức khoẻ trong lĩnh vực sức khoẻ tinh thần. Các mô hình AI tinh vi có thể được sử dụng để thay thế các nhà trị liệu trong các tình huống bệnh nhân không có khả năng tiếp cận với các chuyên gia chăm sóc sức khỏe tâm thần. Các mô hình học máy có thể được huấn luyện để để đáp ứng với bệnh nhân như nhà trị liệu bằng cách hướng dẫn bệnh nhân. Ngoài ra, chatbots cho phép các cá nhân nói chuyện với các con bot để giảm bớt trầm cảm và lo lắng do cô đơn. Các robot xã hội như Paro và

eBear đã được thử nghiệm. Mục đích chính là cung cấp sự đồng hành cho những bệnh nhân cao tuổi có tương tác xã hội hạn chế do tình trạng của họ[8] [9]. Bên cạnh đó, các robot phẫu thuật ngày càng được sử dụng nhiều nhằm hỗ trợ các thủ tục phẫu thuật cho các bác sỹ. Robot phẫu thuật được chia làm hai loại: loại thứ nhất chỉ hỗ trợ bác sĩ trong các thủ tục phẫu thuật và loại thứ hai là những người máy hoạt động độc lập. Rô bốt MAKO và Da Vinci rô bốt là những ví dụ về rô bốt phẫu thuật hỗ trợ bác sĩ[9]. Hệ thống AI PREDICT là một phần mềm hỗ trợ quyết định đồng thời, tích hợp với hệ thống quản lý thực hành để truy xuất dữ liệu bệnh nhân cùng với dữ liệu được nhập vào, nhằm dự đoán xác suất bị bệnh tim mạch trong vòng 5 năm tới[10].

2. Các vấn đề liên quan tới đạo đức trong nghiên cứu và sử dụng AI

Bên cạnh những khía cạnh tích cực của việc sử dụng AI trong hệ thống chăm sóc sức khỏe, vấn đề đạo đức trong việc nghiên cứu và sử dụng AI là một chủ đề quan trọng, thu hút được nhiều quan tâm từ cộng đồng. Vòng đời của một hệ thống AI là sự lặp lại của bốn quá trình: (1) Nghiên cứu phát triển, (2) Thí nghiệm đánh giá; (3) Đưa vào sử dụng; (4) Phản hồi và cải tiến. Trong bốn quá trình này, (1) và (3) là hai quá trình có liên quan nhiều nhất tới khía cạnh đạo đức. Đối với quá trình nghiên cứu và phát triển AI, một vấn đề nổi cộm đấy là việc đảm bảo tính riêng tư, bảo mật của các dữ liệu được sử dụng trong việc huấn luyện và kiểm thử các mô hình học máy. Muốn xây dựng một mô hình học máy chúng ta bắt buộc phải thu thập rất nhiều dữ liệu thật từ bệnh nhân. Những dữ liệu này chứa rất nhiều thông tin cá nhân cần bảo mật. Nếu không có quy trình sử dụng dữ liệu cẩn thận thì thông tin cá nhân rất dễ bị rò rỉ. Trong quá trình thứ (3), điều quan ngại là việc sử dụng AI có thể dẫn tới nhiều hệ lụy về mặt đạo đức. Trước hết, đấy là vấn đề độ chính xác của các quyết định, các thao tác được đưa ra/thực hiện bởi AI. Ví dụ, làm

sao để có thể đảm bảo rằng một cỗ máy hoạt động bằng AI sẽ không làm tổn thương bệnh nhân theo bất kỳ cách nào. Một ví dụ khác là việc đưa ra các chẩn đoán bệnh dựa trên các mô hình AI liệu có độ chính xác tới mức nào? Bởi vì một kết luận sai về tình trạng bệnh của bệnh nhân có thể gây ra những hậu quả khôn lường. Hơn nữa, một khi có sự sai sót xảy ra thì ai là người chịu trách nhiệm: Người tạo ra mô hình AI hay người sử dụng các sản phẩm ứng dụng AI?

Một vấn đề đạo đức khác là việc xuất hiện của các rô bốt có thể sẽ làm suy giảm cơ hội nghề nghiệp của một bộ phận lớn các y bác sỹ, hay làm biến mất một số dịch vụ y tế truyền thống. Ví dụ, sự phát triển của robot phẫu thuật được hỗ trợ bởi AI cuối cùng sẽ có thể khiến một số bác sĩ phẫu thuật phải rời bỏ công việc bởi vì một AI sẽ có thể làm việc làm tốt hơn và rẻ hơn. Tương tự như vậy, việc triển khai AI trong chăm sóc sức khỏe tâm thần sẽ có thể thay thế các bác sĩ và bác sĩ tâm thần đã được đào tạo trong lĩnh vực chăm sóc sức khỏe.

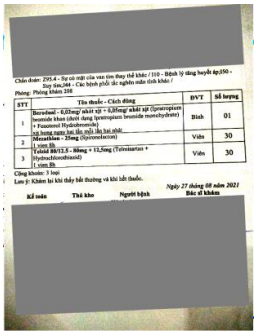
Một vấn đề đạo đức quan trọng khác là làm thế nào để bảo vệ quyền riêng tư và an ninh của bệnh nhân. Hệ thống AI có thể được sử dụng để phân tích thông tin hiện có trong cơ sở dữ liệu để rút ra những thông tin ẩn liên quan tới bệnh nhân. Đây là một vấn đề cần phải được cân nhắc cẩn thận vì nó có khả năng xâm phạm quyền riêng tư của bệnh nhân bằng cách sử dụng thông tin của họ theo những cách không được ủy quyền rõ ràng bởi người bệnh.

3. Một số giải pháp đảm bảo tính riêng tư của bệnh nhân trong dữ liệu y tế

Để bảo vệ thông tin cá nhân của người cung cấp dữ liệu, dữ liệu thường phải trải qua bước xoá bỏ định danh trước khi đưa vào sử dụng. Hình 1 minh hoạ ảnh nội soi được sử dụng trong một ứng dụng AI phát hiện polyp đại tràng. Có thể thấy rằng phần thông tin cá nhân của bệnh nhân đã được bôi đen. Một ví dụ khác trong Hình 2 là ảnh một đơn thuốc đã được che phần thông tin của bệnh nhân. Tuy nhiên, có thể thấy rằng, mặc dù việc phi định danh giúp xoá bỏ thông

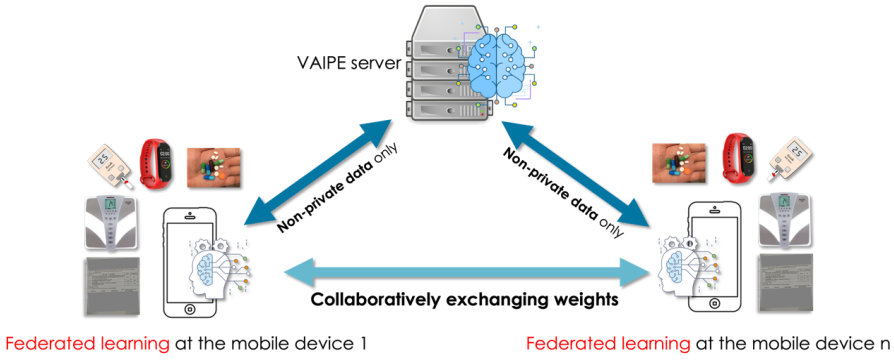
tin cá nhân trong từng dữ liệu, thông tin cá nhân vẫn có thể bị truy ra bằng cách kết hợp nhiều dữ liệu với nhau. Ví dụ, mặc dù thông tin bệnh nhân trên đơn thuốc đã bị xoá, từ tên các viên thuốc trong đơn và cơ sở dữ liệu thuốc của các bệnh viện có thể suy ra được đơn thuốc này xuất phát từ nhà thuốc nào.

Hình 1. Ảnh nội soi đã được che phần dữ liệu người dùng



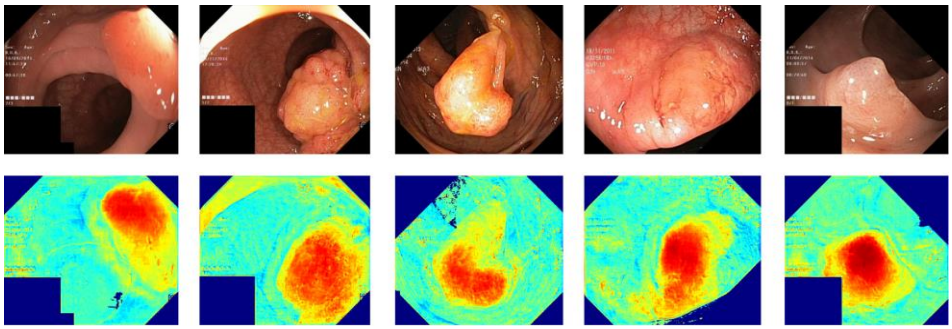
Hình 2. Ảnh đơn thuốc đã được che phần dữ liệu người dùng

Để giải quyết triệt để hơn nữa vấn đề bảo vệ thông tin cá nhân, hiện nay kỹ thuật học liên kết (Federated learning) đã và đang được đưa vào sử dụng. Trong mô hình học liên kết, việc huấn luyện mô hình sẽ được thực hiện trên các thiết bị của chính người dùng thay vì tập trung trên máy chủ. Bằng cách này, chúng ta không cần thu thập dữ liệu của bệnh nhân mà chỉ cần thu thập các kết quả của quá trình huấn luyện (Hình 3).



Hình 3. Mô hình học liên kết

Một giải pháp liên quan tới độ chính xác trong các kết quả và việc ra quyết định dựa trên các hệ thống AI là việc sử dụng những kỹ thuật mới giúp chúng ta có thể giải thích được kết quả của AI (Hình 4). Việc này sẽ giúp cho người sử dụng phân tích độ tin cậy của AI và từ đó đưa ra có hay không nên tin vào kết qủa đưa ra bởi AI, giảm thiểu những sai sót có thể gây ra bởi AI.



Hình 4. Kết quả polyp đại tràng phát hiện bởi AI.
Vùng màu đỏ thể hiện khả năng bị ployp

4. Kết luận

Trí tuệ nhân tạo với những thành tựu vượt bậc đã cho thấy lợi ích không thể phủ nhận của nó trong mọi khía cạnh của đời sống. Tuy nhiên, việc nghiên cứu và ứng dụng trí tuệ nhân tạo đang làm dấy lên những câu hỏi và nghi ngại về mặt đạo đức. Để tránh những hệ luỵ có thể xảy ra, chúng ta cần cẩn trọng trong quá trình nghiên cứu và sử dụng các ứng dụng trí tuệ nhân tạo.

[1] J. Bali, R. Garg, R. Bali, Artificial intelligence (AI) in healthcare and biomedical research:Why a strong computational/AI bioethics framework is required? Indian J. Ophthalmol. 67(1), 3 (2019) http://library.semo.edu:2275/10.4103/ijo.IJO_1292_18

[2] T. Sani, Artificial intelligence: AI in healthcare: Will AI replace doctors? Electronics for you, (2019, Apr 01). Retrieved from https://library.semo.edu:2443/login?url=https://library.semo.edu:4836/docview/2212787461?accountid=38003

[3] . Lakhani, B. Sundaram, Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology (Thomas Jefferson University Hospital) 284(2), 574-582 (2017)

[4] V.H. Buch, I. Ahmed, M. Maruthappu, Artificial intelligence in medicine: Current trends and future possibilities. Br. J. Gen. Pract. 68(668), 143-144 (2018)

[5] Icahn school of medicine at mount sinai to establish world class center for artificial intelligence - hamilton and amabel james center for artificial intelligence and human health, Targeted News Service (2019, June 11). Retrieved from https://library.semo.edu:2443/login?url=https://library.semo.edu:4836/docview/2240407128?accountid=38003
F. Pesapane, C. Volonté, M. Codari, F. Sardanelli, Artificial intelligence as a medical device in radiology: Ethical and regulatory issues in Europe and the United States. Insights Imaging 9(5), 745-753 (2018). https://doi.org/10.1007/s13244-018-0645-y

[6] M.E. Pollack, Intelligent technology for an aging population: The use of AI to assist elders with cognitive impairment. AI Mag. 26(2), 9-9 (2005)

[7] E.J. Topol, High-performance medicine: The convergence of human and artificial intelligence. Nat. Med. 25(1), 44 (2019)

[8] A. Fiske, P. Henningsen, A. Buyx, Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J. Med. Internet Res. 21(5), e13216 (2019). https://doi.org/10.2196/13216

[9] M. Shi, Z. Zhao, The impact of intelligent medicine on health care against the backdrop of big data, in Proceedings of the 2018 Joint International Advanced Engineering and Technology Research Conference (JIAET, 2018). https://doi.org/10.2991/jiaet-18.2018.84

[10] J. Warren, How will AI change health care delivery? N. Z. Med. Student J. 28, 10-12 (2019) Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=138364986&site=ehost-live

Tài liệu tham khảo

1. J. Bali, R. Garg, R. Bali, Artificial intelligence (AI) in healthcare and biomedical research:Why a strong computational/AI bioethics framework is required? Indian J. Ophthalmol. 67(1), 3 (2019) http://library.semo.edu:2275/10.4103/ijo.IJO_1292_18

2. T. Sani, Artificial intelligence: AI in healthcare: Will AI replace doctors? Electronics for you, (2019, Apr 01). Retrieved from https://library.semo.edu:2443/login?url=https://library.semo.edu:4836/docview/2212787461?accountid=38003

3. P. Lakhani, B. Sundaram, Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology (Thomas Jefferson University Hospital) 284(2), 574-582 (2017)

4. V.H. Buch, I. Ahmed, M. Maruthappu, Artificial intelligence in medicine: Current trends and future possibilities. Br. J. Gen. Pract. 68(668), 143-144 (2018)

5. Icahn school of medicine at mount sinai to establish world class center for artificial intelligence - hamilton and amabel james center for artificial intelligence and human health, Targeted News Service (2019, June 11). Retrieved from https://library.semo.edu:2443/login?url=https://library.semo.edu:4836/docview/2240407128?accountid=38003

6. F. Pesapane, C. Volonté, M. Codari, F. Sardanelli, Artificial intelligence as a medical device in radiology: Ethical and regulatory issues in Europe and the United States. Insights Imaging 9(5), 745-753 (2018). https://doi.org/10.1007/s13244-018-0645-y

7. M.E. Pollack, Intelligent technology for an aging population: The use of AI to assist elders with cognitive impairment. AI Mag. 26(2), 9-9 (2005)

8. E.J. Topol, High-performance medicine: The convergence of human and artificial intelligence. Nat. Med. 25(1), 44 (2019)

9. A. Fiske, P. Henningsen, A. Buyx, Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J. Med. Internet Res. 21(5), e13216 (2019). https://doi.org/10.2196/13216

10. M. Shi, Z. Zhao, The impact of intelligent medicine on health care against the backdrop of big data, in Proceedings of the 2018 Joint International Advanced Engineering and Technology Research Conference (JIAET, 2018). https://doi.org/10.2991/jiaet-18.2018.84

11. J. Warren, How will AI change health care delivery? N. Z. Med. Student J. 28, 10-12 (2019) Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=138364986&site=ehost-live

UTILIZING AI IN HEALTHCARE AND ETHICAL ISSUES

Dr. Nguyen Phi Le, Dr. Dinh Viet Sang
School of Information and Communication,
Hanoi University of Science and Technology,
International Research Center for Artificial Intelligence (BK.AI)

1. Application of Artificial Intelligence in medical science

Founded in the 1950s, artificial intelligence (AI) has been described as *"science and technology that seeks to create computing systems as intelligent as humans"*. The goal of artificial intelligence is to build computer systems that can represent or reconstruct the human thinking process. In recent years, along with the explosive use of big data and supercomputer systems, AI has made major strides and been applied in many aspects of life. In particular, healthcare is one of the foremost industries with high demands for AI application. AI has the potential to bring huge benefits in healthcare. Currently in many countries around the world, including Vietnam, the healthcare system is always overloaded and doctors always have to work overtime. In this context, AI application can partially replace doctors' medical examination and treatment as well as accelerate medical examination and treatment by automating some stages. Some of the specific achievements of AI application in healthcare are: IBM Watson Oncology - an automated disease diagnostic system that applies artificial intelligence - can decide drugs for treatment of cancer patients with equal or better efficiency than human experts. Stanford's radiology algorithm may pick up pneumonia better than human radiologists. These examples have shown that AI is an excellent candidate to support medical diagnosis. Some researchers

argue that once the initial cost becomes lower, AI-applied services will be a redoubtable competitor that humans can hardly compete with[1]. In his study[2], Sani wrote that "Now, AI is being employed in diagnosing cancer, tuberculosis, skin, eyes, stroke and other conditions, and it is more precise, accurate, faster and cheaper." In some studies, AI has been used to aid in the detection of tuberculosis on the chest through X-rays. Of the 150 studies that have been conducted, AI "can accurately classify detecting tuberculosis (TB) at chest radiography with an AUC of 0.99"[3]. This can help reduce a patient's waiting time because the patient is diagnosed faster, and can therefore be treated faster[4]. AI can also analyze diagnostic imaging such as MRI, X-radiation and PET which are more accurate and faster than doctors. AI applications can help doctors treat patients before their condition get too severe and will avoid medical errors that may caused by doctors[5].

AI can also assist elders, who take approximately 10 percent of the world's population. In a study involving the elderly, AI was compared to four radiologists in detecting pneumonia in patients, and the conclusion was that the AI outperformed the radiologists[6]. AI is also used to improve mental healthcare. Sophisticated AI models can be used to replace therapists in situations where patients do not have access to mental health professionals. Machine learning models can be trained to respond to patients as therapists by guiding patients. In addition, chatbots or chatterbots allow individuals to talk to bots to alleviate depression and anxiety caused by loneliness. Social robots such as Paro and eBear have been tested. The main aim is to provide companionity for elderly patients with limited social interaction due to their condition[7]. In addition, surgical robots are increasingly being used to support doctors in surgical procedures. Surgical robots are divided into two categories: the first type just assists doctors in surgical procedures and the second type is

independently operated robots. MAKO and Da Vinci robots are examples of surgical robots that assist doctors. THE PREDICT AI system is a simultaneous decision support software, integrated with a practice management system to retrieve patient data along with input data, to predict the probability of cardiovascular disease within the next 5 years[8].

2. Ethical issues in AI research and usage

Besides the positive aspects of using AI in healthcare systems, ethical issues in the study and usage of AI are becoming an important topic, attracting great interest from the community. The life cycle of an AI system is the repetition of four processes: (1) Research and Development, (2) Evaluation Experiment; (3) Putting into use; (4) Feedback and Improvement. Of these four processes, (1) and (3) are the two processes that are most ethically related. For AI research and development, a key issue is ensuring the privacy and security of the data used in the training and testing of machine learning models. To build a machine learning model, we have to collect a lot of real data from patients. This data contains a lot of personal information that needs to be secured. Without a process of strict use of data, personal information is highly susceptible to leakage. In the third process, it is of concern that the use of AI can lead to many ethical implications. First of all, it is a matter of the accuracy of decisions, operations made / performed by AI. For example, how can we ensure that an AI-powered machine won't hurt a patient in any way? Another example is how accurate is making disease diagnoses based on AI models? It is because a false conclusion about the patient's condition can have unintended consequences. Moreover, once a mistake occurs, who would take responsibility: The creators the AI models or the users of AI application products?

Another ethical issue is that the emergence of robots could undermine the career opportunities of a large portion of physicians, or erase some traditional health services. For example, the development of AI-powered surgical robots will eventually cause some surgeons to leave their jobs because an AI will be able to work better and cheaper. Likewise, the deployment of AI in mental health care will be able to replace doctors and psychiatrists who have been trained in this healthcare sector.

Another important ethical issue is how to protect patient privacy and security. AI systems can be used to analyze existing information in the database to draw out hidden information related to patients. This is an issue that needs to be carefully weighed as it has the potential to invade patients' privacy by using their information in ways that are not clearly authorized by them.

3. Some solutions to ensure patients' privacy in medical data

To protect the personal information of the data providers, the data often has to undergo identity deletion before being put into use. Figure 1 illustrates the endoscopic image used in an AI application that detects colon polyps. The patient's personal information can be seen as blacked out. Another example in Figure 2 is a picture of a prescription that has covered the patient's information. However, it can be seen that, although non-identification helps to delete personal information in each data, personal information can still be traced by combining multiple data together. For example, although the patient information on the prescription has been deleted, it is still possible to track which pharmacy this prescription came from by using the names of the pills in the prescription and the drug databases of hospitals.

Figure 1. Endoscopic image of which personal information has been hidden

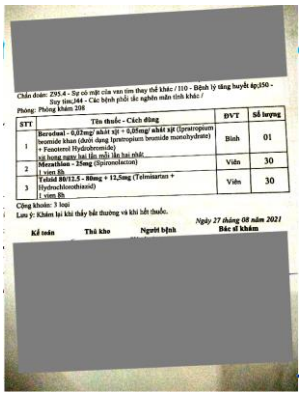


Figure 2. Prescription photo of which personal information has been hidden

In order to better solve the problem of protecting personal information, the federated learning has been being put into use. In the federated learning model, model training will be performed on the user's own devices instead of focusing on the server. By this way, we do not need to collect patients' data but just collect the results of the training process (Fig. 3).

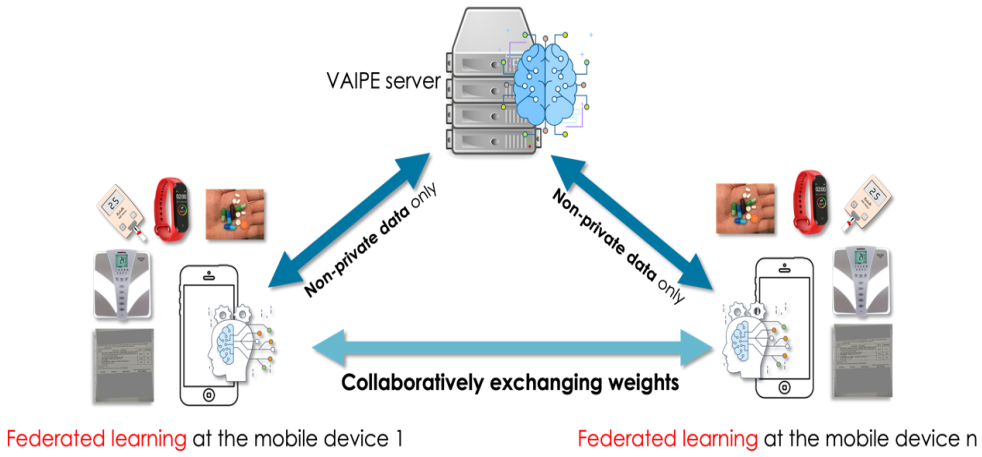


Figure 3. Federated learning model

One solution to accuracy in outcomes and decision-making based on AI systems is the use of new techniques that enable us to explain the results of AI (Figure 4). This will help users analyze the reliability of AI and thereby give whether or not to believe the results given by AI, minimizing the errors that can be caused by AI.

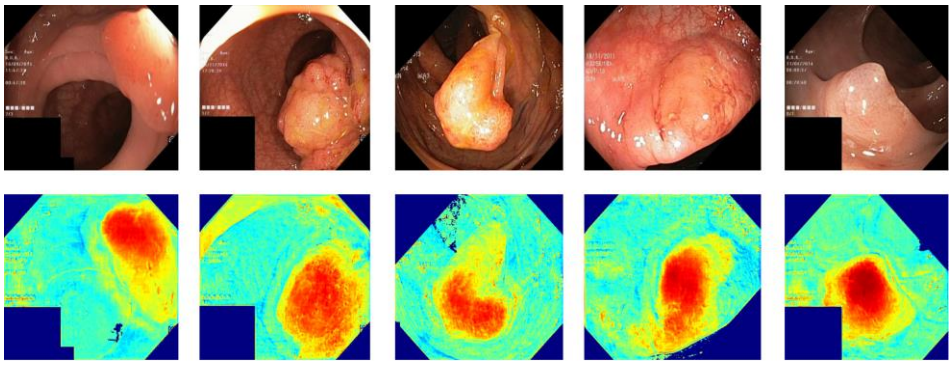


Figure 4. Colon polyp results detected by AI.
The red area represents the possibility of polyps

4. Conclusion

Artificial intelligence with remarkable achievements has shown its undeniable benefits in every aspect of life. However, the research and application of artificial intelligence is raising ethical questions and doubts. To avoid possible consequences, we need to be careful in the process of researching and using artificial intelligence applications.

[1] J. Bali, R. Garg, R. Bali, Artificial intelligence (AI) in healthcare and biomedical research:Why a strong computational/AI bioethics framework is required? Indian J. Ophthalmol. 67(1), 3 (2019) http://library.semo.edu:2275/10.4103/ijo.IJO_1292_18

[2] T. Sani, Artificial intelligence: AI in healthcare: Will AI replace doctors? Electronics for you, (2019, Apr 01). Retrieved from https://library.semo.edu:2443/login?url=https://library.semo.edu:4836/docview/2212787461?accountid=38003

[3] P. Lakhani, B. Sundaram, Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology (Thomas Jefferson University Hospital) 284(2), 574-582 (2017)

[4] V.H. Buch, I. Ahmed, M. Maruthappu, Artificial intelligence in medicine: Current trends and future possibilities. Br. J. Gen. Pract. 68(668), 143-144 (2018)

[5] Icahn school of medicine at mount sinai to establish world class center for artificial intelligence - hamilton and amabel james center for artificial intelligence and human health, Targeted News Service (2019, June 11). Retrieved from https://library.semo.edu:2443/login?url=https://library.semo.edu:4836/docview/2240407128?accountid=38003

F. Pesapane, C. Volonté, M. Codari, F. Sardanelli, Artificial intelligence as a medical device in radiology: Ethical and regulatory issues in Europe and the United States. Insights Imaging 9(5), 745-753 (2018). https://doi.org/10.1007/s13244-018-0645-y

[6] E.J. Topol, High-performance medicine: The convergence of human and artificial intelligence. Nat. Med. 25(1), 44 (2019)

[7] A. Fiske, P. Henningsen, A. Buyx, Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J. Med. Internet Res. 21(5), e13216 (2019). https://doi.org/10.2196/13216

[8] J. Warren, How will AI change health care delivery? N. Z. Med. Student J. 28, 10-12 (2019) Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=138364986&site=ehost-live

References

1. J. Bali, R. Garg, R. Bali, Artificial intelligence (AI) in healthcare and biomedical research: Why a strong computational/ AI bioethics framework is required? Indian J. Ophthalmol. 67(1), 3 (2019) http://library.semo.edu:2275/10.4103/ijo.IJO_1292_18

2. T. Sani, Artificial intelligence: AI in healthcare: Will AI replace doctors? Electronics for you, (2019, Apr 01). Retrieved from https://library.semo.edu:2443/login?url=https://library.semo.edu:4836/docview/2212787461?accountid=38003

3. P. Lakhani, B. Sundaram, Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology (Thomas Jefferson University Hospital) 284(2), 574-582 (2017)

4. V.H. Buch, I. Ahmed, M. Maruthappu, Artificial intelligence in medicine: Current trends and future possibilities. Br. J. Gen. Pract. 68(668), 143-144 (2018)

5. Icahn school of medicine at mount sinai to establish world class center for artificial intelligence - hamilton and amabel james center for artificial intelligence and human health, Targeted News Service (2019, June 11). Retrieved from https://library.semo.edu:2443/login?url=https://library.semo.edu:4836/docview/2240407128?accountid=38003

6. F. Pesapane, C. Volonté, M. Codari, F. Sardanelli, Artificial intelligence as a medical device in radiology: Ethical and regulatory issues in Europe and the United States. Insights Imaging 9(5), 745-753 (2018). https://doi.org/10.1007/s13244-018-0645-y

7. M.E. Pollack, Intelligent technology for an aging population: The use of AI to assist elders with cognitive impairment. AI Mag. 26(2), 9-9 (2005)

8. E.J. Topol, High-performance medicine: The convergence of human and artificial intelligence. Nat. Med. 25(1), 44 (2019)

9. A. Fiske, P. Henningsen, A. Buyx, Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J. Med. Internet Res. 21(5), e13216 (2019). https://doi.org/10.2196/13216

10. M. Shi, Z. Zhao, The impact of intelligent medicine on health care against the backdrop of big data, in Proceedings of the 2018 Joint International Advanced Engineering and Technology Research Conference (JIAET, 2018). https://doi.org/10.2991/jiaet-18.2018.84

11. J. Warren, How will AI change health care delivery? N. Z. Med. Student J. 28, 10-12 (2019) Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=138364986&site=ehost-live

VẤN ĐỀ ĐẠO ĐỨC TRÍ TUỆ NHÂN TẠO VÀ NHỮNG TÁC ĐỘNG XÃ HỘI

TS. Hoàng Vũ Linh Chi
Viện Xã hội học, Viện Hàn lâm KHXH Việt Nam

1. Đặt vấn đề

Trước khi có AI, con người đang sống trong một không gian nhị phân bao gồm không gian vật lý và không gian xã hội của con người. Trong không gian nhị phân này, trật tự cho các hoạt động của con người được quyết định bởi sự tương tác giữa con người với nhau và giữa con người với vật thể và con người đóng vai trò là người hình thành và thống trị các trật tự xã hội. Với sự phát triển nhanh chóng của siêu dữ liệu, điện toán đám mây, Internet vạn vật, các thiết bị di động thông minh trên các lĩnh vực khác nhau của xã hội loài người và thúc đẩy sự ra đời của kỷ nguyên thông minh, đưa con người đến một không gian mới - không gian được đánh dấu bởi không gian vật lý, không gian xã hội của con người và không gian mạng. Trong không gian bậc ba, trật tự của xã hội loài người sẽ luôn được tái cấu trúc. Cho dù con người có nhận thức được sự thay đổi đó hay không, thì những ảnh hưởng sâu sắc đến đời sống xã hội của con người mà AI mang lại sẽ dần được chấp nhận trong mọi tầng lớp xã hội.

Nhu cầu cấp thiết cần phải giải quyết những thách thức mà công nghệ AI đặt ra liên quan đến các thuật toán mang tính thiên vị, bảo vệ quyền riêng tư của con người, dữ liệu cá nhân, những nguy cơ tạo ra sự loại trừ và bất bình đẳng dưới hình thức mới, các vấn đề về phân chia lợi ích và rủi ro, trách nhiệm giải trình, tác động đến việc làm và tương lai của công việc, nhân quyền, an ninh và rủi ro

của việc sử dụng là mối quan tâm của nhiều tổ chức và chính phủ quan tâm đến những tác động đạo đức của trí tuệ nhân tạo.

2. Trí tuệ nhân tạo và đạo đức trí tuệ nhân tạo

Trí tuệ nhân tạo, còn được gọi là AI, có thể được định nghĩa là một tập hợp các công nghệ khác nhau có thể được kết hợp lại với nhau để cho phép máy móc hoạt động với những gì có vẻ giống như mức độ thông minh của con người. Điều này bao gồm việc học các quy tắc cần thiết để đưa ra các quyết định và lý do nhất định để đi đến kết luận nhất định, học hỏi từ kinh nghiệm trong quá khứ và tự sửa chữa. Russel và Norvig là các tác giả của cuốn sách "Trí tuệ nhân tạo: cách tiếp cận hiện đại" (Artificial Intelligence: a Modern Approach) định nghĩa AI là "*nghiên cứu và thiết kế các tác nhân thông minh, trong đó tác nhân thông minh là một hệ thống nhận thức được môi trường và thực hiện các hành động nhằm tối đa hóa cơ hội thành công*". Theo Rusel và Norvig hiện nay có bốn trường phái AI. Trường phái thứ nhất, các nhà nghiên cứu tập trung vào việc tạo ra những cỗ máy có suy nghĩ giống như con người. Trường phái này tìm cách tái tạo theo cách nào đó, các quá trình, biển hiện và kết quả của tư duy con người lên máy móc. Trường phái thứ hai tập trung vào việc tạo ra máy móc hoạt động giống con người. Nó tập trung vào hành động giống như những robot trên thế giới hiện nay chứ không tập trung vào quá trình để đạt được hành động đó. Trường phái thứ ba tập trung vào việc phát triển những cỗ máy hoạt động hợp lý. Tính hợp lý có quan hệ mật thiết với tính tối ưu. Cuối cùng, trường phái thứ tư tập trung vào phát triển những cỗ máy có thể suy nghĩ theo lý trí. Việc lập kế hoạch hoặc ra quyết định sẽ được thực hiện một cách tối ưu. Tối ưu ở đây có liên quan đến vấn đề mà hệ thống đang cố gắng giải quyết.

Yếu tố cơ bản nhất chung cho tất cả các trường phái là AI liên quan đến việc nghiên cứu, thiết kế và xây dựng các tác nhân thông

minh có thể đạt được mục tiêu. Một AI linh hoạt và có thể học hỏi từ kinh nghiệm cũng như ý thức, lập kế hoạch và hành động dựa trên cấu hình của nó, có thể được cho là thông minh hơn một AI chỉ có một bộ quy tắc hướng dẫn một số các hành động cố định. Tuỳ vào bối cảnh và phương pháp tiếp cận khác nhau các nhà sáng chế có xu hướng nhấn mạnh một số yếu tố này hơn những yếu tố khác. Ví dụ: các nhà phát triển hệ thống chuyên gia coi AI là kho kiến thức chuyên môn mà con người có thể tham khảo, trong khi các nhà phát triển hệ thống học máy coi AI là thứ có thể khám phá ra kiến thức mới. Như chúng ta thấy, mỗi cách tiếp cận đều có điểm mạnh và điểm yếu riêng.

Các nhà khoa học và chuyên gia AI đồng ý rằng AI đã ảnh hưởng nhanh chóng và đáng kể đến xã hội loài người và cách con người tương tác với nhau. Nó sẽ tiếp tục như vậy, trong quá trình phát triển, AI đưa ra những thách thức lớn về đạo đức và chính trị xã hội đòi hỏi phải có một phân tích triết học và đạo đức kỹ lưỡng. Tác động xã hội của nó cần được nghiên cứu để tránh bất kỳ tác động tiêu cực nào. Chúng ta đang chạy đua với thời gian, cần thiết lập các nguyên tắc đạo đức trước khi công nghệ bắt kịp con người. Đã có rất nhiều các cuộc bàn luận về tính hai mặt của trí tuệ nhân tạo, bởi sự phát triển nhanh chóng của trí tuệ nhân tạo và máy học mang lại những lợi ích tiềm năng to lớn. Bên cạnh tính tích cực của trí tuệ nhân tạo trong một số lĩnh vực như y tế, giáo dục, không gian… tuy nhiên một số nhà khoa học lo ngại về mặt tiêu cực của trí tuệ nhân tạo trong cuộc chạy đua chết người của trí tuệ nhân tạo.

Theo Mark Coeckelbergh, đạo đức của AI là sự thay đổi công nghệ và tác động của nó đối với cuộc sống cá nhân, nhưng cũng là những chuyển đổi trong xã hội và nền kinh tế. Các vấn đề về thành kiến và phân biệt đối xử đã chỉ ra rằng AI có liên quan và sẽ ảnh hưởng đến sự thay đổi nền kinh tế và tiếp đó có lẽ là cấu trúc xã hội của các xã hội (Coeckelbergh, 2020). Theo Brynjolfsson và McAfee

(2014), chúng ta đã bước vào Kỷ nguyên máy móc thứ hai, trong đó máy móc không chỉ là vật dụng hỗ trợ con người, như trong Cách mạng công nghiệp, mà còn là vật thay thế con người. Khi tất cả các ngành nghề và công việc sẽ bị ảnh hưởng bởi AI, xã hội của chúng ta đã được dự đoán sẽ thay đổi đáng kể khi các công nghệ từng được mô tả trong khoa học viễn tưởng bước vào thế giới thực (Andrew McAfee and Erik Brynjolfsson., 2017).

3. Mối quan hệ giữa đạo đức trí tuệ nhân tạo và tác động xã hội

3.1. Các câu hỏi đặt ra về mặt đạo đức trí tuệ nhân tạo

Trí tuệ nhân tạo có thể mang lại rất nhiều lợi ích, hỗ trợ bổ sung cho các hoạt động của con người, thay thế con người làm những công việc khó khăn hay nhàm chán. Giáo sư Andrew Ng. người sáng lập ra Coursera đã nói rằng "Cũng giống như cuộc Cách mạng công nghiệp đã giải phóng nhân loại khỏi sự cực nhọc về thể chất. Tôi nghĩ rằng AI có tiềm năng giải phóng nhân loại khỏi sự cực nhọc về tinh thần". Trí tuệ nhân tạo có thể được sử dụng để cải thiện các dịch vụ công và dịch vụ thương mại. Ví dụ, nhận dạng hình ảnh có thể giúp chẩn đoán các bệnh như ung thư và Alzheimer. Tuy nhiên, trí tuệ nhân tạo cũng tạo ra rất nhiều những rủi ro như đánh giá thấp các kỹ năng của con người, loại bỏ trách nhiệm, giảm sự kiểm soát và xói mòn quyền tự quyết của con người (Floridi, 2018). Chính những vấn đề này đã tạo ra những thách thức mới chưa từng có liên quan đến những vấn đề về đạo đức. Bởi tiềm năng to lớn của những công nghệ AI trong mọi lĩnh vực có thể đem lại lợi nhuận khổng lồ cho các công ty nắm được dữ liệu của khách hàng.

Diễn đàn Kinh tế thế giới đã đưa ra 9 vấn đề đạo đức liên quan đến trí tuệ nhân tạo 1. Mất việc làm; 2. Bất bình đẳng. Làm thế nào để phân phối của cải được tạo ra bởi máy móc; 3. Tính nhân văn. Máy móc ảnh hưởng đến hành vi và tương tác của con người như

thế nào? 4. Làm thế nào để tránh được những lỗi gây ra bởi trí tuệ nhân tạo. 5. Người máy phân biệt chủng tộc. Làm thế nào để có thể loại bỏ sự thiên vị trong AI ?; 6. Bảo mật. Làm cách nào để chúng ta giữ cho AI an toàn trước kẻ thù?; 7. Làm thế nào để chúng ta được bảo vệ khỏi những hậu quả không mong muốn?; 8. Điểm kỳ dị. Làm thế nào để chúng ta kiểm soát một hệ thống thông minh phức tạp?; 9. Quyền của người máy. Làm thế nào để chúng ta xác định cách đối xử nhân đạo của AI? (Bossmann., 2016)

Viện nghiên cứu Alan Turing đã liệt kê một loạt các tác hại tiềm ẩn mà các hệ thống AI có thể gây ra. (1) Định kiến và phân biệt đối xử, (2) Từ chối Quyền tự chủ, quyền của cá nhân, (3) Kết quả không minh bạch, không thể giải thích được hoặc không chính đáng, (4) Xâm phạm quyền riêng tư, (5) Cô lập và hòa tan kết nối xã hội, (6) Kết quả không đáng tin cậy, không an toàn hoặc chất lượng kém (Leslie, 2019). Tuy nhiên, Mueller đã liệt kê 10 vấn đề chính liên quan đến những vấn đề nảy sinh trong quá trình sử dụng một số công nghệ AI gồm: (1) Quyền riêng tư & Giám sát; (2) Thao túng hành vi; (3) Độ mờ của hệ thống AI; (4) Sự thiên vị trong hệ thống quyết định; (5) Tương tác giữa con người-Robot; (6) Tự động hóa và Việc làm; (7) Hệ thống tự chủ; (9) Đạo đức máy móc; (9) Tác nhân đạo đức nhân tạo và (10) Điểm kỳ dị. Tác giả cho rằng ngoài việc "sử dụng có trách nhiệm" cũng cần phải đề cập đến việc "thiết kế có trách nhiệm" trong lĩnh vực này (Müller, 2020).

Những câu hỏi thường được đặt ra liên quan đến các vấn đề đạo đức của trí tuệ nhân tạo:

✓ Làm cách nào để chúng ta sắp xếp mục tiêu của các hệ thống AI tự hành với mục tiêu của chúng ta?

✓ Tương lai của AI có gây ra mối đe dọa hiện hữu cho nhân loại không?

✓ Làm cách nào để con người ngăn các thuật toán có những thành kiến phản cảm về mặt đạo đức?

✓ Có nên sử dụng AI tự hành để tiêu diệt trong chiến tranh?

✓ Có được phép yêu một hệ thống AI không?

✓ AI như xe tự lái nên sử dụng những quy tắc đạo đức nào?

✓ Hệ thống AI có thể ngăn chặn các tác hại đạo đức không?

✓ Hệ thống AI có thể là tác nhân đạo đức không? Nếu vậy, chúng ta nên quy trách nhiệm cho chúng như thế nào?

Ngoài ra tuỳ vào cách tiếp cận mà các câu hỏi được nêu ra khi bàn đến vấn đề đạo đức, tuy nhiên nói chung các câu hỏi tập trung vào điều gì nên được xem là đúng và sai trong việc phát triển và triển khai AI và robot. Ví dụ: Ủy ban châu Âu yêu cầu đưa đạo đức vào việc phát triển và sử dụng các công nghệ mới trong các chương trình cũng như xây dựng các hướng dẫn đạo đức để sử dụng và phát triển AI liên quan đến các quyền cơ bản của con người. Kế hoạch của Uỷ ban châu Âu còn hướng tới mục tiêu thiết lập tiêu chuẩn đạo đức toàn cầu hướng tới trở thành hướng dẫn hàng đầu thế giới về AI đáng tin cậy và có đạo đức.

Theo quan điểm của Tổ chức Giáo dục, Khoa học và Văn hoá Liên hợp quốc (UNESCO), cần các chính sách quốc tế và quốc gia và các khuôn khổ pháp lý để đảm bảo rằng những công nghệ mới nổi này mang lại lợi ích cho nhân loại nói chung. AI cần lấy con người làm trung tâm và vì lợi ích lớn hơn của con người chứ không phải ngược lại. UNESCO đề xuất phát triển một công cụ toàn diện thiết lập tiêu chuẩn toàn cầu để cung cấp cho AI một nền tảng đạo đức vững chắc, không chỉ bảo vệ mà còn thúc đẩy quyền con người và phẩm giá con người. Nếu được thông qua, nó sẽ là một la bàn hướng dẫn đạo đức và một nền tảng quy chuẩn toàn cầu cho phép xây dựng một sự tôn trọng mạnh mẽ đối với các quy định của pháp luật trong thế giới kỹ thuật số (UNESCO, 2020).

Dù chúng ta có thể nghĩ gì về những nguyện vọng này, chắc chắn đúng là còn thiếu các hướng dẫn và quy tắc thực tế, và cần được thống nhất về các hệ thống tự chủ. Ngày nay, hầu hết các khu vực pháp lý trên khắp thế giới chỉ mới bắt đầu điều tra các khía cạnh quy định của AI. Hàng loạt các viện nghiên cứu về đạo đức AI và các tác động của nó được thành lập tại rất nhiều các trường đại học nổi tiếng trên thế giới. Ngoài ra các tổ chức và các công ty công nghệ lớn trên thế giới cũng đang dần thiết lập các nguyên tắc cũng như các vị trí việc làm liên quan đến đạo đức AI nhằm phục vụ cho các sản phẩm, ứng dụng của mình. Do đó, đạo đức trở nên quan trọng ở nhiều tầng của cuộc thảo luận chính sách. Đó là một chủ đề dành cho kỹ sư thiết kế hệ thống, bao gồm cả việc sinh viên học cách xây dựng hệ thống AI. Đây cũng là một chủ đề để xã hội đánh giá tác động của công nghệ AI đối với cuộc sống hàng ngày của người dân. Đạo đức trở thành câu hỏi quan trọng đối với các nhà hoạch định chính sách trong các cuộc thảo luận về AI và công nghệ robot. Lưu ý rằng các khía cạnh đạo đức không chỉ được thảo luận trong các câu hỏi về quy định. Quan trọng hơn nhiều, các câu hỏi đạo đức làm nền tảng cho việc thiết kế các hệ thống AI từ việc xác định ứng dụng đến các chi tiết triển khai. Đạo đức trong AI do đó rộng hơn nhiều và liên quan đến các lựa chọn thiết kế rất cơ bản và cân nhắc về xã hội mà chúng ta muốn sống.

3.2. Những tác động xã hội của trí tuệ nhân tạo

3.2.1. Tác động của AI đối với quyền tự chủ của con người

Quyền tự chủ bao gồm hành động phù hợp với các mục tiêu và giá trị của một người, khác với việc sử dụng quyền tự chủ đơn giản là từ đồng nghĩa với độc lập hoặc kiểm soát (Soenens và cộng sự, 2007). Mặc dù quyền tự chủ của con người từ lâu đã trở nên quan trọng đối với triết học và các ngành khoa học xã hội, nhưng trong lĩnh vực kỹ thuật và khoa học máy tính có xu hướng tập trung vào quyền

tự chủ của máy móc. Một nguyên lý kinh điển của đạo đức sinh học là nguyên tắc tự chủ: các cá nhân có quyền tự quyết định về mình. Với AI, tình hình trở nên phức tạp hơn: khi chúng ta áp dụng AI và tác nhân thông minh của nó, chúng ta sẵn sàng nhường một phần quyền quyết định của mình cho máy móc. Do đó, khẳng định nguyên tắc tự chủ trong bối cảnh của AI có nghĩa là tạo ra sự cân bằng giữa quyền ra quyết định mà chúng ta giữ lại cho chính mình và quyền mà chúng ta giao cho các tác nhân nhân tạo.

Một trong những ảnh hưởng của AI đến quyền tự chủ của con người liên quan đến việc ra quyết định. Khi các hệ thống AI trở nên tự chủ hơn và thay thế con người cũng như thay thế con người đưa ra quyết định trong cách ứng xử ngày càng tăng, chúng ta sẽ có nguy cơ mất khả năng đưa ra các quy tắc sống, quyết định hoặc định hình cuộc sống của mình. Ví dụ: có thể tham khảo ý kiến của cố vấn y tế AI (ví dụ: xem Ada.com). Chúng ta có thể chọn từ bỏ khả năng và sự sẵn lòng của mình để biết và hiểu về cơ thể và bệnh tật của chính mình, để lựa chọn một hệ thống tự động thay thế bác sĩ đa khoa quản lý tình hình bệnh tật, tình trạng sức khoẻ và đưa ra các quyết định quan trọng liên quan đến sức khoẻ của mình. Một chuyên gia dinh dưỡng AI cũng có thể gây ra mối đe dọa tương tự, đưa ra các yêu cầu về những gì họ nên ăn và những gì họ nên tránh. Hơn nữa, công nghệ AI có thể có sức mạnh thực sự buộc thực hiện chế độ ăn kiêng này thông qua tủ lạnh thông minh. Con người có thể từ bỏ việc quản lý chế độ ăn uống của mình cho một hệ thống/ứng dụng như vậy và mất khả năng tự chăm sóc bản thân hoặc quản lý chế độ ăn uống của mình do ngày càng phụ thuộc vào các hệ thống đó. Dần dần, con người sẽ mất đi khả năng hành động một cách tự chủ.

Ngoài ra, thao túng và lừa dối rõ ràng là những trường hợp mà quyền tự chủ của con người bị suy giảm do việc sử dụng các hệ thống AI. Hệ thống AI có thể ảnh hưởng và hình thành sở thích và niềm tin của chúng ta như thế nào. Một số nghiên cứu cho thấy cách

mọi người thay đổi sở thích của mình phù hợp với các đề xuất mà họ nhìn thấy bằng thuật toán đề xuất. Chẳng hạn như những trường hợp Youtube, Cambrige Analytica.

Câu hỏi về sự kiểm soát không chỉ đề cập đến các hệ thống AI tự chủ mà còn đề cập đến các thuật toán tự ra quyết định. Khi con người hay người sử dụng không thể yêu cầu thực hiện lại sự ra quyết định này hoặc rất mất thời gian để có thể thay đổi được quyết định mà máy móc/công nghệ đã thực hiện.

3.2.2. Gia tăng định kiến và phân biệt đối xử

Khi các thuật toán được sử dụng thay cho việc ra quyết định của con người, có lẽ không có gì ngạc nhiên khi chúng đôi khi có nguy cơ tái tạo những thành kiến và bất bình đẳng hiện có trong xã hội. Do đó, các quy tắc không công bằng đã vô tình được mã hóa cứng vào các hệ thống AI, do đó dẫn đến các quyết định phân biệt đối xử. Có rất nhiều ví dụ về sự định kiến, sai lệch của các thuật toán bởi các quyết định dựa trên mẫu tìm được dựa trên dữ liệu lớn. Trong khi các giá trị xã hội, chuẩn mực và thái độ đối với chủng tộc và giới tính vẫn đang phát triển, dấu ấn của chúng vẫn còn trên internet từ đó phần lớn dữ liệu đào tạo cho các thuật toán học máy được lấy ra. Ví dụ, các nghiên cứu đã chứng minh rằng ‘đàn ông’ được liên kết với sếp, chủ tịch, lãnh đạo và giám đốc, trong khi ‘phụ nữ’ được liên kết với người trợ giúp, trợ lý, nhân viên và phụ tá. Do đó, sự thúc đẩy của xã hội đối với sự công bằng và công bằng hơn đã vấp phải những giá trị lịch sử về nghèo đói, giới tính, chủng tộc và dân tộc bị khóa trong dữ liệu lớn. (Barr, 2015; Barrett, 2014; Bass, D. & Huet, 2017). Điển hình là trường hợp Google nhầm lẫn gán nhãn người châu Phi là khỉ đột đã tạo ra một cuộc tranh luận về vấn đề đạo đức của AI. Ngoài ra, các nhà phát triển và các nhà khoa học dữ liệu đã từ lâu nói rằng sự thiên lệch thường dẫn đến những mô hình sai lệch. Có thể kể đến ví dụ về việc Google dạy các thuật toán nhận

diện giày hay tác động bất cân xứng mà đại dịch Covid -19 đang gây ra cho người da màu trong việc tiếp cận dịch vụ chăm sóc sức khỏe (Eliane Röösli, Brian Rice, 2021) đã cho thấy sự hạn chế của các thuật toán trong AI. Một trong những nguyên nhân dẫn đến sự thiên lệch này là do những người lập trình đang sống trong thế giới công nghệ do người da trắng làm chủ, họ thiếu thông tin, dữ liệu về những chủng tộc khác hay những dữ liệu từ nam giới và kết quả do các thuật toán quyết định cuối cùng không phải là kết quả tối ưu.

Sự phức tạp trong thiết kế, chế tạo và lập trình robot, vấn đề trọng tâm về đạo đức ở đây được tạo ra là khả năng truy xuất nguồn gốc của robot, khả năng truy xuất nguồn gốc là khả năng giúp cho robot có thể theo dõi được những nguyên nhân tạo ra tất cả những hành động trong quá khứ. Với những robot có khả năng tự quyết, khả năng đưa ra quyết định và khả năng học hỏi thì việc truy xuất nguồn gốc này chính là mối lo ngại, bởi những robot này không chỉ được lập trình để làm những công việc cụ thể, mà nó còn có thể tự học và tự phát triển bản thân trong sự tương tác với môi trường. Chính điều này tạo ra một lỗ hổng hớn trong khung đạo đức hiện hành đối với khả năng truy xuất nguồn gốc này của robot.

Chưa kể đến một nhân tố khác đó là cạm bẫy giải pháp công nghệ. Thực tế cho thấy công nghệ số đã thay đổi hoàn toàn cách chúng ta sống, làm việc và vui chơi, tuy nhiên không thể biện minh cho việc lạm dụng và ứng dụng ồ ạt các giải pháp công nghệ một cách phi lý trong đời sống công cộng. Một số nhà nghiên cứu lo ngại về một tư tưởng mới nổi lên rằng mọi vấn đề xã hội đều có thể được "sửa chữa" thông qua công nghệ. Đặc biệt trong thời kỳ Covid, các công nghệ nhận diện khuôn mặt, công nghệ áp dụng cho các hoạt động hành chính được sử dụng ngày càng gia tăng đã trở thành chủ đề tranh luận ở rất nhiều nước thành viên Liên minh châu Âu về cách thực hiện, ứng dụng của các hệ thống ADMs trong đời sống con

người. Bởi những ứng dụng của AI hoàn toàn là "black box" (hộp đen) và nếu không thể có khả năng biết chính xác hệ thống ADM làm gì, tại sao và cuối cùng được triển khai, tất cả các nỗ lực khác nhằm cân bằng giữa hệ thống và các quyền cơ bản của con người sẽ không bao giờ đạt được. (Fabio Chiusi, Sarah Fischer, Nicolas Kayser-Bril and Matthias Spielkamp, 2020).

Định kiến và phân biệt đối xử là những rủi ro cố hữu của bất kỳ hoạt động kinh tế hoặc xã hội nào, chính vì vậy việc ra quyết định của con người không tránh khỏi những sai lầm và định kiến. Tuy nhiên, nếu sự thiên vị tương tự khi xuất hiện trong AI có thể có một tác động lớn hơn nhiều, ảnh hưởng và phân biệt đối xử với nhiều người nếu không có cơ chế kiểm soát xã hội chi phối hành vi của con người. Bởi những tác động không mong muốn không diễn ra hay dự đoán được ở trong giai đoạn thiết kế mà do tác động thực tế của các mối tương quan hay mô hình mà hệ thống xác định khi dựa trên tập dữ liệu lớn.

3.2.3. Tác động đến bất bình đẳng

Từ các cuộc cách mạng công nghiệp trước, nỗi lo ngại về việc tự động hoá sẽ thay thế những công việc hiện tại và tạo ra những thiệt hại cho thị trường lao động. Những thay đổi về việc làm liên quan đến tự động hóa và số hóa sẽ không chỉ thể hiện thông qua việc mất việc làm mà còn về chất lượng công việc, vì AI được kỳ vọng sẽ tạo ra nhiều hình thức việc làm mới và đa dạng. Sự thay đổi có thể đồng nghĩa với việc hàng triệu người lao động bị mất việc làm và mất thời gian làm việc. Một số người có thể mất việc làm vĩnh viễn và một số người sẽ phải chuyển đổi việc làm khi công việc của họ sẽ được AI thay thế. Các phân tích dự đoán rằng AI và các công nghệ mới khác sẽ tiếp tục mang lại lợi ích cho những người lao động có kỹ năng cao hơn, người lao động có kỹ năng thấp và trung bình sẽ phải đối mặt với áp lực mất việc làm bởi công nghệ. Những người thất

nghiệp sẽ mất khả năng kiếm thu nhập, điều này tạo ra nhiều bất bình đẳng hơn trong một xã hội có thu nhập ngày càng tăng do thu nhập bình quân gia tăng do năng suất của công nghệ AI (McAfee and Brynjolfsson, 2017). Trong nghiên cứu mới đây của NBER (Mỹ) công nghệ tự động hóa là động lực chính dẫn đến bất bình đẳng thu nhập ở Hoa Kỳ trong 40 năm qua. Trí tuệ nhân tạo đã gây ra sự sụt giảm lương từ 50% đến 70% ở Mỹ, kể từ năm 1980, có thể là do sự giảm lương của những người lao động cổ xanh đã bị thay thế hoặc suy giảm bởi tự động hóa. Trí tuệ nhân tạo, người máy và các công nghệ phức tạp mới đã gây ra khoảng cách lớn về sự giàu có và bất bình đẳng thu nhập. Có vẻ như vấn đề này sẽ tăng tốc, những người có bằng sau đại học nhận thấy mức lương của họ tăng lên, trong khi những người lao động có trình độ học vấn thấp lại giảm đáng kể. Thu nhập thực tế của nam giới không có bằng cấp 3 hiện thấp hơn 15% so với năm 1980 (Acemoglu et al., 2021).

Công nghệ AI và robot được kỳ vọng sẽ cho phép các công ty tối ưu hoá hoạt động kinh doanh của mình, giảm thiểu các chi phí cận biên và hoạt động hiệu quả hơn và năng suất hơn. Điều này chắc chắn sẽ có nghĩa là doanh thu sẽ được chia cho ít người hơn, làm gia tăng bất bình đẳng xã hội. Do đó, các cá nhân nắm giữ quyền sở hữu trong các công ty do AI điều khiển sẽ được hưởng lợi một cách không cân xứng (Acemoglu et al., 2021; European Parliament, 2020).

Bất bình đẳng không chỉ diễn ra giữa các cá nhân người lao động mà còn xảy ra giữa các quốc gia. Mặc dù AI được dự đoán sẽ thúc đẩy tăng trưởng kinh tế toàn cầu, nhưng hầu hết mức tăng trưởng được dự đoán ở Mỹ và Trung Quốc - hai quốc gia hàng đầu trong lĩnh vực AI. Các quốc gia có thu nhập thấp và trung bình có khả năng chỉ thu được lợi ích khiêm tốn, đồng thời chịu gánh nặng của việc mất việc làm do AI gây ra. Hầu hết các công cụ và ngành công nghiệp AI có khả năng tập trung ở một số quốc gia trong khi các

nước nghèo nhất sẽ có rất ít cơ hội khai thác những công nghệ này cho nền kinh tế của họ. Theo PwC, trong số 15,7 nghìn tỷ USD tài sản mà AI sẽ tạo ra trên toàn cầu vào năm 2030, 70% sẽ được tích lũy bởi Trung Quốc và Mỹ (Hagerty & Rubinov, 2020). Các quốc gia có thu nhập thấp và trung bình đang chứng kiến sự mở rộng nhanh chóng của khu vực lao động phi chính thức do AI điều khiển. Người lao động tại các quốc gia này sẽ trở thành "người lao động thời đại kỹ thuật số" làm công việc nhấp chuột" được trả lương thấp, những người theo dõi và gắn nhãn ảnh để hỗ trợ việc học máy.

3.2.4. Tác động đến giá trị, chuẩn mực xã hội

Nhìn vào lịch sử nhân loại, chúng ta nhận thấy rằng các xã hội đã xây dựng những hệ thống đạo đức công phu nhằm hỗ trợ sự phát triển của con người, nhưng cũng có những hệ thống đạo đức gây ra những nỗi đau khổ sâu sắc. Ngay cả trong xã hội hiện đại thì vẫn còn những xã hội tồn tại những hệ thống đạo đức gây ra những bất bình đẳng giữa các nhóm người, giữa nam và nữ, giữa các tôn giáo. Các khuôn khổ đạo đức đã bắt nguồn từ các xã hội trên toàn thế giới, bao gồm tư tưởng Nho giáo, Thần đạo và Ấn Độ giáo, cũng như các khuôn khổ tôn giáo của Do Thái giáo, Cơ đốc giáo và Hồi giáo, trong số nhiều khuôn khổ khác. Không chỉ có vô số các khuôn khổ đạo đức tồn tại trên khắp các nền văn hóa, mà những biến thể trong những khuôn khổ đạo đức này càng làm tăng thêm tính đa dạng của chúng (khác biệt giữa tư tưởng Công giáo và Tin lành trong Cơ đốc giáo, tư tưởng Sunni và Shiite trong Hồi giáo, tư tưởng Đại Thừa và Tiểu Thừa trong Phật giáo..). Chính vì lẽ đó mà những khái niệm liên quan đến vấn đề đạo đức có thể được hiểu khác nhau giữa các nền văn hoá hay các cộng đồng khác nhau.

Một ví dụ gây ra tranh cãi về vấn đề đạo đức ảnh hưởng đến các chuẩn mực xã hội liên quan đến vấn đề này là mối quan hệ giữa con người và máy móc ở phương diện tình cảm. Những người phản

đối robot tình dục cho rằng chúng sẽ làm suy giảm những giá trị của con người, đặc biệt là phụ nữ và tạo ra những định kiến có hại về phụ nữ phục tùng (Coeckelbergh, 2020). Cũng có những lo ngại rằng việc cho những người thiếu kỹ năng xã hội tiếp cận với robot tình dục sẽ khiến họ không bận tâm đến việc đạt được các kỹ năng xã hội và dần sẽ dẫn đến bị loại trừ trong xã hội. Hay các quan hệ khác giữa con người và robot, liệu máy móc có phải là đối tượng của tình yêu khi xem xét trường hợp Akihiko Kondo kết hôn với Hatsune Miku - một nhân vật tổng hợp giọng nói Vocaloid được minh họa phổ biến xuất hiện dưới dạng hình ba chiều trong một khối hình trụ được gọi là Gatebox.

Vấn đề về đạo đức có thể được hiểu khác nhau giữa các nền văn hoá, hệ tư tưởng và dư luận. Truyền thống phương Tây có xu hướng coi trọng quyền riêng tư của cá nhân nhưng điều này không được coi là vấn đề quá quan trọng ở phương Đông. Nho giáo nhấn mạnh đến lợi ích tập thể nhiều hơn đến lợi ích cá nhân, do vậy quan niệm về quyền riêng tư cá nhân theo truyền thống ít được chú ý hơn và đôi khi mang tính tiêu cực. Lấy ví dụ trường hợp Trung Quốc, do hệ thống chính trị khác nhau, nên vấn đề đạo đức của Trung Quốc có một số điểm khác biệt với quan điểm của các nước phương Tây. Rất nhiều ứng dụng công nghệ AI đã được sử dụng tại Trung Quốc như nhận diện khuôn mặt, thu thập dữ liệu sinh trắc học của người Duy Ngô Nhĩ ở Tây Tạng, hay hệ thống chấm điểm công dân đã gây ra những lo ngại về đạo đức AI của nước này. Bởi mục tiêu chính của chính phủ Trung Quốc là “phúc lợi xã hội” hơn là phúc lợi cá nhân do vậy quyền riêng tư và quyền cá nhân sẽ không phải là mục tiêu ưu tiên hàng đầu (Roberts và cộng sự, 2021).

4. Gợi mở chính sách liên quan đến trí tuệ nhân tạo cho Việt Nam

Tháng 1/2021, Thủ tướng Chính phủ Việt Nam đã ban hành Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng trí tuệ

nhân tạo đến năm 2030 (Chiến lược AI); đặt mục tiêu đẩy mạnh nghiên cứu, phát triển và ứng dụng AI, đưa AI trở thành lĩnh vực công nghệ quan trọng của Việt Nam. Đến năm 2030, phấn đấu đưa Việt Nam vào nhóm danh sách bốn nước dẫn đầu trong Hiệp hội các quốc gia Đông Nam Á (ASEAN) và nhóm 50 nước dẫn đầu trên thế giới về nghiên cứu, phát triển và ứng dụng AI; xây dựng được 10 thương hiệu AI có uy tín trong khu vực. Trên thực tế tại Việt Nam, AI cũng bắt đầu được ứng dụng trong nhiều lĩnh vực như: y tế, giáo dục, nông nghiệp, giao thông và thương mại điện tử. Việt Nam cần có một lộ trình phát triển các chính sách phù hợp tập trung vào những lĩnh vực cụ thể, có kế hoạch tận dụng những cơ hội mà công nghệ mang lại để có thể thực hiện được Chiến lược đã đề ra.

Vai trò của các nhà khoa học xã hội

Chiến lược quốc gia về nghiên cứu phát triển và ứng dụng trí tuệ nhân tạo đến năm 2030 đã đưa ra một số định hướng đầy tham vọng cho Việt Nam trong một thập kỷ tới tận dụng sự phát triển công nghệ để phát triển kinh tế. Tuy nhiên, Chiến lược AI của Việt Nam đã không đề cập đến vai trò của đạo đức trí tuệ nhân tạo trong sự phát triển AI trong tương lai. Nếu không xem xét đến các vấn đề đạo đức quan trọng như quyền riêng tư và giám sát, thành kiến và phân biệt đối xử cũng như vai trò phán xét của con người, sự phát triển của AI trong nước có thể chỉ mang lại lợi ích cho một nhóm nhỏ người và có thể mang lại tác hại cho những người khác (Nga Than, 2021). Hơn nữa, các nhà khoa học xã hội dường như đã bị bỏ quên trong chiến lược này trong khi để đảm bảo thiết kế an toàn, tránh thành kiến thiên lệch và phân biệt đối xử, các nhà phát triển công nghệ rất cần các nhà khoa học xã hội có kiến thức nền tảng từ tâm lý học thực nghiệm, khoa học nhận thức, kinh tế học, khoa học chính trị, tâm lý xã hội, v.v. có kinh nghiệm về nhận thức, hành vi và đạo đức con người, tạo ra AI có trách nhiệm (responsible AI) nhằm tránh hoặc giảm thiểu sự thiên vị, nâng cao tính công bằng, trách nhiệm giải trình.

Quy hoạch phát triển vùng

Với sự phát triển của công nghệ như hiện nay, trong tương lai rất gần, các hệ thống trí tuệ nhân tạo sẽ ảnh hưởng đến đời sống của chúng ta và ngày càng trở nên phổ biến và có sức lan toả mạnh mẽ trong xã hội. Những đột phá về công nghệ dường như sẽ đến nhanh hơn dự kiến, vì vậy cần phải có chính sách cấp thiết, kịp thời để bắt kịp với những thay đổi nhan chóng của xã hội. Trước mắt cần có các giải pháp xây dựng cơ sở hạ tầng công nghệ trí tuệ nhân tạo đáp ứng tốt yêu cầu phát triển khoa học công nghệ. Để làm được vậy cần có lộ trình cụ thể trong từng giai đoạn phát triển cho việc xây dựng nền công nghiệp mới. Cần có một quy hoạch tổng thể về sự phát triển của trí tuệ nhân tạo để định hướng sự phát triển của các vùng, địa phương trên cả nước trên mọi mặt cơ sở hạ tầng, nội dung phát triển công nghệ, địa điểm phát triển, nguồn nhân lực để có kế hoạch đầu tư có mục tiêu vào nghiên cứu và phát triển trên diện rộng. Xây dựng tập trung chuyển đổi cơ sở hạ tầng và chuyển đổi cơ cấu thị trường lao động có giá trị gia tăng thấp sang thị trường sản xuất có giá trị gia tăng cao với các khoản đầu tư có mục tiêu vào nghiên cứu và phát triển về công nghệ trí tuệ nhân tạo. Ngoài ra cũng cần có quy hoạch cụ thể đối với các vùng để có các hướng xây dựng các ngành công nghiệp phụ trợ phục vụ khác nhau liên quan đến các ngành công nghiệp mũi nhọn.

Đầu tư

Trước hết, cần gia tăng đầu tư vào các lĩnh vực trọng tâm liên quan đến trí tuệ nhân tạo như IoT, dữ liệu lớn, an ninh mạng, điện toán đám mây, cơ sở hạ tầng có liên quan. Sự phát triển của AI cần một lượng lớn vốn tài trợ đến từ nhiều nguồn khác nhau, chẳng hạn như các công ty đầu tư mạo hiểm quốc tế, các nhà đầu tư mạo hiểm địa phương, tài trợ của chính phủ hoặc lợi nhuận của chính các công ty. Theo một báo cáo điều tra của EDBI và Kearney, hiện nay mức

đầu tư của Việt Nam vào AI trên đầu người đang ở mức thấp, dưới 1 USD trong khi đó, Mỹ ở mức 155 USD và Trung Quốc là 21 USD (Dylan Loh, 2020). Với mức đầu tư AI trên đầu người thấp sẽ khiến Việt Nam mất đi cơ hội trong cuộc cạnh tranh sắp tới và có thể bị tụt hậu về công nghệ.

Thể chế

Công nghệ trí tuệ nhân tạo sẽ được tích hợp trong mọi mặt của cuộc sống trong thời gian ngắn và sẽ là xu hướng phát triển tương lai, có tác động đa chiều đến mọi mặt của cuộc sống, thay đổi cấu trúc xã hội, cơ cấu lao động cần do đó cần xây dựng các nguyên tắc đạo đức ngay từ khi thiết kế, triển khai và sử dụng, tránh việc thể chế phải chạy theo công nghệ. Cần có những bước chuẩn bị cho việc phát triển các thể chế có liên quan đến công nghệ AI trước khi công nghệ đi vào thực tế và ảnh hưởng đến đời sống con người. Các nguyên tắc này cần có sự thống nhất và được lồng ghép vào các khung pháp lý hiện hành. Một số thể chế cần phải được quan tâm lưu ý trước mắt đó là bảo vệ dữ liệu cá nhân, quyền của người lao động.

Nguồn nhân lực

Cho dù Việt Nam không thiếu nhân tài kỹ thuật, nhưng giáo dục đào tạo AI còn khá mới mẻ. Các trường đại học kỹ thuật hàng đầu Việt Nam mới chỉ tuyển sinh được 1 và 2 khoá đầu tiên về trí tuệ nhân tạo và mới chỉ có 5 cơ sở đào tạo nhân lực trong lĩnh vực này (OECD, 2021). Làm việc trong lĩnh vực này chủ yếu là nhóm sinh viên tốt nghiệp công nghệ thông tin, các nhà khoa học dữ liệu hoặc kỹ sư phần mềm. Hầu hết lực lượng lao động công nghệ vẫn đang làm việc trong lĩnh vực gia công phần mềm. Ngành công nghiệp này đang dựa trên phân khúc tài năng, nơi các tài năng được sử dụng phát huy và nhận được khen thưởng xứng đáng, do đó các tài năng hay các kỹ sư lành nghề thường tìm kiếm các cơ hội nghề nghiệp ở nước ngoài, nơi có mức lương cao hơn đáng kể. Hơn nữa, những

cơ hội này sẽ cho phép họ tham gia tích cực vào việc nghiên cứu, phát triển và triển khai các công nghệ AI hiện đại ở nhiều quốc gia phát triển AI hơn. Chính vì lẽ đó, để thực hiện chiến lược này, cần có những sáng kiến, có đầu tư thích đáng về giáo dục chuyên ngành, tìm kiếm, nuôi dưỡng và duy trì tài nhân lực phục vụ đất nước.

Vai trò chủ thể

Cũng giống như bao Chiến lược khác, Chiến lược phát triển trí tuệ nhân tạo cũng chỉ ra những mục tiêu và vai trò chủ thể thực hiện và phối hợp gồm 16 bộ và Viện Hàn lâm khoa học tự nhiên. Tuy nhiên, Chiến lược đã không chỉ ra những chủ thể khác tham gia vào quá trình phát triển là những doanh nghiệp khởi nghiệp, người thụ hưởng và người dùng phổ biến của công nghệ này. Do vậy, cần xác định rõ vai trò của các chủ thể thực hiện chính sách phát triển trí tuệ nhân tạo là nhà nước hay phi nhà nước. Bởi dẫn đầu trong các hoạt động phát triển AI hiện nay là các các tập đoàn công nghệ lớn trong nước như Vingroup, FPT, VietAI.

Văn hoá và giáo dục

Những thử nghiệm với học máy (machine learning) hiện đang gia tăng và cho thấy tiềm năng to lớn mà phương thức này mang lại trong lĩnh vực âm nhạc, phim ảnh và văn học. Trong tương lai, nhiều sản phẩm văn hóa là kết quả của công nghệ này. Cần sớm có chiến lược cho phát triển văn hóa thời đại AI, tôn trọng các giá trị con người, duy trì bản sắc văn hoá dân tộc. Trong bối cảnh thế giới công nghệ dẫn đầu bởi một số nước lớn như Mỹ, Trung Quốc, châu Âu, nếu thiếu yếu tố văn hoá trong chiến lược phát triển AI sẽ khiến chúng ta mất dần văn hoá dân tộc và điều này cũng sẽ làm tổn hại đến cấu trúc xã hội. Ngoài ra, cũng cần có ưu tiên phát triển giáo dục đào tạo thích nghi với văn hoá của thời đại AI, khuyến khích một xã hội học tập, trang bị kỹ năng số, có khả năng giải quyết vấn đề nhanh và sáng tạo.

Nhận thức của công chúng về AI

Cần nâng cao nhận thức của công chúng về AI. Tại Việt Nam, AI được nhìn nhận một cách tích cực. Nó được coi là động lực xúc tác cho tiến bộ kinh tế và công nghệ. Trong tâm trí công chúng, khái niệm về AI là gì, nó được sử dụng như thế nào và nó ảnh hưởng đến ai vẫn chưa rõ ràng. Chưa nhận ra rằng công nghệ AI đã hiện hữu trong cuộc sống, trong các hoạt động cá nhân của họ như Grab, Tiki đều do thuật toán điều hướng các hoạt động. Do thúc đẩy số hóa và công nghiệp 4.0, người Việt Nam coi AI chỉ là một công cụ dành riêng cho các ngành công nghiệp như một số hoạt động xử lý ngôn ngữ tự nhiên và thị giác máy tính được sử dụng cho các mục tiêu kinh doanh xa hơn.

Nhận thức sâu sắc về AI là rất cần thiết vì nó mở rộng quan điểm của công chúng về vai trò của AI trong việc mang lại lợi ích hoặc gây hại cho cuộc sống của họ. Trong đại dịch hiện nay, các vấn đề ứng dụng công nghệ AI như máy ATM gạo, Bluezone, QR code về tiêm chủng vắc xin, thẻ xanh… là những bước khởi đầu trong việc giải quyết các vấn đề xã hội cấp thiết trước mắt mà lu mờ những vấn đề có ảnh hưởng lâu dài hơn đến quyền riêng tư, an ninh mạng…

Đạo đức của trí tuệ nhân tạo đang chưa thu hút được sự chú ý của truyền thông và các cuộc thảo luận chính sách công cũng như vắng bóng trong việc đào tạo kỹ thuật. Những gì sinh viên học ở các trường đại học khối kỹ thuật mới chỉ được đào tạo về đạo đức trong khoa học máy tính. Các trường cao đẳng và đại học nên đầu tư vào việc không chỉ học hỏi từ việc học và giảng dạy chương trình này, cũng nên đầu tư vào việc nghiên cứu, đặc biệt là khoa học xã hội để xem xét các tác động xã hội của công nghệ ở Việt Nam (Nga Than, 2021).

Cuối cùng, xây dựng và phát triển AI có lợi cho xã hội, lấy con người làm trung tâm là mục tiêu cuối cùng mà chúng ta cần hướng

đến. Bởi bất kỳ công nghệ nào được áp dụng để thực hiện các mục đích của một nhóm người cụ thể, sẽ không phải lúc nào cũng có thể loại bỏ thành kiến, vì nó là một phần của sự khác biệt văn hóa. Thành kiến và khuôn mẫu không chỉ được nhúng trong dữ liệu hoặc trong các thuật toán, mà còn trong cách mà các chính phủ, công ty và người dùng tương tác với máy móc. Do đó, điều cần thiết là phát triển các chiến lược vượt ra ngoài quy tắc đạo đức trừu tượng đơn thuần và thiết kế các chính sách công để đảm bảo rằng các hệ thống AI - và các tác nhân khai thác chúng - có thể kiểm tra được và có trách nhiệm giải trình.

Tài liệu tham khảo

1. Acemoglu, D., Restrepo, P., Autor, D., Gancia, G., Lemieux, T., Rogerson, R., Rossi-Hansberg, E., & Ross, S. (2021). *Tasks and Inequality*. http://www.nber.org/papers/w28920

2. Andrew McAfee and Erik Brynjolfsson. (2017). *Machine, Platform, Crowd: Harnessing Our Digital Future*. W. W. Norton.

3. Barr, A. (2015). Google mistakenly tags black people as ‘gorillas,’ showing limits of algorithms. *The New York Times*.

4. Barrett, D. (2014). Holder Cautions on Risk of Bias in Big Data Use in Criminal Justice. *Wall Street Journal*. https://www.wsj.com/articles/u-s-attorney-general-cautions-on-risk-of-bias-in-big-data-use-in-criminal-justice-1406916606

5. Bass, D. & Huet, E. (2017). Researchers Combat Gender and Racial Bias in Artificial Intelligence. *Bloomberg*. https://www.bloomberg.com/news/articles/2017-12-04/researchers-combat-gender-and-racial-bias-in-artificial-intelligence

6. Bossmann., J. (2016). *Tope 9 Ethical Issues in Artifical Intelligence*.

7. Coeckelbergh, M. (2020). *AI ethics*. The MIT Press.

8. Dylan Loh. (2020). ASEAN faces wide AI gap as Vietnam and Philippines lag behind. *Nikkei Asia*. https://asia.nikkei.com/Business/Technology/ASEAN-faces-wide-AI-gap-as-Vietnam-and-Philippines-lag-behind

9. Eliane Röösli, Brian Rice, T. H.-B. (2021). Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19. *Journal of the American Medical Informatics Association*, *28*(1), 190-192. https://doi.org/https://doi.org/10.1093/jamia/ocaa210

10. European Parliament. (2020). *The ethics of artificial intelligence: Issues and initiatives - Think Tank*. https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2020)634452

11. Fabio Chiusi, Sarah Fischer, Nicolas Kayser-Bril and Matthias Spielkamp. (2020). *Automating Society Report 2020* (N. K.-B. and M. S. Fabio Chiusi, Sarah Fischer (ed.)). AlgorithmWatch gGmbH and Bertelsmann Stiftung. https://automatingsociety.algorithmwatch.org

12. Floridi, L. and et al. (2018). AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*, 689-707. https://doi.org/10.1007/s11023-018-9482-5

13. Hagerty, A., & Rubinov, I. (2020). Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence, *https://Arxiv.Org/*.

14. Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. https://doi.org/10.5281/zenodo.3240529

15. Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. In *The Stanford Encyclopedia of Philosophy*.

16. OECD (2021). AI Strategies and Policies in Vietnam. OECD.AI. Truy cập ngày 1 tháng 10 năm 2021.

17. Roberts, H., Cowls, J., Morley, J. et al. (2021). The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & Soc*, *36*, 59-77. https://doi.org/10.1007/s00146-020-00992-2

18. Soenens and et al. (2007). Conceptualizing parental autonomy support: adolescent perceptions of promotion of independence versus promotion of volitional functioning. *Developmental Psychology*, *43*(3), 633-646. https://doi.org/10.1037/0012-1649.43.3.633

19. Thủ tướng chính phủ (2021). *Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng trí tuệ nhân tạo đến năm 2030*.

20. UNESCO. (2020). *Elaboration of a Recommendation on the ethics of artificial intelligence*. https://en.unesco.org/artificial-intelligence/ethics#recommendation

ETHICAL ISSUES OF ARTIFICIAL INTELLIGENCE AND ITS SOCIAL IMPLICATIONS

Dr. Hoang Vu Linh Chi
Institute of Sociology, Vietnam Academy of Social Sciences

1. Introduction

Before the emergence of Artificial Intelligence (AI), humans were living in a binary space consisting of physical and social spaces. In this binary space, the order for human activities was determined by the human-to-human and human-to-object interactions, in which people played the role of the shaper and the ruler. With the rapid development of metadata, cloud computing, Internet of things, smart mobile devices in various fields of human society, which promote the birth of the intelligent era, people are taken to a new space marked by physical space, human social space and cyberspace. In the cubic space, the order of human society will constantly be restructured. Whether people are aware of that change or not, the profound effects on human social life that AI brings will gradually be accepted in all walks of life.

There is an urgent need to address the challenges posed by AI, related to biased algorithms, protection of people's privacy, personal data, the risk of new forms of exclusion and inequality, issues on sharing of benefits and risks, accountability, impact on jobs and the future of work, human rights, security and risks of AI usage, which are the concerns of many organizations and governments interested in the ethical implications of artificial intelligence.

2. AI and ethics of AI

Artificial intelligence (AI) can be defined as a collection of different technologies that allow machines to act with an intelligence

level like a human being. This includes learning the rules needed to make certain decisions and reasons for certain conclusions, as well as learning from past experiences and self-correction. Russel and Norvig, the authors of "Artificial Intelligence: a Modern Approach", defined AI as "*the study and design of intelligent agents, where an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success*". According to Russell and Norvig, there are currently four approaches to AI. The first approach focuses on creating machines that think like humans. With this approach, researchers try to somehow reproduce processes, manifestations and results of human thinking onto machines. The second approach focuses on creating machines that act like humans. It focuses on actions like the robots today, not on the process to come to these actions. The third approach focuses on developing machines that act rationally. Rationality is closely related to optimization. Finally, the fourth approach focuses on developing machines that think rationally. The planning or decision making will be done optimally, which relates to the problem that the system is trying to solve.

The most common fundamental element in all approaches is that AI consists of the research, design, and construction of intelligent agents that can achieve goals. A flexible AI with the ability to learn from experiences as well as consciousness, to plan and act based on its configuration, can be evaluated as smarter than an AI that has only one set of rules to guide some fixed actions. Depending on different contexts and approaches, inventors tend to emphasize some factors more than others. For example, expert system developers see AI as a storehouse of expert knowledge that humans can refer to, while machine learning system developers see AI as something that can discover new knowledge. As we can see, each approach has its own strengths and weaknesses.

AI scientists and experts agree that AI has rapidly and significantly influenced human society as well as the way humans interact with each other. During its development, AI presents major ethical and socio-political challenges that require a thorough ethical and philosophical analysis. Its social impact needs to be researched to avoid any negative effects. We are in a race against time to establish ethical principles before technology catches up with people. There have been lots of arguments about the duality of artificial intelligence. The rapid development of artificial intelligence and machine learning offers huge potential benefits including its positive impacts on several fields such as health, education, space, etc. However, some scientists concern about the negative side of artificial intelligence in its deadly race.

According to Mark Coeckelbergh, the ethics of AI is about technological change and its impact on individual lives, and also the transformations in society and the economy. Issues of prejudice and discrimination have shown that AI is involved and would influence the changes in the economy and subsequently the social structure of societies (Coeckelbergh, 2020). According to Brynjolfsson and McAfee (2014), we have entered the Second Machine Age, in which machines are not only human aids, as in the Industrial Revolution, but also human substitutes. When all fields and jobs are affected by AI, our society is predicted to change dramatically as technologies depicted in science fiction enter the real world (Andrew McAfee and Erik Brynjolfsson, 2017).

3. The relationship between artificial intelligence ethics and social impacts

3.1. Artificial intelligence ethical concerns

Artificial intelligence could bring huge benefits, supporting human activities, substituing humans in doing difficult or boring

jobs. Professor Andrew Ng., the founder of Coursera said “As the Industrial Revolution liberated the humankind from physical toil, I think AI could potentially liberate the humankind from mental toil.” Artificial intelligence could be used to improve public and commercial services. For example, image recognition helps diagnose diseases like cancer and Alzheimer. However, AI also generates many risks such as underestimating human skills, refusing responsibility, reducing control and eroding human self-determination (Floridi, 2018). There are unprecedented challenges related to ethical issues because of the great potential of AI technologies in all fields that can bring huge profits to companies with access to customer data.

The World Economic Forum has raised the top 9 ethical issues in AI: 1. Unemployment. What happens after the end of jobs?; 2. Inequality. How to distribute the wealth created by machines?; 3. Humanity. How do machines affect human behaviour and interaction?; 4. Artificial stupidity. How to avoid mistakes caused by AI?; 5. Racist robots. How do we eliminate AI bias?; 6. Security. How do we keep AI safe from adversaries?; 7. Evil genies. How can we be protected against unintended consequences?; 8. Singularity. How can we control a complex intelligent system?; 9. Robot rights. How do we define the humane treatment of AI? (Bossmann., 2016).

The Alan Turing Institute listed a wide range of potential harms caused by AI systems. (1) Bias and discrimination**,** (2) Denial of individual autonomy and rights**,** (3) Non-transparent, unexplainable or unjustifiable outcomes**,** (4) Invasion of privacy**,** (5) Isolation and disintegration of social connection**,** (6) Unreliable, unsafe, or poor-quality outcomes (Leslie, 2019). Meanwhile, Mueller listed 10 main AI problems arising in the use of some AI technologies: (1) Privacy & Surveillance; (2) Manipulating of behavior; (3) Opacity of AI system; (4) Bias in decision system; (5) Human-Robot interaction;

(6) Automation and Employment; (7) Autonomous system; (8) Machine ethics; (9) Artificial ethical agent; (10) Singularity. The author argues that in addition to "responsible use", it is also necessary to mention "responsible design" in this area (Müller, 2020).

Common debates regarding the ethical issues of AI:

- How do we align the aims of autonomous AI systems with our own?

- Does the future of AI pose an existential threat to humanity?

- How do we prevent learning algorithms from acquiring morally objectionable biases?

- Should autonomous AI be used to kill in warfare?

- How should AI systems be embedded in our social relations? Is it permissible to fall in love with an AI system?

- What sort of ethical rules should AI like a self-driving car use?

- Can AI systems suòer moral harms? And if so, of what kinds?

- Can AI systems be moral agents? If so, how should we hold them accountable?

- How should we live with and understand minds that are alien to our own?

In addition, debates are raised depending on different approaches, but in general the questions focus on what should be considered correct/ incorrect in the development and implementation of AI and robot. For example, the European Commission requires ethics in developing and using new technologies in its programmes, as well as forming ethical guidelines related to fundamental human rights. The European Commission's plan also aims to set a global ethical standard towards becoming the world's leading guide for trustworthy and ethical AI.

From the viewpoint of the United Nations Educational, Scientific and Cultural Organization (UNESCO), international and national policies and regulatory frameworks are required to ensure that these emerging technologies bring benefits for humanity. AI shall be human-centered and for the greater benefits of human being, rather than vice-versa. UNESCO recommends a comprehensive global standard-setting instrument to provide AI a strong ethical foundation that not only protects but also promotes human rights and human dignity. If adopted, it will be an ethical guiding compass and a global normative base allowing to buil strong respect for the rule of law in the digital world (UNESCO, 2020).

No matter what we think about these wishes, it is obvious that there is a lack of practical guidelines and rules, and a need for consensus on autonomous systems. Today, most jurisdictions worldwide are just beginning to investigate the regulatory aspects of AI. Numerous research institutes on AI ethics and its impact have been established within many famous universities worldwide. In addition, large technology organizations and companies are also gradually establishing principles and job positions related to AI ethics that are intended for their products and applications. Thus, ethics becomes important at many levels of policy discussion. It's a topic for system-design engineers, including students learning to build AI systems. This is also a topic for society to assess the impacts of AI on people's daily lives. Ethics becomes an important question for policymakers in discussions about AI and robotics. It should be noted that ethical aspects are not only discussed in regulatory questions. Much more importantly, the ethical questions underpin the design of AI systems from application definition to implementation details. Therefore, ethics in AI is much broader and involves fundamental design choices and considerations about the society we want to live in.

3.2. *Social impacts of artificial intelligence*

3.2.1. AI impacts on human autonomy

Autonomy involves acting in accordance with one's goals and values, which is distinct from the use of autonomy as simply a synonym for either independence or being in control autonomy involves acting in accordance with one's goals and values, which is distinct from the use of autonomy as simply a synonym for either independence or being in control autonomy involves acting in accordance with one's goals and values, which is distinct from the use of autonomy as simply a synonym for either independence or being in control

Autonomy involves acting in accordance with one's goals and values, which is distinct from the use of autonomy simply as a synonym for either independence or being in control (Soenens et al., 2007). While human autonomy has long been important to philosophy and the social sciences, in the field of engineering and computer science there has been a tendency to focus on machine autonomy. A classic tenet of bioethics is the principle of autonomy: individuals have the right to make their own decisions. With AI, the situation becomes more complicated: when we apply AI and its intelligent agents, we are willing to cede part of our decision-making power to the machines. Thus, asserting the principle of autonomy in the context of AI means striking a balance between the decision-making power we retain to ourselves and the rights we delegate to artificial agents.

One of the AI effects on human autonomy involves decision making. As AI systems become more autonomous and increasingly replace humans in behavior decision making, we are threatened of losing our ability to make the rules of life, make decisions or shape our life. For example, we can consult with an AI medical advisor (see

Ada.com). We can choose to give up our ability and willingness to know and understand our own bodies and illnesses, in favor of an automated system that replaces the general practitioner who manages the illness, health status and make important decisions regarding our health. An AI dietitian could pose a similar threat, making requests about what they should eat and what they should avoid. Moreover, AI technology may really force the execution of this diet through smart refrigerators. People might give up managing their diets to such a system/application and lose their ability to take care of themselves or manage their own diet due to increasing dependence on these systems. Gradually, people will lose the ability to act autonomously.

In addition, manipulation and deception are clearly cases where human autonomy is impaired due to the use of AI systems. How AI systems can influence and shape our interests and beliefs? Several studies indicate how people change their preferences according to appeared recommendations of algorithm such as the cases of Youtube, Cambrige Analytica.

The concerns about control refers not only to autonomous AI systems, but also to algorithms that make their own decisions. The problem is raised when humans or users cannot request the re-execution of decision making process or it is very time consuming to change the decision that the machine/technology has made.

3.2.2. Increasing bias and discrimination

When algorithms are used in decision-making instead of human, it is not surprising that sometimes there are the risk of recreating existing biases and inequalities in society. As a result, unfair rules are inadvertently hard-coded into AI systems, thus leading to discriminatory decisions. There are many examples of biased algorithms since decisions were based on big data. While social

values, norms and attitudes towards race and gender are still evolving, their imprint remains on the internet from which most of the training data for machine learning algorithms is derived. For instance, studies demonstrated that 'men' are associated with bosses, presidents, leaders and directors, while 'women' are associated with helpers, assistants and employees. As a result, society's push for equity and more equity has stumbled upon historical values of poverty, gender, race and ethnicity locked in big data (Barr, 2015; Barrett, 2014; Bass, D. & Huet, 2017). Typically, the case of Google mistakenly labeling Africans as gorillas has created a debate about the ethics of AI. In addition, developers and data scientists have said that bias often leads to skewed models. For example, Google's shoe recognition algorithms or the disproportionate impact of Covid-19 pandemic on colored people in accessing healthcare service (Eliane Röösli, Brian Rice, 2021) has shown the limitations of AI algorithms. One of the reasons for this bias is that programmers live in a white-owned technology world, they lack information and data about other races or data from men and the decisions made by algorithms are ultimately not optimal.

Robot design, manufacturing and programming are complex. The focused ethical issue is the traceability of robot. Traceability is the ability that help the robot to track the causes of all past actions. For robots with self-determination, decision-making and self-learning capabilities, the traceability is a concern because these robots are not just programmed to do specific jobs, but they can also learn and develop themselves while interacting with the environment. It generates a big gap in the current ethical framework for this traceability of robots.

One other factor is the technology solution pitfall. It is the fact that digital technology has completely changed the way we live, work and play, however, we cannot justify the irrational abuse and massive

application of technological solutions in our social lives. Some researchers are concerned about the emerging idea that all social problems can be "fixed" through technology. Especially in the Covid era, face recognition technologies or technologies applied in administrative activities are increasingly used, which has become a topic of debate in many European Union member countries about the implementation and application of ADM systems in human life. Because the applications of AI are completely “black box” and if it is not possible to know exactly what the ADM system does and why, then when the applications ultimately are deployed, all other efforts to balance between the system and basic human rights will never be achieved (Fabio Chiusi, Sarah Fischer, Nicolas Kayser-Bril and Matthias Spielkamp, 2020).

Bias and discrimination are inherent risks of any economic or social activity, so human decision-making is not immune to mistakes and prejudices. However, if the same bias is presented in AI, there could be a much larger impact, affecting and discriminating against many people if there are no social control mechanisms governing human behavior. Undesirable effects do not occur and cannot be predicted at the design stage, but due to the actual effects of correlations or models that the system identifies when based on big data.

3.2.3. Impacts on inequality

Since previous industrial revolutions, there is concern that automation will displace existing jobs and cause damage to the labor market. Employment changes related to automation and digitization can be observed not only in job losses but also in job quality, as AI is expected to create new and diverse types of employment. The change could mean millions of workers losing their jobs and working time. Some people may lose their jobs permanently, and some will have to switch jobs when their jobs are replaced by AI. The analyses

predict that AI and other new technologies will continue to benefit higher-skilled workers, while low and medium-skilled workers will face pressure to lose their jobs because of technology. The unemployed will lose their ability to get income, which creates more inequality in society where incomes are increasing due to rising per capita incomes thanks to higher productivity of AI technology (McAfee and Brynjolfsson, 2017). In a recent study by NBER (USA), automation technology has been the main driver of income inequality in the United States over the past 40 years. Artificial intelligence has caused a 50% to 70% decline in wages in the US, since 1980, possibly due to a drop in the wages of blue-collar workers who have been displaced or decimated by automation. Artificial intelligence, robotics and new complex technologies have caused huge wealth and income inequality. It seems likely that this problem will accelerate, with post-graduate degree holders seeing a rise in their wages, while salary of workers with low education level falling dramatically. Real income for men without a high school diploma is now 15% lower than it was in 1980 (Acemoglu et al., 2021).

AI and robotics technologies are expected to enable companies to optimize their business operations, reduce marginal costs, and operate more efficiently and productively. Thus, revenue will be divided among fewer people, which expands social inequality. As a result, individuals holding ownership in AI-driven companies will benefit disproportionately (Acemoglu et al., 2021; European Parliament, 2020).

Inequality does not only occur between individual workers but also countries. While AI is predicted to drive global economic growth, most of the growth is predicted in the US and China - the two leading countries in the field of AI. Low and middle-income countries are likely to reap only modest benefits while bearing the brunt of AI-induced job losses. Most of the AI instruments and industries are

likely concentrated in a few countries while the poorest countries will have little chance of harnessing these technologies for their economies. According to PwC, of the $15.7 trillion in assets that AI will create globally by 2030, 70% will be accumulated by China and the US (Hagerty & Rubinov, 2020). Low and middle-income countries are seeing a rapid expansion of the AI-driven informal labor sector. Workers in these countries will become low-paid "digital clickers" who monitor and label photos to aid machine learning.

3.2.4. Impacts on social values and standards

Looking at human history, we find that societies have built elaborate ethical systems to support human development, but there are also ethical systems that cause deep suffering. Even in modern society, there are still societies with ethical systems that cause inequalities between groups, men and women, and between religions. Ethical frameworks have taken root in societies around the world, including Confucian, Shinto and Hindu, as well as the religious frameworks of Judaism, Christianity and Islam, among many other frameworks. Not only do a multitude of ethical frameworks exist across cultures, but variations in these frameworks add to their diversity (difference between Catholic and Protestant thought in Christianity, Sunni and Shia thought in Islam, Mahayana and Hinayana thought in Buddhism, etc.). It is the reason why ethical concepts can be understood differently across cultures or communities.

One controversial example of how ethics affects social norms is the emotional relationship between human and machine. Opponents of sex robots argue that they degrade human values, especially women, and create harmful stereotypes about submissive women (Coeckelbergh, 2020). There are also concerns that giving access to sex robots for people lacking social skills will distract them from

gaining social skills and eventually lead to social exclusion. There may be other human-robot relationships, whether machines are objects of love, considering the case of Akihiko Kondo marrying Hatsune Miku, a popular illustrated Vocaloid voice synthesizer appearing as a hologram in a cylinder called a Gatebox.

Ethics can be understood differently across cultures, ideologies and public opinion. Western traditions tend to respect individual privacy, but it is not so important in the East. Confucianism emphasizes collective interests more than individual interests, so the concept of individual privacy has traditionally received less attention and sometimes negative meaning. Take China as an example, because of different political systems, perspectives of some ethical issues of China differ from Western countries. Many applications of AI technology used in China such as facial recognition, biometric data collection of the Uighurs in Tibet, or a citizen scoring system have raised concerns about the AI ethics in this country. Since the main goal of the Chinese government is “social welfare” rather than individual welfare, privacy and individual rights will not be top priority (Roberts et al., 2021).

4. AI-related policy recommendations for Viet Nam

In January 2021, the Prime Minister of Viet Nam issued the National Strategy on Research, Development and Application of Artificial Intelligence to 2030 (AI Strategy); aims to promote research, development and application of AI, making AI an important technology of Viet Nam. By 2030, strive to put Viet Nam among top four countries in the Association of Southeast Asian Nations (ASEAN) and in the group of world's top 50 countries in research, development and application of AI; Setting up 10 reputable AI centers in the region. In reality, AI has also begun to be applied in many fields in Viet Nam such as healthcare, education, agriculture, transportation

and e-commerce. Viet Nam needs to have a roadmap to develop appropriate policies focusing on specific areas, a plan to take advantage of opportunities that technology brings in order to implement the proposed Strategy.

The role of social scientists

The National Strategy for Research, Development and Application of Artificial Intelligence to 2030 has set out some ambitious directions for Viet Nam over the next decade to utilize technological developments for economic development. However, Viet Nam's AI Strategy did not mention the role of AI ethics in the future of AI development. Without taking into account important ethical issues such as privacy and surveillance, bias and discrimination, and human judgement, the domestic development of AI may only be beneficial to a small group of people and bring harm to others (Nga Than, 2021). Moreover, social scientists seem to have been neglected in this strategy while to ensure secure design, avoid bias and discrimination, technology developers desperately need to consult social scientists with background knowledge from experimental psychology, cognitive science, economics, political science, social psychology etc., and experience in cognition, behavior and ethics of people, in order to create responsible AI, avoid or minimize bias, improve fairness and accountability.

Regional development plan

With the current development of technology, in the very near future, AI systems will affect our lives, become increasingly popular and pervasive in society. Technological breakthroughs seem to come sooner than expected, so it is necessary to have urgent and timely policies to keep up with the rapid changes of society. In the immediate future, it is necessary to have solutions to build artificial intelligence technology infrastructure to meet the requirements of

science and technology development. To do so, a specific roadmap is required in each development stage for building a new industry. There should be a master plan on the development of artificial intelligence to orient the development of regions and localities across the country in all aspects of infrastructure, contents, locations, human resources for targeted investments in research and development on a large scale. Focus on transforming infrastructure and restructuring low-value-added labor markets into high-value-added manufacturing markets with targeted investments in research and development of artificial intelligence technology. In addition, it is also necessary to have specific plans for regions in order to have directions for building different supporting industries for key industries.

Investment

First of all, it is essential to increase investment in key areas related to artificial intelligence such as IoT, big data, cybersecurity, cloud computing and relevant infrastructure. The development of AI requires a large amount of funding coming from various sources, such as international venture capital firms, local venture capitalists, government funding or corporate self-profits. According to a survey by EDBI and Kearney, currently, Viet Nam's per capita investment in AI is at a low level, below 1 USD, while it is 155 USD in the US and 21 USD in China (Dylan Loh, 2020). With such low per capita investment in AI, Viet Nam will lose opportunities in the upcoming competition and may lag behind in technology.

Institutions

Artificial intelligence technology will be integrated in all aspects of life in a short time and will be a development trend in the future with multi-dimensional impacts on all aspects of life, changing social and labor structure. Therefore, it is necessary to develop ethical principles right from the design, implementation and use to avoid the

need for institutions to follow technology. Preparatory steps are needed for the development of institutions related to AI technology before the technology becomes practical and affects people's lives. These principles need to be unified and integrated into existing legal frameworks. Some institutions that need immediate attention are the protection of personal data and the rights of workers.

Human resources

Although there is no shortage of technical talents in Viet Nam, AI education and training is still quite new. The leading technical universities in Viet Nam have only recruited 1 or 2 first courses on artificial intelligence and there are only 5 human resource training institutions in this field (OECD, 2021). People working in this field are mainly among information technology graduates, data scientists or software engineers. Most of the tech workforce is still working in software outsourcing. This industry is based on the talent segment, where the talent is used, promoted and rewarded deservely, so the talent or skilled engineers are often looking for career opportunities abroad where imcome are significantly higher. Furthermore, these opportunities will enable them to actively participate in the research, development and implementation of modern AI technologies in more AI developed countries. Therefore, to implement this strategy, it is necessary to have appropriate initiatives and investmenst in specialized education, to search, nurture and maintain human resources to serve the country.

Role of actors

Like many other Strategies, the AI Development Strategy also outlines the goals and roles of the implementing and coordinating actors, including 16 ministries and the Viet Nam Academy of Sciences and Technology. However, the Strategy does not identify other actors involved in the development process such as startups,

beneficiaries and popular users of this technology. Therefore, it is necessary to clearly define the role of actors implementing policies on artificial intelligence development as state or non-state. Large domestic technology corporations such as Vingroup, FPT and VietAI are currently leading in AI development activities.

Culture and education

Experiments with machine learning are now on the rise and show the huge potential in the fields of music, film and literature. In the future, many cultural products will be the outcome of this technology. It is necessary to soon have a strategy for cultural development in the AI era, respecting human values and maintaining national cultural identity. In the context of the technology world led by some big countries like the US, China and Europe, the lack of cultural factors in the AI development strategy will gradually lessen our national culture and damage the social structure. In addition, priority should be given to developing education and training that is adapted to the culture of the AI age, encouraging a learning society, equipped with digital skills and capable of fast and creative problem solving.

Public perception of AI

There is a need to raise public awareness of AI. In Viet Nam, AI is viewed positively. It is considered to be the catalyst for economic and technological progress. In the public mind, the concept of what AI is, how it is used, and who it affects are still unclear. A part of citizen hasn’t realized that AI technology is already present in their lives, their personal activities like Grab, Tiki, of which activities are all navigated by algorithms. Due to the promotion of digitization and industry 4.0, the Vietnamese see AI as just an instrument dedicated to industries, for instance, natural language processing and computer vision activities are used for further business purposes.

In-depth awareness of AI is essential as it broadens the public's view of the role AI plays in benefiting or harming their lives. During the current pandemic, applications of AI technology such as rice ATMs, Bluezone, QR codes on vaccinations, green cards, etc. are the early steps in solving urgent immediate social problems, which can overshadow issues that have a more lasting impact on privacy, cybersecurity…

The AI ethics is under-represented in the media and public policy discussions and absent in technical training. What students learn about ethics at technical universities is only in computer science. Colleges and universities should invest in not only learning and teaching, but also research, especially in social sciences, to consider social impacts of technology in Viet Nam (Nga Than, 2021).

Finally, building and developing AI that benefits society, taking people as the center is our ultimate goal. Because any technology that applies to fulfill the purposes of a particular group of people will not always be possible to eliminate bias, as it is part of cultural differences. Bias and stereotypes are not only embedded in data or in algorithms, but also in the way governments, companies and users interact with machines. It is therefore essential to develop strategies that go beyond mere abstract ethics and design public policies to ensure that AI systems - and the actors that exploit them - are checkable and accountable.

References

1. Acemoglu, D., Restrepo, P., Autor, D., Gancia, G., Lemieux, T., Rogerson, R., Rossi-Hansberg, E., & Ross, S. (2021). *Tasks and Inequality*. http://www.nber.org/papers/w28920

2. Andrew McAfee and Erik Brynjolfsson. (2017). *Machine, Platform,*

Crowd: Harnessing Our Digital Future. W. W. Norton.

3. Barr, A. (2015). Google mistakenly tags black people as 'gorillas,' showing limits of algorithms. *The New York Times*.

4. Barrett, D. (2014). Holder Cautions on Risk of Bias in Big Data Use in Criminal Justice. *Wall Street Journal*. https://www.wsj.com/articles/u-s-attorney-general-cautions-on-risk-of-bias-in-big-data-use-in-criminal-justice-1406916606

5. Bass, D. & Huet, E. (2017). Researchers Combat Gender and Racial Bias in Artificial Intelligence. *Bloomberg*. https://www.bloomberg.com/news/articles/2017-12-04/researchers-combat-gender-and-racial-bias-in-artificial-intelligence

6. Bossmann., J. (2016). *Tope 9 Ethical Issues in Artifical Intelligence*.

7. Coeckelbergh, M. (2020). *AI ethics*. The MIT Press.

8. Dylan Loh. (2020). ASEAN faces wide AI gap as Viet Nam and Philippines lag behind. *Nikkei Asia*. https://asia.nikkei.com/Business/Technology/ASEAN-faces-wide-AI-gap-as-Vietnam-and-Philippines-lag-behind

9. Eliane Röösli, Brian Rice, T. H.-B. (2021). Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19. *Journal of the American Medical Informatics Association*, *28*(1), 190-192. https://doi.org/10.1093/jamia/ocaa210

10. European Parliament. (2020). *The ethics of artificial intelligence: Issues and initiatives - Think Tank*. https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2020)634452

11. Fabio Chiusi, Sarah Fischer, Nicolas Kayser-Bril and Matthias Spielkamp. (2020). *Automating Society Report 2020* (N. K.-B. and M. S. Fabio Chiusi, Sarah Fischer (ed.)). AlgorithmWatch gGmbH and Bertelsmann Stiftung. https://automatingsociety.algorithmwatch.org

12. Floridi, L. and et al. (2018). AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*, 689-707. https://doi.org/10.1007/s11023-018-9482-5

13. Hagerty, A., & Rubinov, I. (2020). Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence. https://Arxiv.Org/.

14. Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. https://doi.org/10.5281/zenodo.3240529

15. Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. In *The Stanford Encyclopedia of Philosophy*.

16. OECD (2021). AI Strategies and Policies in Viet Nam. OECD.AI. Truy cập ngày 1 tháng 10 năm 2021.

17. Prime Minister (2021). *National strategy on research, development and application of artificial intelligence to 2030*

18. Roberts, H., Cowls, J., Morley, J. et al. (2021). The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & Soc*, *36*, 59-77. https://doi.org/10.1007/s00146-020-00992-2

19. Soenens and et al. (2007). Conceptualizing parental autonomy support: adolescent perceptions of promotion of independence versus promotion of volitional functioning. *Developmental Psychology*, *43*(3), 633-646. https://doi.org/10.1037/0012-1649.43.3.633

20. UNESCO. (2020). *Elaboration of a Recommendation on the ethics of artificial intelligence*. https://en.unesco.org/artificial-intelligence/ethics#recommendation

Phần 2: Xây dựng quy chuẩn về đạo đức cho Trí tuệ Nhân tạo

Part 2: Problems and Mapping Possible Norms for AI

ĐẠO ĐỨC AI VÀ GỢI Ý CHÍNH SÁCH CHO VIỆT NAM

PGS.TS. Cao Thu Hằng
Tạp chí Cộng sản

1. Lịch sử phát triển xã hội loài người cho thấy, khoa học, kỹ thuật và công nghệ có vai trò quan trọng đối với sự phát triển nhân loại. Mỗi bước tiến của khoa học, kỹ thuật, công nghệ dẫn đến những bước tiến trong các lĩnh vực kinh tế, xã hội. Những năm gần đây, sự phát triển vượt bậc của các nhân tố này đã tạo ra những bước tiến đáng kể cho sự phát triển xã hội, đặc biệt là việc ứng dụng các thành tựu của chúng trong các lĩnh vực của đời sống xã hội, trong đó có thể kể đến trí tuệ nhân tạo. Trí tuệ nhân tạo là trí tuệ của máy móc do con người lập trình nên, có cơ chế hoạt động gần giống như bộ não người. Nó có thể triển khai dưới dạng hệ thống dữ liệu, cũng có thể được ứng dụng vào thiết bị nào đó, và qua đó, thiết bị đó có thể hoạt động như một bộ não của con người (ứng dụng công nghệ AI). Khác với các chương trình phần mềm khác chỉ thực hiện các thao tác do con người lập trình, đặc trưng của AI là khả năng tự học tương tự như con người, cải thiện cách thức hoạt động, đáp ứng khi nhận được thông tin phản hồi từ môi trường bên ngoài trong các tình huống, cho phép giải quyết vấn đề trong tình huống thiếu thông tin quan trọng hoặc xử lý tình huống mới[1].

Với đặc trưng đó, có thể thấy, trí tuệ nhân tạo đã tạo ra một cuộc cách mạng trong xã hội, tạo điều kiện thuận lợi cho việc thay đổi cách thức sản xuất, thay đổi vai trò của con người trong sản xuất. Nếu như trước kia người lao động là tác nhân chủ yếu trong dây chuyền sản xuất, thì ngày nay, các rô-bốt với việc ứng dụng công nghệ AI đã thay thế; việc sử dụng công nghệ AI, tự động hóa không

chỉ tạo ra bộ máy sản xuất gọn nhẹ, mà còn tạo ra bộ máy quản lý cũng gọn nhẹ theo. Trong lĩnh vực kinh tế, nhờ trí tuệ nhân tạo, người ta nắm bắt tốt hơn thị hiếu khách hàng thông qua việc tổng hợp đánh giá hành vi cá nhân của từng người. Người ta cũng có thể dùng AI để quảng cáo theo mục tiêu bằng việc thu nhận hoạt động của con người; có thể đọc được cảm xúc của con người thông qua nét mặt, cử chỉ. Trong lĩnh vực công nghiệp ô tô, người ta cũng tích hợp AI vào các phương tiện giao thông để giúp giảm thiểu tai nạn giao thông. Trí tuệ nhân tạo cũng được ứng dụng nhiều trong chăm sóc sức khỏe bệnh nhân, như việc sử dụng trợ lý y tế ảo. Trong lĩnh vực giáo dục, AI tạo giúp được cá nhân hóa người học, giúp từng người học bài, nâng cao hiệu quả học tập phù hợp với khả năng của mỗi cá nhân. Trong lĩnh vực quân sự, AI có thể được ứng dụng sản xuất vũ khí tự hành để tiêu diệt mục tiêu định trước, để chống khủng bố, chống bạo loạn.

Như vậy, trí tuệ nhân tạo đã tạo ra nhiều lợi ích cho con người trong hoạt động kinh tế, chăm sóc sức khỏe, thay đổi cách quản trị, quản lý xã hội, cách tiếp cận và thỏa mãn nhu cầu cá nhân. Nhờ đó, nó tạo cơ hội cho các quốc gia thay đổi mô hình tăng trưởng nhằm làm cho đất nước ngày càng vững mạnh hơn.

Bên cạnh việc mang đến những cơ hội, trí tuệ nhân tạo cũng đặt ra nhiều thách thức cho con người và xã hội. Việc sử dụng trí tuệ nhân tạo, rô-bốt thay thế con người ở nhiều công đoạn sản xuất có thể làm cho một bộ phận lao động, ngành nghề bị cắt giảm. Việc sử dụng nhận dạng hình ảnh rất hữu ích trong việc tự động hóa một số thao tác của con người (mở khóa điện thoại, khóa nhà, laptop…), trong phòng, chống tội phạm, nhưng các quyền riêng tư của con người có thể sẽ bị xâm phạm. Nó cũng đặt ra vấn đề sử dụng các loại vũ khí trong chiến tranh. Vũ khí tự hành có thể tìm được đúng đối phương để tiêu diệt, song nó cũng có thể bắn vào dân thường, trường học, cơ sở y tế… AI có thể được sử dụng để phát tán phần

mềm độc hại, tấn công mạng, gây thiệt hại cơ sở vật chất, hạ tầng cơ sở của xã hội, như điện, nước, các dịch vụ thiết yếu… Việc tích hợp công nghệ AI vào vũ khí quân sự có thể nâng cao hiệu quả việc chống khủng bố, nhưng người ta cũng chưa thể lường hết được những hậu quả xảy ra khi vũ khí tự hành được các tổ chức tội phạm, khủng bố nắm giữ. Nhờ AI, thông tin được truyền bá đến công chúng nhanh chóng, nhưng nó cũng có thể tạo ra các thông tin giả gây ảnh hưởng đến các quốc gia, chia rẽ tinh thần đoàn kết của các quốc gia, dân tộc.

Có thể thấy, AI đã hỗ trợ, tạo cho con người, xã hội loài người có cơ hội sống cuộc sống tốt đẹp hơn, nhưng đồng thời nó cũng có thể gây ảnh hưởng nghiêm trọng tới đời sống của chúng ta. Bản chất tự do của con người liệu có còn khi trí tuệ nhân tạo được ứng dụng để thao túng tâm lý, hướng lái ý thức của mỗi người. Quyền tự quyết của con người liệu có còn khi quyền riêng tư bị xâm phạm. Thế giới loài người sẽ đi về đâu nếu như các tổ chức tội phạm, khủng bố nắm trong tay vũ khí tự hành giết người hàng loạt; khi lạm dụng công nghệ AI để tấn công cơ sở hạ tầng của các quốc gia trên thế giới…

Tất cả những điều này đặt ra vấn đề đạo đức trong thiết kế, phát triển, ứng dụng công nghệ AI trong các lĩnh vực đời sống xã hội; đặt ra yêu cầu về giới hạn được phép/ không được phép thực hiện việc thiết kế, phát triển, ứng dụng công nghệ AI. Theo chúng tôi, với trí tuệ nhân tạo, vấn đề đạo đức cần được đặt ra ngay từ khâu thiết kế, phát triển. Cần phải làm rõ, việc thiết kế, phát triển chương trình AI nào đó vì mục tiêu gì, giới hạn nào cho hoạt động của AI. Vấn đề đạo đức cũng cần đặt ra trong quá trình ứng dụng công nghệ AI vào đời sống. Bởi có thể việc thiết kế, phát triển AI có mục tiêu ban đầu tốt đẹp, vì lợi ích tiến bộ của xã hội, song việc ứng dụng công nghệ AI cũng có thể mang tính hai mặt. Người ta ứng dụng công nghệ AI có thể vì mục tiêu tốt đẹp, song cũng có thể vì những toan tính mà đi ngược lại với tiến bộ xã hội, ảnh hưởng đến các quyền và phẩm giá con người.

Từ cách tiếp cận khác nhau, việc đề xuất đạo đức AI để AI mang lại những giá trị tích cực cho con người và xã hội và hạn chế những khía cạnh tiêu cực trong quá trình thiết, phát triển và ứng dụng AI ngày càng được chú trọng[2]. Các đề xuất cũng chỉ ra rằng, đạo đức AI chính là các nguyên tắc, chuẩn mực quy định hành vi của con người trong thiết kế, lập trình, ứng dụng công nghệ AI nhằm mục tiêu bảo đảm phát triển, tiến bộ, công bằng xã hội và vì phẩm giá con người.

2. Việt Nam đang đẩy mạnh phát triển khoa học, công nghệ và đổi mới sáng tạo nhằm mục tiêu đến năm 2025: là nước đang phát triển, có công nghiệp theo hướng hiện đại; đến năm 2030: là nước đang phát triển, có công nghiệp hiện đại và đến năm 2045: trở thành nước phát triển[3]. Để đạt được mục tiêu đó, việc ứng dụng những thành tựu của Cách mạng công nghệ lần thứ tư, với cốt lõi là trí tuệ nhân tạo (AI) được Việt Nam chú trọng, được coi là lĩnh vực then chốt của một số cấp bộ, ngành, cộng đồng doanh nghiệp. Ở Việt Nam, khung chính sách liên quan đến AI cũng đang dần được hoàn thiện. Ngày 16-1-2021, Thủ tướng Chính phủ Việt Nam đã ký quyết định số 127/QĐ-Ttg *Ban hành Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng Trí tuệ nhân tạo đến năm 2030*, trong đó khẳng định, AI sẽ "là một lĩnh vực công nghệ nền tảng của Cách mạng công nghiệp lần thứ tư, góp phần quan trọng tạo bước phát triển đột phá về năng lực sản xuất, nâng cao năng lực cạnh tranh quốc gia, thúc đẩy phát triển kinh tế tăng trưởng bền vững". Chiến lược cũng khẳng định, "đẩy mạnh nghiên cứu, phát triển ứng dụng trí tuệ nhân tạo, đưa trí tuệ nhân tọa tở thành lĩnh vực công nghệ quan trọng của Việt Nam trong cuộc Cách mạng công nghiệp lần thứ tư. Đến năm 2030, Việt Nam trở thành trung tâm đổi mới sáng tạo, phát triển các giải pháp và ứng dụng trí tuệ nhân tạo trong khu vực ASEAN"[4]. Mục tiêu này cho thấy, Việt Nam mong muốn và hướng tới trở thành một quốc gia có trình độ phát triển cao về AI, trở thành trung tâm AI của khu vực[5]. AI đã và đang thay đổi thế giới và cũng sẽ

mang đến những thay đổi cho Việt Nam, là cơ hội để Việt Nam phát triển bứt phá; tuy nhiên, cũng cần phải thấy rõ cả thách thức của AI đối với Việt Nam.

Về cơ hội

Việc triển khai nghiên cứu, ứng dụng công nghệ AI giúp Việt Nam có nhiều cơ hội phát triển kinh tế nhanh và bền vững; đẩy nhanh việc tái cơ cấu nền kinh tế, phát triển các ngành nghề trong lĩnh vực giao thông, y tế, giáo dục. Hiện nay, ở Việt Nam, AI đã được sản xuất nghiên cứu, ứng dụng trong một số lĩnh vực khác nhau của đời sống xã hội. Đó là sử dụng AI trong việc kiểm tra lỗi trong quá trình sản xuất hàng hóa; nội soi hình ảnh dùng trong ngành y tế giúp chẩn đoán, điều trị bệnh chính xác hơn so với phương pháp truyền thống, tính toán hành vi người tiêu dùng, chuyển đổi số, ngân hàng số. Đặc biệt, trong công tác phòng, chống dịch COVID-19 ở Việt Nam vừa qua, công nghệ trí tuệ nhân tạo đã đóng vai trò quan trọng, giúp giảm gánh nặng cho đội ngũ y tế, lực lượng phòng, chống dịch thông qua các ứng dụng thiết thực. Việt Nam đã có những sản phẩm ứng dụng công nghệ AI được đánh giá là hỗ trợ đắc lực cho đội ngũ phòng, chống dịch.

Đơn cử như ứng dụng AI trong việc truy vết người tiếp xúc đã phát huy hiệu quả khi có ổ dịch hoặc ca lây nhiễm mới trong cộng đồng, giúp đội ngũ y tế điều tra lịch trình, lấy thông tin dịch tễ. Đặc biệt, khi số lượng ca lây nhiễm trong cộng đồng lớn, việc nhập liệu từ các bản khai giấy mất nhiều thời gian, dễ nhầm lẫn, công nghệ chuyển hình ảnh thành văn bản với sự hỗ trợ của AI đã giúp số hóa toàn bộ tờ khai, phiếu điều tra dịch tễ và đưa lên hệ thống. Khi dịch bệnh diễn biến phức tạp hơn, AI đã hiện hữu ngay trong các khu cách ly, bệnh viện với robot tự động giúp khử khuẩn, giao hàng, đưa thuốc cho người bệnh... Hệ thống bản đồ dịch tễ, phần mềm đánh giá nguy cơ lây nhiễm cũng phát huy hiệu quả tại các địa phương có

dịch. Trên bản đồ này, người dùng theo dõi được nơi người bệnh đã đến, các khu vực cách ly, khu vực bệnh viện... một cách trực quan[6].

Về thách thức

Thứ nhất, việc lạm dụng AI xuất hiện chủ yếu từ 2 phương diện: thứ nhất, trong quá trình thiết kế, phát triển AI; thứ hai, trong quá trình ứng dụng công nghệ AI. Có thể mục đích ban đầu trong thiết kế, phát triển AI là tốt đẹp, song quá trình ứng dụng công nghệ AI có thể gây ảnh hưởng tích cực hay tiêu cực đối với con người và xã hội. Điều này phụ thuộc vào cách thức con người nhận thức và ứng dụng nó. Ở Việt Nam, nhất là vùng sâu, vùng xa, vùng đồng bào dân tộc thiểu số, AI là vấn đề còn mới mẻ. Điều đó đặt ra vấn đề lạm dụng AI trong ứng dụng các lĩnh vực của đời sống xã hội, dẫn đến quyền riêng tư, tính bảo mật… của con người bị vi phạm mà người dân không nhận biết được hết.

Thứ hai, dù sớm nhận thức được vai trò, vị trí của AI đối với sự phát triển đất nước, nhưng với nền tảng xuất phát điểm chưa cao, Việt Nam cũng chưa sản xuất, ứng dụng được nhiều công nghệ AI trong cuộc sống. Vai trò dẫn dắt của doanh nghiệp Nhà nước trong lĩnh vực này chưa rõ nét. Cùng với đó, cơ sở hạ tầng về công nghệ thông tin của Việt Nam còn thấp. Một số sản phẩm công nghệ lõi chưa phát triển, vẫn phụ thuộc vào bên ngoài[7]. Điều đó tiềm ẩn những nguy cơ bị cài phần mềm gián điệp, dẫn đến việc kiểm soát các hành vi liên quan đến thiết kế, phát triển, ứng dụng AI sẽ khó khăn, đặc biệt là nếu như Nhà nước không xây dựng được hệ thống chế tài đủ mạnh, hiệu lực, hiệu quả và nếu như mỗi cá nhân, tổ chức không nhận thức rõ được ranh giới giữa việc cung cấp thông tin, giám sát tội phạm với vi phạm quyền riêng tư, quyền tự quyết…

Thứ ba, công nghệ AI giúp tạo ra một lượng thông tin và khả năng chia sẻ thông tin nhanh trên các mạng xã hội. Là một trong 10 nước đứng hàng đầu thế giới về số lượng người dùng mạng xã hội,

với gần 64 triệu tài khoản Facebook và gần 35 triệu tài khoản Youtube[8], có thể thấy, trong bối cảnh hiện nay, có một lượng thông tin khổng lồ đầy ắp trên các mạng xã hội, tin thực có, tin không thực cũng có, tin có giá trị khoa học thì nhiều, nhưng tin phản khoa học cũng không ít; những thông tin thực và giả đan xen nhau như vậy đã gây ảnh hưởng đến đời sống con người và xã hội. Mặt tích cực của lượng thông tin này là, nhờ lượng thông tin khổng lồ, nhờ tiến bộ khoa học, mỗi người Việt Nam cũng dễ dàng hơn trong việc tiếp cận những tri thức mới của nhân loại, nhất là các tri thức đa dạng về ngành nghề, lĩnh vực, các tri thức về khoa học liên ngành… để có thể phát triển, hoàn thiện bản thân, đáp ứng nhu cầu phát triển của xã hội. Mặt tiêu cực là, các thông tin giả về sự phát triển kinh tế - xã hội ở Việt Nam, những thông tin sai sự thật, xuyên tạc, vu khống, xúc phạm uy tín của cơ quan, tổ chức, danh dự, nhân phẩm của cá nhân. Một số đối tượng, bằng ứng dụng công nghệ AI, có thể tạo ra, xây dựng các chiến dịch thông tin giả nhằm gây ảnh hưởng xấu đến môi trường văn hóa, con người và sự phát triển của Việt Nam.

3. Truyền thống đạo đức nhân đạo, lòng yêu thương và quý trọng con người[9], vì nghĩa"[10] đã tạo nền tảng cho người Việt Nam có được nhân sinh quan tốt đẹp trong giải quyết các vấn đề của cuộc sống, từ mối quan hệ giữa cá nhân với cá nhân đến mối quan hệ giữa cá nhân với cộng đồng, đến việc giải quyết các vấn đề liên quan đến chiến tranh và hòa bình... Đây là những yếu tố thuận lợi cho việc thực hiện đạo đức AI ở Việt Nam. Đó cũng là một trong những nguyên nhân dẫn đến việc, cho đến hiện tại, AI chưa gây ảnh hưởng tiêu cực đến Việt Nam. Tuy nhiên, kinh nghiệm thế giới cùng với xu hướng phát triển ngày càng nhanh chóng của AI, từ góc độ đạo đức, việc nhận diện, phòng ngừa những tác hại của AI đối với xã hội và con người là quan trọng. Theo đó, ở Việt Nam hiện nay, cần thiết:

Thứ nhất, đẩy mạnh công tác giáo dục, nâng cao trình độ dân trí cho người dân.

AI mang lại cả thách thức và cơ hội cho mỗi quốc gia và mỗi cá nhân. Việc có được nền tảng tri thức tốt sẽ giúp mỗi người có tiếp cận được cơ hội tốt đẹp do sự phát triển khoa học - kỹ thuật nói chung, AI nói riêng mang lại, không chỉ trong lĩnh vực kinh tế, tìm kiếm cơ hội việc làm nhằm giảm bớt tình trạng bất bình đẳng, mà còn nâng cao nhận thức của mỗi người trong việc bảo vệ bản thân trước tác động/ ảnh hưởng mang tính hai mặt của AI; đồng thời, có khả năng nêu lên quan điểm và bảo vệ các vấn đề liên quan đến cá nhân, quyền con người, quyền riêng tư, phẩm giá con người một cách tốt nhất.

Tính chung, Việt Nam là quốc gia có trình độ học vấn cao, nhưng không phân bổ đồng đều giữa các khu vực, vùng, miền của đất nước. Khu vực vùng sâu, vùng xa, vùng đồng bào dân tộc thiểu số có trình độ học vấn thấp hơn các khu vực khác trong cả nước. Theo điều tra năm 2019, tỷ lệ người dân tộc thiểu số có trình độ trung học phổ thông trở lên chiếm 20,2% tổng số người dân tộc thiểu số từ 15 tuổi trở lên, thấp hơn tỷ lệ này của toàn quốc là 16,3%; có 9,1% người dân tộc thiểu số từ 15 tuổi trở lên có trình độ chuyên môn kỹ thuật, thấp hơn nhiều so với tỷ lệ dân số có trình độ chuyên môn kỹ thuật của cả nước (19,2%), tỷ lệ hộ được tiếp cận Internet của đồng bào dân tộc thiểu số chiếm 61,3% tổng số hộ dân tộc thiểu số[11]. Do vậy, bên cạnh việc tiếp tục nâng cao trình độ dân trí cho người dân cả nước, cần tập trung nhiều nguồn lực hơn nữa trong việc nâng cao trình độ dân trí, học vấn khu vực này nhằm tạo ra sự công bằng trong tiếp cận các cơ hội mà AI mang lại, qua đó, cũng tạo cho người dân có khả năng bảo vệ mình trước những rủi ro mà AI gây ra.

Thứ hai, xây dựng bộ quy tắc đạo đức AI, đẩy mạnh công tác giáo dục đạo đức nói chung và đạo đức trong lĩnh vực AI nói riêng.

Trí tuệ nhân tạo, dù ngày càng được hoàn thiện, có khả năng bắt chước các hành vi của con người, thay thế con người trong một

số lĩnh vực, thì xét đến cùng, cũng là sản phẩm do con người tạo ra. Như vậy, trí tuệ nhân tạo sẽ được phát triển như thế nào, được ứng dụng ra sao - là do con người quyết định. Điều đó có nghĩa là những người thiết kế, phát triển, ứng dụng AI là các chủ thể chính để những lợi ích tốt đẹp mà AI mang lại cho con người và xã hội sẽ được thực hiện. Họ cũng là chủ thể chính trong thực hiện các quy tắc đạo đức AI. Việc được trang bị kiến thức về đạo đức nói chung, đạo đức trong hoạt động lĩnh vực AI nói riêng khiến cho đông đảo đội ngũ này có được nhận thức đúng đắn, qua đó, có hành vi đạo đức đúng đắn trong hoạt động thiết kế, phát triển, ứng dụng AI. Có được đạo đức đúng đắn về AI, người thiết kế, phát triển hay người ứng dụng công nghệ AI, trong hoạt động nghề nghiệp của mình, sẽ nhằm mục tiêu phục vụ nhu cầu chân chính của con người và xã hội loài người. Thấm nhuần đạo đức tốt đẹp sẽ thúc đẩy họ vượt qua những trở ngại để tận tâm phụng sự xã hội và con người. Trong trường hợp ngược lại, tức là nếu họ thiếu đi sự phát triển đạo đức tương xứng với sứ mệnh nghiên cứu, phát triển, ứng dụng AI thì thậm chí còn gây ảnh hưởng tiêu cực đến sự phát triển xã hội và con người. Chẳng hạn như, sự thù địch con người, lòng ham muốn quyền lực, tiền bạc, danh vọng một cách không chính đáng cùng những thói xấu đạo đức khác sẽ khiến những cá nhân, tổ chức tiến hành những nghiên cứu, phát triển, ủng hộ những ứng dụng công nghệ AI vào việc chống lại con người, loài người. Thực tiễn lịch sử nghiên cứu và ứng dụng thành tựu khoa học và công nghệ cho thấy, có một số người, một số tổ chức thường khuyến khích và tài trợ cho những nghiên cứu chống lại con người, loài người. Trong trường hợp như vậy, khoa học và công nghệ, mà cụ thể ở đây là AI, từ chỗ là biểu hiện của sức mạnh bản chất con người, phương tiện hùng mạnh góp phần phát triển xã hội và con người lại thể hiện ra như là sức mạnh hủy diệt văn minh và sát hại con người. Chính lý do như vậy, việc xây dựng bộ quy tắc đạo đức AI, giáo dục đạo đức nói chung, đạo

đức AI cho mọi người, đặc biệt là những người thiết kế, phát triển, ứng dụng AI, sẽ trang bị cung cấp tri thức, củng cố tình cảm, niềm tin vào những giá trị tốt đẹp của xã hội, của con người, từ đó, thúc đẩy mỗi người phát triển, ứng dụng AI một cách tích cực[12].

Trong bối cảnh Việt Nam hiện nay, việc xây dựng bộ quy tắc đạo đức AI ở tầm quốc gia cần bao gồm các quy tắc minh bạch, công bằng, trách nhiệm, giải trình - đây là các chuẩn mực đạo đức cần có của người thiết kế, lập trình chương trình và ứng dụng AI trong cuộc sống. Việc xây dựng các quy tắc đạo đức AI để bảo đảm AI được thiết kế, phát triển, ứng dụng theo các quy tắc này. Trên cơ sở bộ quy tắc đạo đức AI quốc gia, mỗi lĩnh vực, ngành nghề có thể cụ thể hóa cho phù hợp với đặc thù nghề nghiệp.

Cùng với việc xây dựng bộ quy tắc đạo đức AI, cần đẩy mạnh công tác giáo dục đạo đức nói chung, đạo đức AI nói riêng, thông qua nhà trường, xã hội, các phương tiện truyền thông đại chúng…

Thứ ba, hoàn thiện hệ thống pháp luật.

Là một phương thức điều chỉnh hành vi con người, pháp luật xác định những giới hạn cho hành động của họ và mức độ trừng phạt những vi phạm. Thông qua các quy phạm do nhà nước ban hành, pháp luật điều tiết hành vi con người một cách cưỡng chế để đảm bảo cho sự ổn định nhất định của xã hội. Vì vậy, người ta gọi pháp luật là đạo đức tối thiểu, còn chuẩn mực đạo đức - điều chỉnh hành vi của con người một cách tự nguyện, tự giác thông qua lương tâm (và cả dư luận xã hội) - là pháp luật tối đa.

Chính vì mối quan hệ như vậy mà pháp luật là một trong những biện pháp để khẳng định một chuẩn mực nào đó, biến nó thành thói quen, từ đó biến nó thành nhu cầu của con người, và hơn nữa, biến nó thành chuẩn mực đạo đức[13]. Banzenade có lý khi nhận xét rằng: "Xưa nay, pháp luật bao giờ cũng là một trong những biện pháp để

khẳng định một chuẩn mực đạo đức và biến nó thành thói quen. Chuẩn mực càng khó bao nhiêu thì vai trò của pháp luật càng quan trọng bấy nhiêu. Vì vậy không thể buông lỏng pháp luật nếu việc này chưa được chuẩn bị bằng sự tiến bộ đạo đức của xã hội"[14].

Để khoa học-công nghệ đáp ứng tối đa vai trò động lực cho sự phát triển xã hội, mỗi quốc gia đều xây dựng các bộ luật nhằm điều chỉnh hoạt động nghiên cứu và ứng dụng khoa học-công nghệ. Những yêu cầu pháp luật vừa định hướng cho hoạt động khoa học-công nghệ nói chung, vừa điều chỉnh hoạt động của những người hoạt động trong nghiên cứu, ứng dụng khoa học - công nghệ nói riêng. Các bộ luật này tạo ra những điều kiện thuận lợi cho khoa học - công nghệ phát triển, những người hoạt động trong lĩnh vực khoa học - công nghệ hành nghề; đồng thời, chúng cũng tạo ra một hành lang pháp lý đảm bảo những nghiên cứu, ứng dụng thành tựu khoa học - công nghệ không đi chệch các mục tiêu kinh tế, xã hội, đạo đức và phát triển con người. Với tư cách là khía cạnh đạo đức thuộc lĩnh vực khoa học - công nghệ, hiện nay, nhiều nước trên thế giới đã bắt đầu xây dựng các khung khổ pháp lý điều chỉnh các hoạt động liên quan trực tiếp đến AI[15]. Tuy nhiên, ở Việt Nam, bởi tính mới mẻ của AI nên vấn đề này chưa được đề cập một cách cụ thể, trực tiếp trong các bộ luật, mà vẫn nằm trong khuôn khổ các điều chỉnh của khoa học - công nghệ nói chung. Do đó, nếu như Việt Nam đã có quyết tâm cao trong việc coi AI là một nhân tố quan trọng trong phát triển kinh tế - xã hội đất nước, thì để tạo sự đồng bộ, hạn chế tác động tiêu cực và xử lý nghiêm minh các hành vi vi phạm trong lĩnh vực này thì việc Nhà nước đồng thời cần xây dựng các bộ luật, các quy định pháp lý điều chỉnh các hành vi liên quan đến AI là hết sức cần thiết. Trong đó, cần xác định bản chất pháp lý của AI, những vấn đề về bảo vệ quyền riêng tư, bảo vệ dữ liệu; về quyền sở hữu (những sản phẩm do AI tạo ra); về trách nhiệm của chủ sở hữu, sử dụng AI trong trường hợp AI gây ra những sai phạm, thiệt hại cho con người và xã hội…

Thứ tư, tăng cường hợp tác quốc tế trong hoạt động nghiên cứu và ứng dụng công nghệ AI.

Có thể thấy, việc thiết kế, phát triển, ứng dụng AI không chỉ nằm trong phạm vi cá nhân, mà còn ở cả phạm vi tổ chức, quốc gia, quốc tế. Do vậy, cần phải xem xét bình diện đạo đức AI không chỉ ở phạm vi cá nhân, mà cả ở cấp độ tổ chức, quốc gia và toàn cầu. Hơn nữa, quá trình toàn cầu hóa hiện nay khiến cho sự hợp tác đã trở thành nguyên tắc ứng xử mà quốc gia nào cũng phải chấp nhận nếu muốn phát triển. Trong bối cảnh đó, hàng loạt các vấn đề mà AI đặt ra: như chiến tranh, khủng bố quốc tế, tấn công mạng trên phạm vi toàn cầu, bất bình đẳng… - là những vấn đề liên quan đến tất cả các quốc gia và việc giải quyết các vấn đề này trên bình diện khoa học đòi hỏi sự hợp tác và tương trợ lẫn nhau giữa các nhà khoa học, các tổ chức, các quốc gia trên thế giới.

Với Việt Nam, hoạt động tăng cường hợp tác quốc tế sẽ giúp Việt Nam tiếp cận dễ dàng với những thành tựu trong việc giải quyết các vấn đề AI trên thế giới, bao hàm cả khía cạnh khoa học - công nghệ và đạo đức, thông qua đó, Việt Nam nâng cao được năng lực giải quyết các vấn đề liên quan đến phát triển, ứng dụng AI; kế thừa có chọn lọc trong xây dựng bộ quy tắc đạo đức AI, xây dựng hành lang pháp lý phù hợp với thông lệ quốc tế và bối cảnh lịch sử cụ thể của Việt Nam.

Thực hiện hợp tác quốc tế trong lĩnh vực đạo đức AI có nghĩa Việt Nam tham gia xây dựng, ký kết các công ước quốc tế, khu vực trong lĩnh vực đạo đức, pháp luật liên quan đến AI; cùng với các quốc gia khác thực hiện và xử lý nghiêm các hành vi vi phạm các chuẩn mực đạo đức, pháp luật quốc tế trong lĩnh vực AI, nhằm bảo đảm cho việc thiết kế, phát triển, ứng dụng AI là vì phẩm giá con người trên phạm vi toàn thế giới.

* *
*

Trong lịch sử nhân loại, mỗi bước phát triển của khoa học - công nghệ đều mang đến những lợi ích vô cùng to lớn cho xã hội, như động cơ hơi nước, điện thoại, máy vi tính… Nhưng một số sản phẩm được tạo ra từ sự phát triển khoa học - công nghệ, bên cạnh những lợi ích mà chúng đem lại, cũng đã đặt ra các vấn đề đạo đức mà loài người phải đối mặt, song bằng nỗ lực của các bên, cho đến nay, chúng ta đã giải quyết được thành công, như năng lượng hạt nhân, công nghệ gen… Với AI, dù rằng sự phát triển hiện nay của nó mới giai đoạn đầu như dự báo[16], thì với những kinh nghiệm xử lý các vấn đề trên, cùng với sự minh bạch, công bằng, trách nhiệm, giải trình, cả ở phía những nhà thiết kế, phát triển và ứng dụng AI, cùng sự kết hợp với các nhà hoạch định chính sách, chúng ta tin tưởng các vấn đề đạo đức của AI sẽ được các bên thực hiện nghiêm túc và việc nhận diện, thực hiện đạo đức AI ở Việt Nam sẽ góp phần vào sự phát triển bền vững của nhân loại, vì những giá trị chân chính của con người.

---

[1] Xem: Nguyễn Thanh Thủy, Hà Quang Thụy, Phan Xuân Hiếu, Nguyễn Trí Thành (2018), *Trí tuệ nhân tạo trong thời đại số: Bối cảnh thế giới và liên hệ với Việt Nam*, https://tapchicongthuong.vn/bai-viet/tri-tue-nhan-tao-trong-thoi-dai-so-boi-canh-the-gioi-va-lien-he-voi-viet-nam-55038.htm, truy cập ngày 20/9/2021.

[2] Như các bản quy tắc: ASILOMAR AI PRINCIPLES (xem: https://futureoflife.org/ai-principles/?cn-reloaded=1, truy cập ngày 20/10/2021), Our Code of Ethics for AI (xem: www.capgemini.com, truy cập ngày 20/10/2021), Ethics Guidelines for Trustworthy AI (set up by the European Commission), Bosch code of ethics for AI (In Brief) (February 19, 2020)…

[3] Xem: Đảng Cộng sản Việt Nam (2021), *Văn kiện Đại hội đại biểu toàn quốc lần thứ XIII*, Nxb. Chính trị quốc gia Sự thật, Hà Nội, t. I, tr.36.

[4] Xem: Thủ tướng Chính phủ, Quyết định số 127/QĐ-TTg ngày 16 tháng 01 năm 2021 Ban hành Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng Trí tuệ nhân tạo đến năm 2030.

[5] Phương Dung, *Chủ tịch HĐQT FPT và 'bố già AI' nói về tương lai AI*, nguồn: https://vietnamnet.vn/vn/giao-duc/chu-tich-hdqt-fpt-va-bo-gia-ai-noi-ve-tuong-lai-ai-769088.html, truy cập ngày 8.10.2021.

[6] Xem: Hoàng Giang, *Từng bước đưa Việt Nam trở thành điểm sáng về trí tuệ nhân tạo*, nguồn: http://baochinhphu.vn/Khoa-hoc-Cong-nghe/Tung-buoc-dua-Viet-Nam-tro-thanh-diem-sang-ve-tri-tue-nhan-tao/437608.vgp, truy cập ngày 2/10/2021.

[7] Xem: Vũ Trọng Lâm, Vũ Thị Hương (2021), "Tăng cường quản lý nhà nước, bảo đảm an toàn thông tin trong truyền thông xã hội ở Việt Nam hiện nay", *Tạp chí Cộng sản*, số 976, tr.28.

[8] Xem: Quang Đông: "*Tin giả và "người đọc thông thái"",* Báo Nhân Dân điện tử, https://nhandan.com.vn/dien-dan/tin-gia-va-nguoi-doc-thong-thai-614893, ngày truy cập 30-6-2021

[9] Xem: Vũ Khiêu (chủ biên) (1974), *Đạo đức mới*, Nxb Khoa học xã hội, Hà Nội, tr.74-86

[10] Xem: Trần Văn Giàu (1980), *Giá trị tinh thần truyền thống của dân tộc Việt Nam*, Nxb Khoa học xã hội, Hà Nội, tr.94

[11] Ủy ban Dân tộc & Tổng cục Thống kê (2020), *Kết quả điều tra thu thập thông tin về thực trạng kinh tế - xã hội của 53 dân tộc thiểu số năm 2019*, Nxb. Thống kê, Hà Nội, tr. 79, 96.

[12] Xem thêm: Viện Hàn lâm Khoa học xã hội Việt Nam (2020), *Đạo đức trong nghiên cứu khoa học: Kinh nghiệm thế giới và bài học cho Việt Nam*, Đề tài nghiên cứu cấp Bộ trọng điểm.

[13] Xem thêm: Cao Thu Hằng (2016), *Xây dựng nhân cách con người Việt Nam từ góc độ truyền thống*, Nxb. Chính trị quốc gia Sự thật, Hà Nội.

[14] G.Bandzeladze (1983), *Đạo đức học*, tập.1, Nxb Giáo dục, Hà Nội, tr.177.

[15] Xem: A.Atabekov, O.Yastrebov (2018), "Legal Status of Artificial Intelligence Across Countries: Legislation on the Move", *European Research Studies Journal*, XXI (4), pp.773-782.

[16] Xem: *Tìm hiểu công nghệ AI là gì? Cách phân loại công nghệ AI*, nguồn: https://www.semtek.com.vn/al-la-gi/, truy cập ngày 9/10/2021. Theo đó, có loại công nghệ AI có thể phân tích những động thái khả thi và lựa chọn hành động chiến lược nhất. Loại này không có ký ức và không thể sử dụng những kinh nghiệm trong quá khứ để tiếp tục huấn luyện trong tương lai. Hoặc có loại công nghệ AI với bộ nhớ hạn chế, sử dụng những kinh nghiệm trong quá khứ để đưa ra các quyết định trong tương lai. Hoặc có loại công nghệ AI này có thể tự mình suy nghĩ và học hỏi những thứ xung quanh để áp dụng cho chính bản thân nó cho một việc cụ thể. Loại công nghệ AI này chưa khả thi trong thời gian hiện tại và có loại công nghệ AI tự nhận thức, có ý thức về bản thân và hành xử như con người. Chúng thậm chí còn có cảm xúc và hiểu được cảm xúc của những người khác. Tất nhiên, loại công nghệ AI này vẫn chưa khả thi.

Tài liệu tham khảo

1. Đỗ Thị Kim Hoa (2009), "Xây dựng đạo đức kinh doanh - cơ sở cho việc thực hiện trách nhiệm xã hội của doanh nghiệp", *Tạp chí Triết học*, số 10.

2. Đoàn Xuân Mượu (1999), *Tiến bộ khoa học nhìn về phía trái*, Nxb. Khoa học xã hội, Hà Nội.

3. Nguyễn Thu Nghĩa (2014), "Mấy vấn đề về đạo đức doanh nhân trong cơ chế thị trường định hướng xã hội chủ nghĩa ở Việt Nam hiện nay", *Tạp chí Triết học*, số 11.

4. Steven John Thompson (2021), *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*, Published in the United States of America by IGI Global.

5. Christoph Bartneck, Christoph Lutge, Alan Wagner, Sean Welsh, *An Introduction to Ethics in Robotics and AI,* at http://www.springer.com/series/10184

6. Vincent c. Muller (Ed.) (2013), *Philosophy and Theory of Artificial Intelligence,* Springer.

7. Patrick Lin, Ryan Jenkins, and Keith Abney (Edited)(2017), *Robot ethics 2.0 from autonomous cars to artificial intelligence*, Oxford University Press.

8. Joachim von Braun, Margaret S. Archer, Gregory M. Reichberg, Marcelo Sánchez Sorondo (Editors) (2021), *Robotics, AI, and Humanity: Science, Ethics, and Policy*, Springer.

9. Raymond E. Spier (Edited) (2002), *Science and Technology Ethics*, Routledge.

ETHICS OF ARTIFICIAL INTELLIGENCE AND POLICY IMPLICATIONS FOR VIET NAM

Assoc.Prof.Dr. Cao Thu Hang
Communist Review

1. The history of human society development shows that science, engineering and technology play an important role in human development. Each advancement of science, engineering, and technology leads to advancements in economic and social fields. In recent years, the remarkable development of these factors has created significant strides for social development, especially the application of their achievements in various fields of social life, including artificial intelligence (AI). AI is the intelligence of machines programmed by humans, with working mechanism almost like the human brain. It can be deployed as a data system, can also be applied to certain devices and allow them to act humanly (application of AI technology). Unlike other software programs that only perform human-programmed operations, AI is characterized by its ability of self-learning like humans, ability to improve the way it works and to respond when receiving feedback from the environment in certain circumstances. It allows problem solving where important information is missing or dealing with new situations[1].

With that feature, it can be seen that artificial intelligence has made a revolution in society, creating favorable conditions for changing the way of production, as well as the role of people in production. In the past, workers were the main actors in the production line; nowadays, robots with the application of AI

technology have replaced that position; Using AI technology, automation creates not only a compact production machine, but also a compact management system. In the economic field, thanks to artificial intelligence, people better grasp customer tastes through the synthesis and evaluation of each person's individual behavior. AI can also be used for targeted advertising activities by capturing human activities and read human emotions through facial expressions and gestures. In the automotive industry, they are integrating AI into vehicles to help reducing traffic accidents. Artificial intelligence is also widely applied in patient care, such as virtual medical assistants. In education, AI personalizes learners, assists each person to learn and improves learning efficiency in accordance with each individual's ability. In military, AI can be applied to produce self-propelled weapons to destroy predetermined targets, to fight terrorism and riots.

Thus, artificial intelligence brings great benefits for people in economic activities, health care and causes changes in social governance and management, in approaching and satisfying individual needs. As a result, it creates opportunities for countries to make their growth patterns to be stronger.

Besides opportunities, artificial intelligence also poses many challenges to people and society. The use of artificial intelligence and robots to replace humans in many production stages can cut off a part of jobs. Image recognition is very useful in automating some human actions (unlocking phones, houses, laptops...), in crime prevention, however, human privacy might be infringed. The use of weapons in war is also an issue since on the one hand self-propelled weapons may find the right enemy to destroy, on the other hand it may also shoot at civilians, schools, medical facilities... AI can be used to spread malware, attack networks, or cause damages to physical facilities and infrastructure of society, such as electricity,

water and other essential services. The integration of AI technology into military weapons can improve the effectiveness of counter-terrorism. However, it is impossible to fully predict the consequences when self-propelled weapons are held by criminal and terrorist organizations. Thanks to AI, information is spread to the public quickly, but it can also share fake news that may threaten the countries, divide the solidarity of nations and peoples.

It can be seen that AI has supported and given people and society the opportunity to live a better life, but at the same time, it may have serious impacts on our lives. Will the free nature of people still exist when artificial intelligence is applied to manipulate the psychology and direction of person's consciousness? Does the human self-determination persist when privacy is violated? How will the human world be if criminal and terrorist organizations have mass-killer self-propelled weapons and abuse AI technology to attack the infrastructure of countries around the world?

All of these raise ethical issues in the designing, developing, and applying AI technology in any fields of social life; set requirements on limits (what is allowed or not allowed) in designing, developing, and applying AI technology. From our point of view, with artificial intelligence, ethical issues need to be raised right from the designing and developing stages. It is necessary to clarify the purposes of designing and developing certain AI programs, as well as limits for AI operation. Ethical issues also need to be considered in the process of applying AI technology into life. Though the AI designing and developing may have good initial goals for the sake of society's progress, the application of AI technology might be two-sided. People apply AI technology, maybe for good intention, and also for purposes that go against social progress, affecting human rights and dignity.

From different approaches, the proposal of AI ethics to bring positive values to people and society and reduce negative aspects in the process of designing, developing and applying AI is getting more attention[2]. The proposal also points out that AI ethics are principles and standards regulating human behavior in designing, programming, and applying AI technology in order to ensure the development, progress, social justice and for human dignity.

2. Viet Nam is accelerating the development of science, technology and innovation with the goal that by 2025: to be a developing country with modern industry; by 2030: to be a developed country with modern industry and by 2045: to become a developed country [3] . To achieve those goals, Viet Nam focused on the application of the Fourth Industrial Revolution's achievements, of which the core is artificial intelligence. AI is considered a key field of the Government, ministries, branches and business community. In Viet Nam, the policy framework related to AI is also gradually being fulfilled. On January 16, 2021, the Prime Minister of Viet Nam signed Decision No. 127/QD-Ttg *promulgating the National Strategy on Research, Development and Application of Artificial Intelligence to 2030*, which affirms AI "is the fundamental technology of the Fourth Industrial Revolution, making an important contribution to creating a breakthrough in production capacity and improving national competitiveness, promoting sustainable economic growth". The strategy also affirms, "To promote research, development and application of AI, making it an important technology of Viet Nam in the Fourth Industrial Revolution. By 2030, Viet Nam will develop to be a center for innovation, development of AI solutions and applications in ASEAN"[4] . This goal shows that Viet Nam aspires and aims to become a country with a high level of AI development, becoming the AI center of the region[5] . AI has been changing the world and also Viet Nam. It is an opportunity for Viet Nam to have breakthroughs, but not forget to clearly aware the challenges of AI.

About the opportunities

The research and application of AI technology provide Viet Nam with lots of opportunities for rapid and sustainable economic development, to accelerate the economic restructuring, and to develop industries of transportation, health care and education. Currently, in Viet Nam, AI has been produced, researched and applied in a various areas of social life. AI is applied in checking for defects in the goods production; in endoscopy image in medical industry helping diagnose and treat diseases more accurately than traditional methods; in considering consumer behavior; in digital transformation and digital banking. Especially, in the recent COVID-19 epidemic prevention and control in Viet Nam, artificial intelligence technology has played an important role, contributing to reduce the burden on the medical system and the prevention and control forces through useful AI applications. Viet Nam has AI-powered products that are considered to be effective support for the epidemic prevention and control forces.

For example, the application of AI in contact tracing has been effective when there is an outbreak or new infection in the community, supporting the medical team to investigate travel history and get epidemiological information. In particular, when the number of community infections is high, the data entry from the declarations takes a long time and is easy to make mistakes, Optical Character Recognition (OCR) technology - an AI-based *technology,* has helped to digitize the whole set of declarations, epidemiological investigation sheets and upload to the system. When the epidemic became more complicated, AI has been present in isolation areas, hospitals with automatic robots to help activities of disinfection, goods delivery, and medicine provision to patients. In addition, epidemiological mapping system and infection risk assessment software are forceful in

epidemic areas. With this map, users can track where the patient has been, and visually observe the isolated and hospital areas[6].

About the challenges

Firstly, the abuse of AI appears mainly in two stages: (1) in the process of AI designing and developing; (2) in the process of AI applying. Though the initial purposes of AI designing and developing maybe good, the process of applying AI technology can have either positive or negative impacts on people and society depending on how people perceive and apply it. In Viet Nam, especially in remote, isolated and ethnic minority areas, AI is a quite new concept. That raises the problems of abusing AI application in various areas of social life, leading to people's privacy, security ... being violated without people's realization.

Secondly, in spite of the early recognition of the AI's role and position in the development course of the country, Viet Nam, with a low starting point, has not yet produced and used many AI technologies in practice. The leading role of state-owned enterprises in this area is vague. Moreover, Viet Nam's information technology infrastructure is still poor. Some core technology products have not been developed, still depend on external supply [7]. This poses potential dangers of spyware, which leads to difficulties in controlling behaviors related to AI designing, developing and applying, especially when the State fails to build a strong and effective sanctions system, individuals and organization are unable to distinguish between providing information/monitoring crimes and violating privacy rights/self-determination rights.

Thirdly, AI technology helps create a large amount of information and quickly share information on social networks. Viet Nam is one of the world's top 10 countries in the number of social network users, with nearly 64 million Facebook accounts and nearly

35 million YouTube accounts[8]. In current circumstance, there is a huge amount of information overflowing on social networks, including either real news or fake news, either scientific or anti-scientific information. This confusion has affected human life and society. Positively, thanks to the huge amount of information and scientific progress, Vietnamese are better accessing new knowledge of mankind, especially diverse knowledge of areas/fields as well as interdisciplinary scientific knowledge in order to perfect themselves and meet the development requirements of the society. Negatively, there are fake news about Viet Nam socio-economic development and unfaithful, distorted, slandered information that insults the reputation of agencies, organizations and individuals. Applying AI technology, offenders can run fake news campaigns to negatively influence Vietnamese cultural environment, citizens and development.

3. The tradition of ethics, humanity, love and respect for people[9] , for the sake of justice[10] is the foundation for Vietnamese people to have a good view of life in solving problems, including interpersonal relationship, individual-community relationship and resolving issues related to war and peace, etc. These are favorable factors for the implementation of AI ethics in Viet Nam. That is also one of the reasons why, up to now, AI has not negatively affected Viet Nam. However, learning from world experience, and with the rapidly growing trend of AI, from an ethical perspective, it is important to identify and prevent the harmful effects of AI on society and people. Accordingly, in Viet Nam, it is necessary to:

Firstly, to promote education and raise the people's intellectual level.

AI brings both challenges and opportunities to each country and each individual. Having a good fundamental knowledge will help each person have access to good opportunities brought by the

development of science - technology in general and AI in particular, not only in the economic field, looking for jobs opportunities to reduce inequality, but also in raising awareness of each person in protecting themselves against the dual impact/influence of AI; at the same time, have the best ability to express opinions and protect individuals, human rights, human privacy and dignity.

In general, Viet Nam is a country with a high level of education, but it is not equally distributed among her regions. Remote and ethnic minority areas have lower education levels than other areas in the country. According to the 2019 survey, the proportion of ethnic minorities with upper secondary school education or higher accounts for 20.2% of the total number of ethnic minority population aged 15 and older, 16.3% lower than the national rate; There is 9.1% of ethnic minority people aged 15 and older have professional and technical qualifications, much lower than the national rate (19.2%). The percentage of ethnic minority households with Internet access is 61.3%[11]. Therefore, in addition to continuing to raise the people's intellectual level for the whole country, it is necessary to invest more resources on improving the people's awareness and education level in the ethnic minority areas in order to ensure equity in access to the opportunities that AI brings about, thereby also to give people the ability to protect themselves against the risks posed by AI.

Secondly, to form an AI ethics code, promote ethical education in general and AI ethics in particular.

Although artificial intelligence is gradually perfected to imitate human behaviors and substitute humans in a number of fields, it is just a product of human, which means how artificial intelligence will be developed and applied - is decided by humans. Thus, designers, developers and users of AI are the main subjects to ensure the significant advantages of AI for people and society. They are also the

key subjects in implementing AI ethics rules. Being equipped with knowledge about ethics in general and AI ethics in particular, many people will have right awareness, thereby, to have the right ethical behavior in designing, developing and applying AI. Having right awareness about AI ethics, the AI designers, developers or users will aim to serve the genuine needs of people and human society, when carrying out their professional activities. Instilling good morals will motivate them to overcome obstacles to wholeheartedly serve society and people. In the opposite case, if they lack the ethical development commensurate with the mission of AI research, development and application, they will even negatively affect human and social development. For example, unwarranted hostility to people, lust for power, money, fame and other moral vices will cause individuals and organizations to conduct research and development, support the application of AI technology against human beings. The historical practice of research and application of scientific and technological achievements shows that there are some people and organizations often encourage and sponsor researches against human beings. In such cases, science and technology, specifically AI, from being an expression of the power of human nature, a powerful means of contributing to social and people development, manifests itself as the power to destroy civilization and kill people. For this reason, the composition of AI codes of ethics, education of ethics and AI ethics for everyone, especially those who design, develop, and apply AI, will provide knowledge, strengthen beliefs in the good values of society and people, thereby, promoting each person to develop and apply AI in a positive way[12].

In current situation of Viet Nam, there is a requirement to develop a national AI ethics code including rules of transparency, fairness, responsibility, and accountability, which are the required

ethical standards of AI designers, programmers and users. Developing AI ethics code is to ensure that AI is designed, developed, and applied basing on these rules. Accordingly, the code can be concretized in each field and area to be suitable for particular characteristics.

Along with developing the AI ethics code, it is necessary to promote ethics education in general and AI ethics in particular, within schools, society and on mass media.

Thirdly, to perfect the legal system.

As an instrument to regulate human behavior, laws define the limits of their actions and punishment for violations. Through the rules promulgated by the State, the law coercively regulates human behavior to ensure social relative stability. Therefore, people consider the law as the minimal morality, and the moral standard which voluntarily adjusting human behavior through conscience (and public opinion) as the highest law.

Because of such correlation, the law is one of the measures to affirm a standard, make it a habit, thereby make it a human need, and moreover, make it a moral standard[13]. Banzenade reasonably gave such a comment that: "From time to time, the law has always been one of the instruments to affirm a moral standard and make it a habit. The harder the standard, the more important the role of the law is. Therefore, the law cannot be loosened unless it is set by the moral progress of society".[14]

To maximize the role of science and technology as a driving force for social development, each country has developed laws to regulate research and application of science and technology. The legal requirements both orientate scientific and technological activities in general and regulate the activities of those engaged in

research and application of science and technology in particular. These laws offer favorable conditions for science and technology development and for science and technology workers; at the same time, they also create a legal corridor to ensure that research and application of scientific and technological works do not deviate from economic, social, ethical and human development goals. As an ethical aspect in the field of science and technology, many countries worldwide have begun to develop legal frameworks governing AI-related activities[15]. However, in Viet Nam, because of the novelty of AI, this issue has not been specifically and directly mentioned in the laws, but it is still within the regulatory framework of science and technology. Therefore, when Viet Nam has high determination in considering AI as an important factor of the country's socio-economic development, it is essential to simultaneously develop laws and legal frameworks governing AI-related activities, in order to have the synchronicity, to limit negative impacts and strictly punish violations in this field. In which, it is necessary to determine the legal nature of AI, the issues of privacy protection, data protection; ownership (of AI-generated products); the responsibility of AI owners and users in case AI causes mistakes and damage to people and society.

Fourthly, to strengthen international cooperation in research and application of AI technology.

It can be seen that the designing, developing and applying of AI is not only within the individual scope, but also at the organizational, national and international scope. Therefore, it is necessary to consider AI ethics not only at the individual level, but also at the organizational, national and global levels. Moreover, the current globalization process makes cooperation a principle of conduct that any country must accept in order to develop. In that context, the

series of problems that AI poses, such as: wars, international terrorism, global cyberattacks, inequality... - are relevant to all countries and solving these problems from the scientific aspect requires cooperation and mutual assistance among scientists, organizations, and countries around the world.

For Viet Nam, strengthening international cooperation will help Viet Nam better access achievements in solving AI problems in the world, from scientific - technological and ethical aspects. Through that, Viet Nam can improve its capacity to solve AI-related problems in development and application; selectively inherit achievements in the development of the AI ethics code, building a legal corridor that suits with international practices and the specific historical context of Viet Nam.

Implementing international cooperation in the field of AI ethics means that Viet Nam participates in composing and signing of international and regional conventions in AI-related ethics and laws; cooperates with other countries in implementing and strictly punishing violations of ethical standards and international laws of AI in order to ensure that the designing, developing and applying of AI is for dignity of people on a global scale.

* *

*

In human history, each step of science and technology development has brought enormous benefits to society, such as steam engines, telephones, computers, etc. However, some products, which are also resulted from the development of science and technology, bring either benefits or concerns about the ethical problems that humanity has to face. With much effort of the parties, so far, we have successfully solved those problems, such as nuclear

energy, gene technology, etc. Even though current development of AI is only in the early stages as predicted[16], with the experience of handling the above-mentioned issues, along with transparency, fairness, responsibility, and accountability of AI designers, developers and users, as well as the cooperation with policy makers, we believe that the AI ethics will be taken seriously by all parties and the AI awareness and implementation of AI ethics in Viet Nam will contribute to the sustainable development of humanity and for the sake of pure human values.

---

[1] Nguyen Thanh Thuy, Ha Quang Thuy, Phan Xuan Hieu, Nguyen Tri Thanh (2018), *Trí tuệ nhân tạo trong thời đại số: Bối cảnh thế giới và liên hệ với Việt Nam* (*Artificial intelligence in the digital age: World context and relations with Viet Nam*), https://tapchicongthuong.vn/bai-viet/tri-tue-nhan-tao-trong-thoi-dai-so-boi-canh-the-gioi-valien-he-voi-viet-nam-55038.htm, accessed on September 20, 2021.

[2] As the rules: ASILOMAR AI PRINCIPLES (see: https://futureoflife.org/ai-principles/?cn-reloaded=1, accessed October 20, 2021), Our Code of Ethics for AI (see: www.capgemini.com, accessed on October 20, 2021), Ethics Guidelines for Trustworthy AI (set up by the European Commission), Bosch code of ethics for AI (In Brief) (February 19, 2020).

[3] See: Communist Party of Viet Nam (2021), *Document of the 13th National Congress*, Truth National Politics Publishing House, Hanoi, t. I, p.36.

[4] See: Prime Minister, Decision No. 127/QD-TTg dated January 16, 2021, promulgating the National Strategy on Research, Development and Application of Artificial Intelligence to 2030.

[5] Phuong Dung, *Chủ tịch HĐQT FPT và 'bố già AI' nói về tương lai AI (Chairman of the Board of Directors of FPT and the 'godfather of AI' talk about the future of AI*) https://Viet Namnet.vn/vn/Giao-duc/chu-tich-hdqt-fpt-va-bo-gia-ai-noi-ve-tuong-lai-ai-769088.html, accessed on 8/10/2021

[6] See: Hoang Giang, *Từng bước đưa Việt Nam trở thành điểm sáng về trí tuệ nhân tạ* (*Step by step making Viet Nam a bright spot in artificial intelligence*), source: http://baochinhphu.vn/Khoa-hoc-Cong-nghe/Tung-buoc-dua-Viet-Nam-tro-thanh-diem-sang-ve-tri-tue-nhan-tao/437608.vgp, accessed on October 2, 2021.

[7] See: Vu Trong Lam, Vu Thi Huong (2021*), Tăng cường quản lý nhà nước, bảo đảm an toàn thông tin trong truyền thông xã hội ở Việt Nam hiện nay* (*Strengthening state management, ensuring information security in social media in Viet Nam today*), Communist Review, no. 976, p.28.

[8] See: Quang Dong: "*Tin giả và "người đọc thông thái""* (*"Fake news and "wise readers"*), Nhan Dan Online, https://nhandan.com.vn/dien-dan/tin-gia-va-nguoi-doc-thong-thai-614893, accessed on June 30, 2021.

[9] See: Vu Khieu (chief editor) (1974), *Đạo đức mới* (*New Ethics*), Social Science Publishing House, Hanoi, pp.74-86.

[10] See: Tran Van Giau (1980), *Giá trị tinh thần truyền thống của dân tộc Việt Nam* (*Traditional spiritual values of the Vietnamese nation*), Social Science Publishing House, Hanoi, p.94.

[11] Committee for Ethnic Minorities & General Statistics Office (2020), *Survey result of the socio-economic status of 53 ethnic minorities in 2019*, Statistics Publishing House, Hanoi, p.79, 96.

[12] See: Vietnam Academy of Social Sciences (2020), *Đạo đức trong nghiên cứu khoa học: Kinh nghiệm thế giới và bài học cho Việt Nam* (*Ethics in scientific research: World experience and lessons for Vietnam*), Key ministerial-level research project.

[13] See: Cao Thu Hang (2016), *Xây dựng nhân cách con người Việt Nam từ góc độ truyền thống* (*Building Vietnamese personality from a traditional perspective*), Truth National Politics Publishing House, Hanoi.

[14] G. Bandzeladze (1983), *Ethics*, vol.1, Education Publishing House, Hanoi, p.177.

[15] See: A.Atabekov, O.Yastrebov (2018), "Legal Status of Artificial Intelligence Across Countries: Legislation on the Move", *European Research Studies Journal*, XXI (4), pp.773-782.

[16] See: *Tìm hiểu công nghệ AI là gì? Cách phân loại công nghệ AI* (*What is AI technology? How to classify AI technology*), source: https://www.semtek.com.vn/al-la-gi/, accessed on October 9, 2021. Accordingly, there is the type of AI technology that can analyze the possible moves and choose the most strategic action. This type has no memory and cannot use past experiences to continue training in the future. Or there's the kind of AI technology with limited memory that uses past experiences to make future decisions. Or there is this kind of AI technology that can think and learn on its own to apply itself to a specific case. This type of AI technology is not feasible for the time being, and there is the type of AI technology that is self-aware, has a sense of self, and behaves like a human. They even have emotions and understand the emotions of others. Of course, this kind of AI technology is not yet available.

References

1. Do Thi Kim Hoa (2009), “Xây dựng đạo đức kinh doanh - cơ sở cho việc thực hiện trách nhiệm xã hội của doanh nghiệp” ("Building business ethics - the basis for the implementation of corporate social responsibility"), *Philosophical Review*, Vol. 10.

2. Doan Xuan Muoi (1999), *Tiến bộ khoa học nhìn về phía trái (Scientific progress looking to the left*), Social Sciences Publishing House, Hanoi.

3. Nguyen Thu Nghia (2014), “Mấy vấn đề về đạo đức doanh nhân trong cơ chế thị trường định hướng xã hội chủ nghĩa ở Việt Nam hiện nay" ("Some issues on business ethics in the socialist-oriented market mechanism in Viet Nam today"), *Philosophical Review*, Vol. 11.

4. Steven John Thompson (2021), *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*, Published in the United States of America by IGI Global.

5. Christoph Bartneck, Christoph Lutge, Alan Wagner, Sean Welsh, *An Introduction to Ethics in Robotics and AI*, at http://www.springer.com/series/10184

6. Vincent c. Muller (Ed.) (2013), *Philosophy and Theory of Artificial Intelligence,* Springer.

7. Patrick Lin, Ryan Jenkins, and Keith Abney (Edited) (2017), *Robot ethics 2.0 from autonomous cars to artificial intelligence,* Oxford University Press.

8. Joachim von Braun, Margaret S. Archer, Gregory M. Reichberg, Marcelo Sánchez Sorondo (Editors) (2021), *Robotics, AI, and Humanity: Science, Ethics, and Policy*, Springer.

9. Raymond E. Spier (Edited) (2002), *Science and Technology Ethics,* Routledge.

XÂY DỰNG KHUNG PHÁP LUẬT CHO PHÁT TRIỂN TRÍ TUỆ NHÂN TẠO Ở VIỆT NAM

TS. Ngô Vĩnh Bạch Dương
Viện Nhà nước và Pháp luật, Viện Hàn lâm KHXH Việt Nam

Trí tuệ nhân tạo là một nhánh của khoa học máy tính có những khả năng của trí tuệ con người, tiêu biểu như các khả năng "suy nghĩ", "hiểu ngôn ngữ", và biết "học tập". Thuật ngữ trí tuệ nhân tạo - AI (Artificial intelligence) từ khi được John McCarthy đặt tên từ 1956[1] đã dần dần bước ra khỏi ranh giới không gian học thuật thuần túy và thâm nhập ngày càng phổ biến trong đời sống xã hội dưới các hình thức công nghệ khác nhau. Hàng loạt các ứng dụng sử dụng trí tuệ nhân tạo từ đơn giản như chatbot trò chuyện, cho đến những hệ thống hỗ trợ quyết định (decision support system) với những cơ sở dữ liệu khổng lồ được sử dụng trong một loạt lĩnh vực như: chăm sóc sức khỏe, giáo dục, kinh doanh, pháp lý, tài chính - ngân hàng, giao thông vận tải, sản xuất, dịch vụ.

Lợi ích do các ứng dụng trí tuệ nhân tạo mang lại là điều không cần bàn cãi. Tuy vậy, đâu là thể chế thích hợp để người ta có thể phát triển nó mà không xung đột với các lợi ích và chuẩn mực ứng xử truyền thống? Hình thành nên khung thể chế, trong đó quan trọng nhất là khung pháp luật, đã và đang là nhiệm vụ của tất cả các quốc gia trên thế giới. Với tính cách là một đại lượng can thiệp trực tiếp hoặc gián tiếp vào ứng xử của xã hội, pháp luật được xây dựng để hướng tới những rủi ro đối với các giá trị cần được bảo vệ. Theo đó, một khung pháp luật cho sự phát triển của trí tuệ nhân tạo, *không chỉ kiến tạo cơ hội cho nghiên cứu, phát triển trí tuệ nhân tạo mà còn nhằm hạn chế những rủi ro của nó đối với các giá trị được pháp luật bảo vệ.*

Có nhiều rủi ro tiềm tàng của phát triển và ứng dụng trí tuệ nhân tạo đối với xã hội, tuy nhiên, từ góc độ luật học và quản tri nhà nước, cần thiết phân loại các rủi ro theo ba chiều cạnh: quyền con người, dân chủ và pháp quyền.

1. Nhận dạng các tác động tiêu cực của trí tuệ nhân tạo đối với đời sống xã hội

Các hệ thống trí tuệ nhân tạo được sử dụng rộng rãi trong các lĩnh vực của cá nhân và công cộng đã mang lại những thách thức đáng kể đối với xã hội. Đó là việc thu thập, công bố dữ liệu cá nhân thiếu vắng sự đồng thuận của người mang dữ liệu đó; Đó là việc định hướng thổi bùng các tranh luận thiên lệch trên truyền thông xã hội trước mỗi cuộc bầu cử; Đó cũng có thể là việc ra quyết định chỉ dựa trên suy luận hình thức của máy tính mà không xét đến các quy trình tố tụng công bằng (due process) - một trong những yếu tố cốt lõi của mọi nền pháp quyền. Phần tiếp theo, chúng ta cùng điểm qua những nhóm rủi ro đó. Ngoài ra, việc nghiên cứu triển khai, ứng dụng trí tuệ nhân tạo đã và đang đặt ra những câu hỏi về các chuẩn mực đạo đức, khi mà dạy các máy tính cần thu thập và khai thác những thông tin riêng tư, hoặc phương án hành động của thực thể mang trí tuệ nhân tạo có thể chưa phân biệt được giữa con người với các động vật hoặc vật thể di chuyển được khác.

1.1. Những tác động đến quyền con người

1.1.1. Tác động đến tự chủ, tự quyết của cá nhân

Tự chủ và tự quyết của con người có thể không được tôn trọng hoặc buộc phải giảm mức độ bảo vệ nhằm mục tiêu phát triển các hệ thống trí tuệ nhân tạo . Điều dễ thấy nhất trong phát triển các ứng dụng trí tuệ nhân tạo là thu thập dữ liệu. trí tuệ nhân tạo cần đến cơ sở dữ liệu lớn để học. Những dữ liệu, bên cạnh các dấu hiệu nhân thân như hình ảnh, tuổi tác, giới tính, tình trạng hôn nhân, khuynh hướng tình dục... vốn dĩ được bảo vệ trong luật dân sự truyền thống,

thì những hoạt động, thói quen, các tương tác xã hội hàng ngày của con người cũng là những dữ liệu cần thiết và có thể được thu thập để "dạy" cho máy tính. Từ đó chúng có thể đề xuất, thậm chí trực tiếp thực hiện các hành động tiếp cận người mang dữ liệu. Chẳng hạn, những người hay tương tác một dạng bài viết trên mạng xã hội có thể được trí tuệ nhân tạo của nhà mạng phân tích và cho hiển thị những quảng cáo, những bài viết được cho là có liên quan đến thói quen tương tác của người dùng đó.

Cần xác định rõ ràng rằng, các hành động của con người dựa trên tự do và tự quyết của họ. Việc theo dõi hành vi của họ và làm dữ liều đầu vào cho trí tuệ nhân tạo là vi phạm quyền con người. Cố nhiên, mỗi khi tham gia một nền tảng mạng xã hội hay bất cứ địa điểm hữu hình hoặc trên mạng, người tham gia phải chấp nhận thể chế ứng xử, hay điều kiện sử dựng dịch vụ (term of services) mà nền tảng kia đưa ra. Thế nhưng, các điều khoản sử dụng dịch vụ hầu hết là khó hiểu, dài dòng, nó làm cho việc đọc toàn bộ để chấp nhận chúng trở nên khó khăn và khuynh hướng chấp nhận mọi điều khoản mà không cần đọc đã trở nên phổ biến.

1.1.2. Có thể tạo ra bất bình đẳng và phân biệt đối xử

Mặt khác, các phương án hành động của trí tuệ nhân tạo, cho đến nay, vẫn là mang tính logic. Nó thiếu đi phần ứng xử lịch sự và tế nhị giữa những con người thực. Điều đó có thể dẫn đến trí tuệ nhân tạo đưa ra những dự đoán và công bố những nội dung nhạy cảm, dù có thể chưa xảy ra. Ví dụ, khi tiếp nhận hồ sơ khai trực tuyến, trí tuệ nhân tạo phân tích dữ liệu lớn và dựa vào các biến số của môn tử vi đẩu số, nó có thể phán đoán về người nộp đơn có thể vi phạm pháp luật, tù tội trong tương lai. Không thể kết luận việc đúng hay sai của dự báo nhưng công bố chúng, hoặc đơn giản là loại hồ sơ kia, đã là một vi phạm quyền con người - quyền bất khả xâm phạm về danh dự, nhân phẩm, quyền lao động và tiếp cận việc

làm đàng hoàng. Nó cũng tạo ra những định kiến bất chính, vi phạm nguyên tắc suy đoán vô tội, và cản trở quyền bình đẳng của con người trong tiếp cận các dịch vụ hợp pháp.

Trong khá nhiều trường hợp trên thực tế, trí tuệ nhân tạo có thể tạo ra những phân biệt đối xử, tác động đến các quyền kinh tế, dân sự của con người. Chẳng hạn, thuật toán trên các nền tảng đặt xe ôm thường không phân biệt tài xế nam, hay nữ. Việc các tài xế được nhận ít hợp đồng hơn các đồng nghiệp khác theo chỉ định của trí tuệ nhân tạo xuất phát từ điểm đánh giá của người sử dụng. Tài xế có hình thức ưa nhìn, hoạt ngôn, về cơ bản, được nhận điểm số tốt hơn. Sự khác biệt này, vô hình trung, đã tạo ra sự phân biệt đối xử và trong rất nhiều trường hợp, người gặp bất lợi là nữ giới.

1.1.3. Phá vỡ sự riêng tư

Phát triển trí tuệ nhân tạo, cũng hoàn toàn có thể tác động đến sự riêng tư của đời sống cá nhân. Như trên đã nói, các dữu liệu dùng để “dạy” cho máy tính, robot không chỉ gồm các dữ liệu hình ảnh, nhân khẩu học, mà còn là các hành động quá khứ, thói quen, sở thích, hồ sơ y tế, an sinh xã hội. Thu thập dữ liệu, có nghĩa là con người bị theo dõi mà có thể họ không thể nhận thức được điều đó.

1.1.4. Tác động đến tự do ngôn luận

Trí tuệ nhân tạo, trong không ít các trường hợp, có tác động không tốt đến tự do ngôn luận. Các thuật toán của nền tảng truyền thông xã hội có thể cấm đoán, không cho hiển thị thậm chí khóa tài khoản của người có nhiều phản hồi tiêu cực cho dù thực tế, nội dung bị báo cáo có thể không có tính chất như vậy. Sự việc bức ảnh “Em bé Napalm” của nhiếp ảnh gia Nick Út được nhiều người đăng lại, cách đây không lâu, đã bị thuật toán của Facebook kết luận là khỏa thân và vi phạm quy tắc cộng đồng của nền tảng này. Đáng lưu ý là, việc kết luận đúng, sai ở đây, chủ yếu do các thuật toán của trí tuệ nhân tạo chứ không phải diễn ra trong một phiên tòa công bằng.

1.1.5. Ảnh hưởng đến quyền kinh tế và gây xung đột lợi ích với các ngành nghề truyền thống

Thay vì ứng dụng vào các công việc khó, tính toán lâu, việc nguy hiểm, nhàm chán, thì việc áp dụng quá rộng rãi trí tuệ nhân tạo và các thực thể mang trí tuệ nhân tạo có thể tạo ra tình huống dư thừa lao động, thất nghiệp. Đây là một bài toán khó trong xây dựng chính sách bởi lẽ, việc nghiên cứu và ứng dụng trí tuệ nhân tạo cũng là tự do của các cá nhân, tổ chức. Họ thực hiện với mong muốn, sức lực và tiền bạc của chính mình. Can thiệp để bảo đảm quyền lao động của các cá nhân bằng cách hạn chế tự do của người khác cũng không phù hợp với nguyên tắc pháp quyền.

Ngoài ra, với các giải pháp tính toán, đề xuất giải pháp nhanh chóng, một số ngành nghề sản xuất có thể sẽ bị cạnh tranh mạnh mẽ và thường thì phần thắng sẽ thuộc về các công nghệ mới. Chấp nhận để một só ngành nghề, công việc chấm dứt sự tồn tại của mình là một xu hương không thể tránh khỏi. Vì vậy, cân bằng lợi ích giữa các đơn vị kinh doanh vừa để, trước hết, bảo đảm tự do kinh doanh của họ, vừa không cản trở việc ứng dụng khoa học vào đời sống, sẽ luôn là một câu hỏi khó đối với tất cả các quốc gia. Lựa chọn phương án chính sách phù hợp với từng giai đoạn phát triển với sự tham gia của tất cả các bên là một lựa chọn duy nhất đúng để bảo đảm tự do của người dân, doanh nghiệp và duy trì sự phát triển đất nước.

1.2. Những tác động đến dân chủ

Trí tuệ nhân tạo được học và phân tích có chủ đích của người "dạy" nó. Bởi vậy, những định kiến của trí tuệ nhân tạo luôn là vấn đề đặt ra. Việc suy đoán của trí tuệ nhân tạo có thể hình thành các thông tin giả và ảnh hưởng đến thái độ và quyết định của cử tri đối với một quan chức cụ thể. Khi thông tin giả (disinformation/ fake news) được lan truyền, nó có thể hủy hoại sự nghiệp chính trị của một ai đó nhưng cũng có thể làm lợi cho những chính khách không xứng đáng.

Lịch sử đã chứng kiến những trào lưu (trend) của các mạng xã hội khi những nội dung được tương tác cao sẽ hiện lên trên trang chủ của các tài khoản cá nhân. Cách mạng Hoa nhài Tunisia, Mùa xuân Arab, Bầu cử Tổng thống Mỹ 2016 được cho là có những tác động đáng kể của các thuật toán được xây dựng có chủ đích không công bằng của những nhóm cá nhân nhất định.

Dù vậy, khi các thuật toán được xây dựng trên cơ sở thiện chí và việc sử dụng chúng có trách nhiệm giải trình, chúng có thể nâng cao tính minh bạch của chính quyền, từ đó cải thiện chất lượng tương tác giữa người dân với chính quyền, nâng cao chất lượng quản trị. Có không ít các trang tin điện tử hoặc thậm chí là fanpage trên các nền tảng lớn được thiết kế và quản lý tốt. Chúng trở thành môi trường tương tác giữa cử tri và chính khách, giữa công dân và chính quyền. Quyền được lắng nghe (right to be heard) và sự tham gia (participation) được bảo đảm một cách thực chất và tường minh. Bài học về tương tác trên nền tảng truyền thông xã hội giữa chính quyền và người biểu tình chống tham nhũng ở Hàn Quốc những năm 2016 - 2017 dù chưa phải là ví dụ tốt về trí tuệ nhân tạo nhưng nó thiết thực cho việc dùng trí tuệ nhân tạo cũng như truyền thông xã hội để cải thiện nền dân chủ.

Thu thập dữ liệu công dân và xây dựng cơ sở dữ liệu công dân là một hoạt động khá phổ biến trên thế giới. Ngày nay, nó càng phổ biến hơn với những quyết tâm chính trị về chuyển đổi số, xây dựng quốc gia số. Nó có thể phục vụ cho nhiều mục đích khác nhau, trước hết là thuận tiện cho việc ra các quyết định quản lý. Chẳng hạn, dữ liệu về tiêm chủng phòng Covid - 19 ở Việt Nam được liên thông giữa nhiều nền tảng, cơ sở dữ liệu khác nhau, thuận tiện cho việc theo dõi và phát hiện sớm nguy cơ bùng phát dịch ở các vùng khác nhau. Hoặc cơ sở dữ liệu của cổng dịch vụ công quốc gia, được chia sẻ cho một số ngân hàng để thực hiện việc nộp thuế thuận tiện. Dù vậy, những quan ngại về rủi ro dữ liệu vẫn được đặt ra, khi đối tượng

được chia sẻ có thể sử dụng cho những mục đích khác và điều đó, không chỉ ảnh hưởng đến quyền con người mà còn có tác động xấu đến nền dân chủ. Hơn nữa, camera theo dõi dựng khắp nơi như ở Trung Quốc, thẻ căn cước có gắn chip ở Việt Nam, dù thuận tiện cho quản lý, nhưng trong chừng mực nhất định, nó có thể là một hạn chế tự do của công dân.

1.3. Những tác động đến pháp quyền

Ứng dụng của trí tuệ nhân tạo trong đời sống pháp luật đã trở nên khá rộng rãi. Từ các chatbot giải đáp các câu hỏi liên quan đến pháp luật cho đến tra cứu văn bản, truy xuất hồ sơ, thậm chí hỗ trợ việc ra phán quyết bằng các gợi ý hướng giải quyết các vụ án. Ngoài sự tiện lợi, áp dụng trí tuệ nhân tạo có thể đem tới những rủi ro nhất định cho chế độ pháp quyền, trong cả đời sống dân sự cũng như trong hoạt động của các cơ quan nhà nước.

1.3.1. Hạn chế quyền cơ bản nếu áp dụng tuyệt đối trí tuệ nhân tạo

Phát triển và ứng dụng trí tuệ nhân tạo trong quản lý nhà nước làm này sinh nhu cầu xây dựng cơ sở dữ liệu công dân. Tự do của công dân, vì vậy cũng bị ảnh hưởng, trước tiên là việc bị theo dõi và sau đó là những nguy cơ mất an toàn dữ liệu cá nhân.

Giả thiết các dữ liệu đầu vào được đưa cho hệ thống trí tuệ nhân tạo, nó có thể đưa ra một phán quyết chính xác mà không cần đến thẩm phán, dự thảo một văn bản quy phạm pháp luật không cần các đại biểu quốc hội biểu quyết. Hiển nhiên, ra kết quả nhanh, có độ chính xác cao là điều mà bất kỳ quốc gia nào, người thực thi công vụ nào cũng mong muốn. Tuy vậy, nguyên tắc pháp quyền đòi hỏi mọi hoạt động của các cơ quan nhà nước đều phải tuân thủ các quy định của pháp luật, kể cả việc soạn luật. Đồng thời, nền pháp quyền hiện đại chỉ được thừa nhận nếu nó bảo đảm được dân chủ và các quyền con người. Khoán trắng cho máy móc, thiết bị mang trí tuệ nhân tạo

sẽ làm biến mất việc lấy ý kiến của các bên hữu quan, xóa bỏ việc biểu quyết của những người đại diện nhân dân trong việc thông qua luật hay thủ tiêu tranh tụng, quyền được xét xử công khai, quyền bào chữa bởi tự mình hoặc mời luật sư… trong hoạt động xét xử. Nền pháp quyền, khi đó, không tồn tại bởi nó mất đi những nguyên tắc rường cột.

Bởi vậy, trí tuệ nhân tạo được áp dụng, chỉ nên với tính cách là các hệ thống hỗ trợ quyết định (Decision support system - DSS) cho các cơ quan hành chính, tư pháp và cả lập pháp. Trí tuệ nhân tạo đề xuất các phương án, cơ quan nhà nước nghe người dân, đương sự trình bày ý kiến, bào chữa, bảo vệ để ra quyết định cuối cùng dựa trên cả gợi ý của trí tuệ nhân tạo, cũng như ý kiến của người dân và những phân tích thực tế của cơ quan nhà nước. Bất luận trí tuệ nhân tạo có thông minh đến đâu, nó cũng chỉ nên là trợ lý chứ không phải là người quyết định.

1.3.2. Xác định tư cách pháp lý của trí tuệ nhân tạo và thực thể mang trí tuệ nhân tạo

Phát triển và ứng dụng trí tuệ nhân tạo, dù ở dạng phần mềm hay hữu hình như robot, các máy móc tự vận hành luôn có tác động đến các chủ thể khác. Trong rất nhiều trường hợp, tồn tại những quan hệ quyền và trách nhiệm mà ý chí của chủ sở hữu không có ý nghĩa.. Chẳng hạn, khi trí tuệ nhân tạo hoặc vật thể mang trí tuệ nhân tạo tự nó gây ảnh hưởng đến quyền của người khác, quan hệ nghĩa vụ ở đây do ai gánh vác, trí tuệ nhân tạo hay chủ sở hữu của nó? Ví dụ, xe tự lái gây tai nạn giao thông, rotbot giao dịch làm sập sàn chứng khoán... Đã không có quá nhiều tranh cãi về dạng quan hệ pháp lý này bởi lẽ, nghĩa vụ của chủ sở hữu trí tuệ nhân tạo hoặc thực thể mang trí tuệ nhân tạo hoàn toàn có thể được xác định tương tự như trường hợp để vật nuôi, đồ vật, cây cối gây thiệt hại cho người khác. Cố nhiên, những trường hợp này hoàn toàn có thể

phát sinh những quan hệ pháp lý phái sinh giữa chủ sở hữu đối với nhà sản xuất hoặc người bán sản phẩm trí tuệ nhân tạo cho mình.

Câu chuyện khó khăn hơn khi trí tuệ nhân tạo tự nó sáng tạo ra các sản phẩm vật chất hoặc phi vật chất như tác phẩm nghệ thuật, khoa học, các vật phẩm hữu hình. Xác định chủ sở hữu, người nắm quyền tác giả là người sáng tạo ra trí tuệ nhân tạo, người sở hữu trí tuệ nhân tạo hay ai khác? Điều này chắc chắn sẽ gây tranh cãi bởi nó không được tạo ra từ ý chí của con người cụ thể. Trong khi trí tuệ nhân tạo, thực thể mang trí tuệ nhân tạo không hề là chủ thể của quan hệ pháp luật nào. Đây là điểm cần thiết được nghiên cứu và luật hóa.

Cũng chính từ việc xác định chủ thể quan hệ pháp luật, tình huống có thể dự đoán trước được là chế độ trách nhiệm giữa người nghiên cứu phát triển, người sản xuất, người bán hàng và người sở hữu trí tuệ nhân tạo hoặc thực thể mang trí tuệ nhân tạo sẽ được phân định như thế nào khi có thiệt hại do trí tuệ nhân tạo hoặc thực thể mang trí tuệ nhân tạo gây ra. Hàng loạt các quy định về trách nhiệm sản phẩm, trách nhiệm liên đới, trách nhiệm của người bảo hiểm được cân nhắc áp dụng mà không phải lúc nào cũng được thực hiện một cách thuận lợi.

1.4. Những tác động đến đạo đức và các giá trị xã hội được pháp luật bảo vệ

Về bản chất, đây là cũng là những nội dung của nguyên tắc pháp quyền. Tuy vậy, hoạt động sáng tạo những máy tính, chương trình máy tính có khả năng tự học, cũng bao hàm cả khả năng xuyên tạc các tiêu chuẩn ứng xử của xã hội người. Các tiêu chuẩn đạo đức, vì thế, cần thiết được tách ra thành yêu cầu riêng. Theo đó, việc nghiên cứu, sáng tạo phải nhằm mục đích phục vụ con người, trong các lựa chọn hành động, con người phải được đặt lên trước tiên, cả về sự an toàn cũng như khả năng phục vụ.

Không khó để thấy rằng, các robot được tạo ra, lập trình hành động và/hoặc dạy trí khôn để làm những việc con người gặp khó khăn, nguy hiểm hoặc nhàm chán khi thực hiện. Tuy vậy, các robot tham gia chiến tranh có thể sẽ có những đặc tính trong phương án hành động hết sức cực đoan, vô nhân đạo, chẳng hạn tiêu diệt tù binh, hàng binh - một vi phạm nghiêm trọng luật quốc tế về chiến tranh. Hoặc trong đời sống dân sự, khi buộc phải lựa chọn sự hy sinh của một chiếc xe tự lái, trí tuệ nhân tạo không được phép đâm xe vào con người mà phải lựa chọn phương án khác. Những quy tắc đạo đức này cần thiết phải được người lập trình, dạy máy tính tuân thủ nghiêm ngặt.

2. Quy định pháp luật hiện hành của Việt Nam có liên quan đến trí tuệ nhân tạo

Không có quy định cụ thể nào về trí tuệ nhân tạo ở Việt Nam, tuy nhiên Hiến pháp 2013, Bộ luật Dân sự 2015, Luật Khoa học và Công nghệ 2013, Luật An toàn thông tin mạng 2013, cùng nhiều văn bản quy phạm pháp luật khác cũng đã đề cập những khía cạnh khác nhau của việc nghiên cứu triển khai, ứng dụng trí tuệ nhân tạo.

Quyết định số 418/QĐ-TTg ngày 11/4/2012 phê duyệt Chiến lược phát triển khoa học và công nghệ giai đoạn 2011-2020 được xem là một trong những văn bản đầu tiên xác định công nghệ trí tuệ nhân tạo thuộc nhóm ngành được ưu tiên phát triển: *Đẩy mạnh nghiên cứu làm chủ và chuyển giao công nghệ trong lĩnh vực công nghệ thông tin, sản xuất sản phẩm thương hiệu Việt Nam như: Công nghệ phần mềm và nội dung số; công nghệ thiết kế, chế tạo mạch tích hợp, bộ nhớ dung lượng cao; công nghệ đa phương tiện; công nghệ đa truy nhập; trí tuệ nhân tạo; công nghệ an toàn và an ninh mạng; phát triển hệ thống trung tâm tính toán hiệu năng cao"*[2]

Quyết định số 66/2014/QĐ-TTg của Thủ tướng Chính phủ : Về việc phê duyệt Danh mục công nghệ cao được ưu tiên đầu tư phát

triển và Danh mục sản phẩm công nghệ cao được khuyến khích phát triển nhắc lại công nghệ trí tuệ nhân tạo được ưu tiên phát triển[3].

Ngày 03 tháng 6 năm 2020, Thủ tướng Chính phủ đã ký Quyết định số 749/QĐ-TTg phê duyệt "Chương trình chuyển đổi số quốc gia đến năm 2025, định hướng đến năm 2030". Tiếp theo, ngày 26/1/2021, Thủ tướng Chính phủ ra Quyết định số 127/QĐ-TTg Ban hành Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng Trí tuệ nhân tạo đến năm 2030.

Đây được coi là những cơ sở pháp lý quan trọng và bước đầu cho việc hình thành khung pháp luật cho phát triển trí tuệ nhân tạo ở Việt Nam. Đây là một tập hợp khá sơ sài, tuy vậy chúng cũng thể hiện được các nội dung cơ bản như sau:

- Khẳng định quyền nghiên cứu khoa học, ứng dụng kết quả nghiên cứu và thụ hưởng thành quả của hoạt động đó[4]

- Trí tuệ nhân tạo là lĩnh vực công nghệ được ưu tiên phát triển

- Xây dựng, hoàn thiện chính sách, pháp luật tạo hành lang pháp lý thông thoáng đáp ứng yêu cầu thúc đẩy nghiên cứu, phát triển và ứng dụng trí tuệ nhân tạo vào cuộc sống[5].

Đáng lưu ý là, Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng Trí tuệ nhân tạo đến năm 2030 đã nêu một định hướng quan trọng là xây dựng, hoàn thiện chính sách, pháp luật đối với lĩnh vực này, tuy nhiên đến mục triển khai thực hiện, các nội dung pháp luật lại có vẻ thiếu những nhiệm vụ cần và đủ tương xứng. Cụ thể, chỉ có Bộ Kế hoạch và Đầu tư được giao nhiệm vụ:

- Xây dựng cơ chế đặc thù cho các trung tâm đổi mới sáng tạo để thu hút các quỹ đầu tư mạo hiểm vào các doanh nghiệp khởi nghiệp sáng tạo về trí tuệ nhân tạo ở Việt Nam.

- Hoàn thiện thể chế, các chính sách thu hút đầu tư trực tiếp của nước ngoài (FDI), các tập đoàn công nghệ cao đa quốc gia xây

dựng trung tâm nghiên cứu, phát triển và ứng dụng trí tuệ nhân tạo ở Việt Nam theo Nghị quyết số 50-NQ/TW ngày 20 tháng 8 năm 2019 của Bộ Chính trị về định hướng hoàn thiện thể chế, chính sách, nâng cao chất lượng, hiệu quả hợp tác đầu tư nước ngoài đến năm 2030."

Cùng với Bộ Tư pháp: "*Xây dựng và hoàn thiện bổ sung các văn bản pháp luật về trách nhiệm pháp lý của các đối tượng liên quan tới trí tuệ nhân tạo".*[6]

Dường như, Chiến lược này chưa có cái nhìn bao quát về những rủi ro cần quản lý hoặc kiểm soát. Ngoài việc thu hút đầu tư, xác định trách nhiệm pháp lý của các đối tượng liên quan thì những khía cạnh quan trọng của dân chủ, bảo vệ quyền con người và xây dựng nhà nước pháp quyền cũng như bảo vệ các giá trị đạo đức không được đề cập. Đây là điều hết sức đáng tiếc bởi nghiên cứu, phát triển công nghệ không có quá nhiều rủi ro đối với xã hội mà là việc ứng dụng nó trong đời sống. Chỉ ở nơi đó, những bất cập, không hoàn thiện cùng những mối nguy hiểm của của công nghệ mới bộc lộ.

3. Điểm danh những nội dung cần hoàn thiện

3.1. Hoàn thiện hệ thống quy định về thu thập và xử lý dữ liệu cá nhân

Phát triển công nghệ trí tuệ nhân tạo không thể không thu thập dữ liệu cá nhân để hình thành nên các cơ sở dữ liệu lớn. Tùy thuộc lĩnh vực mà ứng dụng công nghệ đó hướng tới, các thông tin có thể là toàn bộ hoặc chỉ một vài phương diện của các cá nhân con người như bệnh tật đối với các ứng dụng chữa bệnh, thói quen mua sắm, hành vi sử dụng hàng hóa đối với các ứng dụng quảng bá sản phẩm… Việc thu thập thông tin, cần thiết phải có sự đồng thuận của người mang dữ liệu, được báo trước với những điều khoản dễ hiểu để việc chấp nhận nó không cảm tính. Hiện tại, pháp luật Việt Nam mới chỉ dừng lại ở các thông tin nhân khẩu học, hình ảnh, các sinh

hoạt riêng tư và gọi chung là "bí mật đời tư". Điều này cần thiết được mở rộng thành quyền tự quyết đối với các dữ liệu, thông tin của mình. Chia sẻ hay không phải dựa trên quyền tự quyết của chính họ.

Việc thu thập và xử lý thông tin cần tuân thủ những nguyên tắc thu thập thông tin đúng mực *(fair information practice principles - FIPPs)*[7] bao gồm cả những yêu cầu về nội dung như mục đích, giới hạn, chất lượng thu thập thông tin; và thủ tục tin như trình tự truy cập và chấp thuận xử lý thông tin[8]. Các nguyên tắc hoạt động thông tin đúng mực này sau đó đã được sử dụng một cách rộng rãi trong pháp luật bảo vệ quyền con người và tiếp tục phát triển.

Các nguyên tắc FIPPs theo OECD năm 1980 đưa ra để các quốc gia tham khảo:

Thu thập có giới hạn: Việc thu thập dữ liệu cá nhân và bất kỳ dữ liệu nào như vậy cần được thực hiện bằng các phương tiện hợp pháp và công bằng, và khi thích hợp, với những kiến thức hoặc sự đồng ý của đối tượng dữ liệu.

Chất lượng dữ liệu: Dữ liệu cá nhân có liên quan đến mục đích mà chúng có thể được sử dụng, và mức độ cần thiết cho các mục đích, phải chính xác, đầy đủ và phải được cập nhật.

Mục đích thu thập được xác định chi tiết: Các mục đích thu thập dữ liệu cá nhân phải được xác định không muộn hơn thời gian thu thập dữ liệu và sau đó sử dụng giới hạn trong việc thực hiện những mục đích đó; Sẽ là không phù hợp nếu việc thu thập như trên khi thay đổi mục đích.

Sử dụng giới hạn: Dữ liệu cá nhân không được tiết lộ, làm sẵn để sử dụng cho các mục đích khác hơn so với những mục đích ban đầu, ngoại trừ trường hợp có sự đồng ý của chủ dữ liệu hoặc do pháp luật quy định

An toàn dữ liệu: Dữ liệu cá nhân phải được bảo vệ bằng các biện pháp bảo vệ an ninh hợp lý chống lại các rủi ro như mất mát hoặc truy cập trái phép, sử dụng, tiêu huỷ, sửa đổi, bổ sung hoặc tiết lộ dữ liệu.

Nguyên tắc công khai: Cần có một chính sách chung của sự cởi mở về sự phát triển, thực hành và các chính sách đối với dữ liệu cá nhân; phương tiện cho việc thiết lập cơ sở dữ liệu cá nhân, và mục đích chính của việc sử dụng của chúng, cũng như danh tính và địa chỉ người điều khiển dữ liệu phải được xác định trước.

Sự tham gia của cá nhân: Một cá nhân có quyền nhận được xác nhận từ người quản lý dữ liệu về việc có hoặc không có dữ liệu có dữ liệu liên quan đến anh ta; được thông báo khi dữ liệu liên quan đến mình trong khoảng thời gian hợp lý, với một thái độ tử tế và theo hình thức dễ hiểu đối với ông ta, không phải chịu lệ phí một cách quá đáng; được thông báo lý do nếu các yêu cầu trên không được chấp nhận và được quyền phản đối các lý do đó; được phản đối các dữ liệu liên quan đến mình và yêu cầu xóa, sửa chữa hoặc cập nhật dữ liệu đó.

Trách nhiệm giải trình: Người quản lý dữ liệu phải có được trách nhiệm tuân thủ các biện pháp để bảo đảm hiệu lực cho các nguyên tắc đã nêu ở trên.

3.2. Quy định chế độ trách nhiệm sản phẩm

Sau rất nhiều năm cập nhật và sửa đổi Bộ luật Dân sự, chế độ trách nhiệm sản phẩm vẫn không được đưa vào một cách chính thức. Đây là chế độ trách nhiệm mà theo đó người tiêu dùng cuối cùng sản phẩm có thể yêu cầu đòi bất kỳ người nào trong chuỗi cung ứng sản phẩm phải bồi thường thiệt hại nếu mình bị thiệt hại hoặc sản phẩm gây thiệt hại cho người khác mà mình phải bồi thường.

Giả thiết robot gây thiệt hại cho người sử dụng, người tiêu dùng có thể kiện bất kỳ ai, kể cả người bán hàng lẫn người sản xuất, thậm chí cả tác giả thiết kế, người phát triển ứng dụng cho dù họ không có quan hệ hợp đồng với mình. Đây không chỉ là biện pháp xóa bỏ sự bất cân xứng giữa tiêu dùng về vị thế trong mối quan hệ với các hãng công nghệ, mà còn bảo đảm việc nâng cao trách nhiệm giải trình và đạo đức của các nhà phát triển công nghệ.

3.3. Xác định quy tắc đạo đức trong nghiên cứu phát triển công nghệ

Đây được xem là một yêu cầu mơ hồ và có nhiều tranh cãi bởi quan niệm về đạo đức không phải ở đâu và lúc nào cũng giống nhau. Tuy vậy có thể chỉ ra những điểm chung cơ bản là bảo đảm trí tuệ nhân tạo phục vụ con người đúng nghĩa chứ không thể là công cụ phương hại đến loài người:

- Tôn trọng nhân phẩm con ngươì

- Không phương hại đến quyền con người, đặc biệt là quyền tự quyết hay tự định đoạt của họ

- Không phân biệt đối xử, từ bất kỳ phương diện nào

Cần thiết xây dựng bộ quy tắc ứng xử trong nghiên cứu, phát triển AI, thiết lập các thông số kỹ thuật đạo đức trí tuệ nhân tạo. Chỉ khi đáp ứng được các thông số đó, sản phẩm mới được đưa ra sử dụng, phân phối.

---

[1] Sarangi, Saswat, and Pankaj Sharma. *Artificial intelligence: evolution, ethics and public policy*. Routledge India, 2018. p. 18.
[2] Quyết định số 418/QĐ-TTg ngày 11/4/2012, Điều 1, Mục III.
[3] Quyết định số 66/2014/QĐ-TTg, Phụ lục I.
[4] Hiến pháp 2013, Điều 53.
[5] Quyết định số 127/QĐ-TTg Ban hành Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng Trí tuệ nhân tạo đến năm 2030, Mục III.1.

[6] Quyết định số 127/QĐ-TTg Ban hành Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng Trí tuệ nhân tạo đến năm 2030, Mục III.6 và 16.

[7] Chưa có khái niệm tương tự ở Việt Nam nên trong khuôn khổ đề tài này chúng tôi tạm dịch *fair information practices* là *những hoạt động thông tin đúng mực* để diễn tả tính chất tử tế, tôn trọng quyền con người, tôn trọng pháp luật và tôn trọng các giá trị đạo đức, nhân văn của nó.

[8] Fred H. Cate: *The Failure ofFair Information Practice Principles*, in trong cuốn: *Consumer protection in the age of "information economy",* edited by Jane K. Winn, Ashgate Publishing Ltd. 2006, trang 341

BUILDING A LEGAL FRAMEWORK FOR ARTIFICIAL INTELLIGENCE DEVELOPMENT IN VIET NAM

Dr. Ngo Vinh Bach Duong
Institute of State and Law, Vietnam Academy of Social Sciences

Artificial intelligence is a branch of computer science that has the capabilities of human intelligence, typically the abilities to "think", "understand language", and know how to "learn". The term artificial intelligence - AI since it was named by John McCarthy in 1956[1] has gradually stepped beyond the boundary of the purely academic space and penetrated increasingly popular in social life under different forms of technology. A wide range of applications use artificial intelligence from simple chat (chatbots) to decision support systems with huge databases used in a wide range of fields such as healthcare, education, business, legal, finance - banking, transportation, manufacturing, services.

The benefits brought by artificial intelligence applications are undisputed. However, what is the appropriate institution so that it can be developed without conflicting with traditional interests and norms of conduct? Forming an institutional framework, the most important of which is the legal framework, has been the task of all countries in the world. As a force that directly or indirectly interferes in the behavior of society, the laws are designed to address the risks of values that need to be protected. Accordingly, a legal framework for the development of artificial intelligence, *not only creates opportunities for research and development of artificial intelligence, but also aims to limit its risks to the values protected by law*.

There are many potential risks of artificial intelligence development and application to society, however, from the

perspective of jurisprudence and state governance, it is necessary to classify risks according to three dimensions: human rights, democracy and the rule of law.

1. Identifying negative impacts of artificial intelligence on social life

The widespread use of artificial intelligence systems in the private and public spheres has brought significant challenges to society. That is the collection and publication of personal data without the consent of the bearer of that data; It's about igniting biased arguments on social media before every election; It could also mean making decisions based solely on formal computer inference without considering due processes - one of the core elements of all rules of law. In the next section, we take a look at those risk groups. In addition, the research, implementation and application of artificial intelligence has been raising questions about ethical standards, when teaching computers to collect and exploit private information, or the AI entity's course of action may not yet be able to distinguish humans from other animals or moving objects.

1.1. Impacts on human rights

1.1.1. Impacts on self-control and self-determination of individuals

Human self-control and self-determination may be disregarded or forced to reduce protection levels for the purpose of developing artificial intelligence systems. The most obvious thing in developing artificial intelligence applications is data collection. Artificial intelligence needs big databases to learn. In addition to data on identity signs such as image, age, gender, marital status, sexual orientation, etc., which are protected in traditional civil law, activities, habits, daily social interactions of people are also necessary data and can be collected to "teach" the computer. From there, they can

recommend, even directly take actions to reach data bearers. For example, people who often interact with a type of articles on a social network can be analyzed by the network operator's artificial intelligence and display advertisements and articles that are thought to be related to interaction habits of that user.

It should be clearly established that human actions are based on their freedom and self-determination. It is a violation of human rights to monitor their behavior and use as input to artificial intelligence. Of course, every time you join a social networking platform or any other physical or online place, the participant must accept the institution of conduct, or the terms of service that the platform provides. the other offered. However, almost all the terms of services are confusing and lengthy, so that it is difficult to read the whole thing to accept them, and the tendency to accept all terms without reading has become popular.

1.1.2. Possibility to produce inequality and discrimination

On the other hand, the action options of artificial intelligence, so far, are still logical. It lacks the polite and delicate behavior between real people. That could lead to artificial intelligence making predictions and publishing sensitive content, even if it hasn't happened yet. For example, when receiving online filings, the artificial intelligence analyzes big data and based on horoscope variables, it can judge the applicant's potential for breaking the law and imprisonment in the future. It is impossible to conclude whether the forecasts are right or wrong but publishing them or simply refusing the application is already a violation of human rights, which is the inviolability of honor, dignity, the right to work and access to decent work. It also creates unfair prejudices, violates the principle of presumption of innocence, and hinders the equal right of people to access legitimate services.

In many practical cases, artificial intelligence can create discriminations, affecting people's economic and civil rights. For example, algorithms on motorbike booking platforms often do not distinguish between male and female drivers. The fact that the drivers received fewer contracts than their peers as indicated by the artificial intelligence stemming from user ratings. Drivers with good-looking and eloquent appearance basically get better scores. This difference, invisibly, creates discrimination and in many cases puts women at a disadvantage.

1.1.3. Breaking privacy

The development of artificial intelligence can also impact the privacy of personal lives. As mentioned above, the data used to "teach" computers and robots not only includes images, demographics but also past actions, habits, preferences, medical records, social security. Collecting data means people are being tracked and they may not be aware of it.

1.1.4. Impacts on freedom of speech

Artificial intelligence, in many cases, has a negative impact on freedom of speech. Algorithms of social media platforms may ban, prevent, or even lock the accounts of people with a lot of negative feedback, despite the fact that the reported content may not be of such nature. The photo "Napalm girl" by photographer Nick Ut that was reposted by many people, not long ago, was concluded by Facebook's algorithm as nude and violated the platform's community rules. Notably, the conclusion of right and wrong here is mainly due to the algorithms of artificial intelligence, not taking place in a fair trial.

1.1.5. Affecting economic rights and causing conflicts of interest with traditional professions

Instead of applying it to difficult, long computation, dangerous and boring jobs, the over-application of artificial intelligence and AI-bearing entities can create a situation of labor redundancy, unemployment. This is a difficult problem in policy making because the research and application of artificial intelligence is also the freedom of individuals and organizations. They do it with their own desires, energy and money. Intervening to secure the labor rights of individuals by restricting the freedom of others is also inconsistent with the rule of law.

In addition, with computational solutions, quick solution proposals, some manufacturing industries may experience strong competition and often the victory will belong to new technologies. Accepting the fact that some professions and jobs end their existence is an inevitable trend. Therefore, balancing the interests between business units in order to, first of all, ensure their freedom of doing business, while not hindering the application of science to life, will always be a difficult question for all countries. Choosing the appropriate policy options for each stage of development with the participation of all parties is the only right choice to ensure the freedom of people, businesses and maintain the development of the country.

1.2. Impacts on democracy

Artificial intelligence is intentionally learned and analyzed by the person who "taught" it. Therefore, the prejudices of artificial intelligence are always a problem. Artificial intelligence speculation can shape disinformation and influence voters' attitudes and decisions towards a particular official. When disinformation/fake news is spread, it can ruin someone's political career but can also benefit unworthy politicians.

History has witnessed the trends of social networks when highly interactive content will appear on the homepage of individual

accounts. The Tunisian Jasmine Revolution, the Arab Spring, the 2016 US Presidential Election are believed to be significantly affected by purposefully and unfairly constructed algorithms by certain groups of individuals.

However, when algorithms are built on a basis of good faith and their use is accountable, they can improve government transparency and subsequently improves the quality of interaction between citizen and the government, improves the quality of governance. There are many websites or even fanpages on major platforms that are well designed and managed. They become an interactive environment between voters and politicians, between citizens and the government. The right to be heard and participation are substantively and explicitly guaranteed. The lesson on interaction on social media platforms between the government and anti-corruption protesters in Korea in the years 2016-2017 may not be a good example of artificial intelligence, but it is practical for using artificial intelligence as well as social media to improve democracy.

Collecting citizen data and building a citizen database is a fairly common activity around the world. Today, it is even more popular with political determinations about digital transformation and digital nation building. It can serve many different purposes, first of all to facilitate management decisions. For example, data on vaccination against Covid-19 in Viet Nam is linked between many different platforms and databases, making it convenient for monitoring and early detection of the risk of outbreaks in different regions. The database of the national public service portal is shared with several banks for convenient tax payment. However, concerns about data risks still arise when the shared data can be used for other purposes, which not only affects human rights but also has a negative impact on democracy. Moreover, the fact that surveillance cameras are installed everywhere in China or ID cards have chips in Viet Nam is

convenient for management, but to a certain extent, it can be a restriction on the freedom of citizens.

1.3. Impacts on the rule of law

The application of artificial intelligence in legal life has become quite widespread, from chatbots that answer legal questions to look up documents, retrieve records, and even assist in decision making with suggestions for solving cases. In addition to convenience, applying artificial intelligence can bring certain risks to the rule of law, both in civil life as well as in the operations of state agencies.

1.3.1. Limiting basic rights if artificial intelligence is absolutely applied

The development and application of artificial intelligence in state management has created the need to build a citizen database. The freedom of citizens is therefore also affected, first by being tracked and then by the risks of personal data insecurity.

Assuming the input data is fed to the artificial intelligence system, it can make an accurate judgment without a judge, draft a legal document without the voting of parliamentarians. Obviously, any country or public official wants to get fast and accurate results. However, the rule of law requires that all activities of state agencies must comply with the provisions of the law, including the drafting of laws. At the same time, the modern rule of law can only be recognized if it guarantees democracy and human rights. Giving full authority to machines and equipment with artificial intelligence means there will be no consultation of interested parties, no voting of people's representatives in passing laws or eliminating litigation, the right to a public trial, the right to defend by oneself or to invite a lawyer, etc. in trial activities. Thus, the rule of law does not exist because it has lost its key principles.

Therefore, artificial intelligence should be applied only as decision support systems (DSS) for administrative, judicial and legislative agencies. Artificial intelligence proposes options, state agencies listen to people and litigants presenting their opinions, making excuses and defending to make a final decision based on artificial intelligence suggestions, ideas of the people and the actual analysis of state agencies. No matter how smart artificial intelligence is, it should only be the assistant, not the decider.

1.3.2. Determining the legal status of artificial intelligence and AI-embeded entities

The development and application of artificial intelligence, whether in the form of software or tangible as robots, autonomous machines always have an impact on other subjects. In many cases, there exist right and responsibility relationships for which the will of the owner is not meaningful. For example, when artificial intelligence or an AI-carrying entity affects the rights of others, who is responsible for the obligation relationship here, artificial intelligence or its owner? For example, self-driving cars cause traffic accidents, trading robots crash the stock exchange... There has not been too much controversy about this type of legal relationship because the obligations of the owner of the artificial intelligence or the entity carrying the artificial intelligence can be determined in the same way as in the case of pets, objects or trees causing damage to others. Of course, these cases can completely give rise to derivative legal relationships between the owner and the manufacturer or seller of his artificial intelligence product.

The problem is more difficult when artificial intelligence itself creates physical or immaterial products such as works of art, science and tangible items. Determining the owner, copyright holder is the creator of artificial intelligence, the owner of artificial intelligence or

someone else? This is certainly controversial because it was not created by the will of a particular person. While artificial intelligence or the AI-carrying entity is not the subject of any legal relationship. This is a point that needs to be studied and legalized.

Determining the subject of legal relations can help predict in advance how responsibilities will be divided between the researcher, the manufacturer, the seller and the owner of the artificial intelligence or AI-carrying entity when there is damage caused by artificial intelligence or AI-carrying entity. A wide range of provisions on product liability, joint liability, and insurer's liability are considered to apply but are not always implemented favorably.

1.4. Impacts on ethics and social values protected by law

In essence, these are also the contents of the rule of law. However, the creation of self-learning computers and computer programs also includes the ability to distort the standards of behavior of human society. Therefore, ethical standards need to be split into separate requirements. Accordingly, research and creation must aim at serving people and in action choices, people must be put first, both in terms of safety and service ability.

It is not difficult to see that robots are created, programmed to act and/or taught intelligence to do things that humans find difficult, dangerous or boring to do. However, robots participating in war may have extreme and inhumane action options, such as destroying prisoners of war and surrended soldiers - a serious violation of international law on war. In civilian life, when forced to choose the sacrifice of a self-driving car, artificial intelligence is not allowed to crash the car into a person but must choose another option. Computer programmers need to teach artificial intelligence to strictly follow these ethical rules.

2. Current legal regulations related to artificial intelligence in Viet Nam

There are no specific regulations on artificial intelligence in Viet Nam, however, the 2013 Constitution, the 2015 Civil Code, the 2013 Law on Science and Technology, the 2013 Law on Cyberinformation Security and many other legal documents have mentioned different aspects of artificial intelligence research, implementation and application.

Decision No. 418/QD-TTg dated April 11, 2012 approving the Strategy for Science and Technology Development for the period 2011-2020 is considered as one of the first documents identifying artificial intelligence technology in the group industries prioritized for development: *"Promote research, mastery and technology transfer in the field of information technology, produce of Vietnamese branded products such as: Software technology and digital content; technology for designing and manufacturing integrated circuits, high-capacity memory; multimedia technology; multi-access technology; artificial intelligence; safety technology and network security; develop a central high-performance computing system"*[2].

Decision No. 66/2014/QD-TTg of the Prime Minister on the approval of the List of high technologies prioritized for development investment and the List of high-tech products encouraged for development mentions that artificial intelligence technology is prioritized for development[3].

On June 3, 2020, the Prime Minister signed Decision No. 749/QD-TTg approving the *"National digital transformation program to 2025 with orientation to 2030"*. Next, on January 26, 2021, the Prime Minister issued Decision No. 127/QD-TTg promulgating the National Strategy on Research, Development and Application of Artificial Intelligence to 2030.

These are considered as important legal bases and the first step for the formation of a legal framework for artificial intelligence development in Viet Nam. This is a rather sketchy set, but they also show the following basic contents:

- Affirming the right to conduct scientific research, apply research results and enjoy the results of such activities[4]

- Artificial intelligence is a priority technology area for development

- Develop and finalize policies and laws to create an open legal corridor to meet the requirements of promoting research, development and application of artificial intelligence in life[5].

Notably, the National Strategy on Research, Development and Application of Artificial Intelligence to 2030 has stated an important orientation, that is to develop and perfect policies and laws for this field. However, in terms of implementation, the legal contents seem to lack the necessary and adequate tasks. Specifically, only the Ministry of Planning and Investment is tasked with:

- *Develop a specific mechanism for innovation centers to attract venture capital funds into innovative AI startups in Viet Nam*.

- *Perfecting institutions and policies to attract foreign direct investment (FDI), multinational high-tech corporations to build centers for research, development and application of artificial intelligence in Viet Nam according to Resolution No. 50-NQ/TW dated August 20, 2019 of the Politburo on orientations to perfect institutions and policies and improve the quality and efficiency of foreign investment cooperation until 2030*.

Together with the Ministry of Justice: *"Develop and complete additional legal documents on the legal liability of objects related to artificial intelligence".*[6]

It seems that this Strategy does not have an overview of the risks that need to be managed or controlled. Apart from attracting investment, determining the liability of related objects, the important aspects of democracy, the protection of human rights and the construction of the Law-governed State as well as the protection of moral values are not mentioned. This is very unfortunate because the research and development of technology does not have many risks to society but its application in life does, where the inadequacies, incompleteness and dangers of technology are revealed.

3. Contents to be completed

3.1. Completing the system of regulations on collecting and processing personal data

Development of artificial intelligence technology inevitably collects personal data to form big databases. Depending on the field to which the application of such technology is directed, the information may be all or only a few aspects of the individual human being, such as illness for medical applications, shopping habits, behavior of using goods for product promotion applications… It is necessary to obtain the consent of the data bearer when collecting information which is foreshadowed in understandable terms so that its acceptance is not emotional. Currently, Vietnamese law only stops at demographic information, photos, private activities and collectively referred to as "private life secrets". This needs to be extended to the right to self-determination over their data and information. Sharing or not is based on their own self-determination.

Information collection and processing must comply with *fair information practice principles - FIPPs*[7], including content requirements such as purposes, limitations, quality of information collection and information procedures such as procedures for accessing and consenting to information processing[8]. These fair information operations

principles have been widely used in human rights legislation and continue to evolve.

The 1980 OECD FIPPs principles are available to countries for reference:

Limited Collection: The collection of personal data and any such data should be carried out by lawful and fair means, when appropriate with the knowledge or consent of the data subject.

Data quality: Personal data that is relevant to the purposes for which they may be used, and to the necessary extent for the purposes, must be accurate, complete and up to date.

Detailed purpose of collection: Purposes of personal data collection must be identified no later than the time of data collection and then for limited use in the fulfillment of those purposes. It would be inappropriate for such collection when changing the purpose.

Limited Uses: Personal data is not disclosed, made available to use for purposes other than those originally intended, except with the consent of the data owner or as required by law.

Data security: Personal data must be protected by reasonable security safeguards against risks such as loss or unauthorized access, use, destruction, modification, addition or data disclosure.

Disclosure Principle: There should be a general policy of openness regarding the development, practice and policies of personal data; the means for the establishment of personal databases and the primary purpose of their use, as well as the identity and address of the data controller must be determined in advance.

Personal participation: An individual has the right to receive confirmation from the data custodian of the presence or absence of data related to him; to be informed when the data concerns him

within a reasonable period of time, in a manner that is kind and in a form understandable to him, without incurring undue fees; be informed of the reasons if the above requests are not accepted and have the right to object to such reasons; have the right to object to the data relating to them and to request that such data be deleted, corrected or updated.

Accountability: Data custodians should be held accountable for complying with measures to enforce the principles outlined above.

3.2. Regulation on product liability regime

After so many years of updating and revising the Civil Code, the product liability regime is still not officially included. This is a liability regime under which the end consumer of the product can claim damages from anyone in the product supply chain if he suffers damage or the product causes damage to others for which he must compensate.

Assuming the robot causes damage to users, consumers can sue anyone, including sellers and manufacturers, even designers, application developers even though they have no contractual relationship with them. This is not only a measure to eliminate the asymmetry of consumer status in relation to technology firms, but also to ensure enhanced accountability and ethics of technology developers.

3.3. Defining ethical rules in technology research and development

This is considered a vague and controversial requirement because the concept of morality is not always the same. However, it is possible to point out the basic commonalities that are to ensure that artificial intelligence serves humans properly, and cannot be a tool to the detriment of humanity:

- Respect human dignity

- Without prejudice to human rights, especially their right to self-determination

- Non-discrimination from any aspect

It is necessary to develop a code of conduct in AI research and development, and to establish artificial intelligence ethical specifications. Only when those parameters are met can the product be used and distributed.

---

[1] Sarangi, Saswat, and Pankaj Sharma. *Artificial intelligence: evolution, ethics and public policy*. Routledge India, 2018. p. 18.
[2] Decision No. 418/QD-TTg dated 11/4/2012, Article 1, Section III.
[3] Decision No. 66/2014/QĐ-TTg, Appendix I.
[4] The 2013 Constitution, Article 53.
[5] Decision No. 127/QD-TTg promulgating National strategy on research, development and application of Artificial Intelligence to 2030, Section III.1.
[6] Decision No. 127/QD-TTg promulgating National strategy on research, development and application of Artificial Intelligence to 2030, Section III.6 and 16.
[7] There is no similar concept in Viet Nam, so within the framework of this topic, we temporarily use *fair information practices* as *proper information activities* to express kindness, respect for human rights, respect for the law and respect its moral and humanistic values.
[8] Fred H. Cate: *The Failure of Fair Information Practice Principles*, in: *Consumer protection in the age of "information economy",* edited by Jane K. Winn, Ashgate Publishing Ltd. 2006, page 341

TRÍ TUỆ NHÂN TẠO VÀ VẤN ĐỀ BẢO HỘ QUYỀN TÁC GIẢ

TS. Trần Văn Biên
Viện Nhà nước và Pháp luật, Viện Hàn lâm KHXH Việt Nam
PGS.TS. Vũ Thị Hải Yến

1. Trí tuệ nhân tạo và tác phẩm được tạo ra bởi trí tuệ nhân tạo

Trí tuệ nhân tạo (Artificial Intelligence - AI) hiểu đơn giản là những cỗ máy thông minh, mô phỏng trí tuệ của con người, được lập trình để suy nghĩ giống con người và bắt chước hành động của họ. Thuật ngữ này cũng có thể được áp dụng cho bất kỳ máy móc nào thể hiện những đặc điểm liên quan đến trí óc con người như học tập, ra quyết định hay giải quyết các vấn đề[1]. Nếu như những người máy đầu tiên trước đây chỉ có thể mô phỏng lại những hành vi của con người, thì ngày nay, thế hệ người máy hiện đại đã có những khả năng vượt trội, có trí thông minh như hoặc thậm chí còn hơn con người, có khả năng suy nghĩ, hành động… như con người. Trong cuốn sách "Artificial Intelligence A Modern Approach", các tác giả đã phân tích bốn yếu tố đặc trưng của AI, đó là: i) Suy nghĩ của con người (Thinking humanly); ii) Suy nghĩ hợp lý (Thinking rationally); iii) Hành động của con người (Acting humanly); iv) Hành động hợp lý (Acting rationally)[2]. Các chuyên gia về trí tuệ nhân tạo còn chỉ ra nhiều hơn những đặc tính cơ bản của trí tuệ nhân tạo, đó là: i) Tính sáng tạo; ii) Tính không thể dự đoán trước; iii) Tính độc lập và tự chủ trong hoạt động, không có sự can thiệp của con người; iv) Tính hợp lý; v) Khả năng tự học tập và không ngừng cải tiến, phát triển thông qua sự tương tác với môi trường; vi) Khả năng thu thập dữ liệu và truyền đạt; vii) Tính hiệu quả, chính xác; viii) Khả năng tự do sử dụng các phương án thay thế[3].

Với những năng lực vượt trội đó, ngày nay, trí tuệ nhân tạo đã có thể cho ra đời những tác phẩm văn học, nghệ thuật khiến cho chúng ta phải thực sự ngỡ ngàng và thán phục. Năm 2016, một dự án hợp tác giữa Microsoft và Bảo tàng Rembrandt đã công bố một tác phẩm là bức chân dung 3D có tên "The Next Rembrandt" do máy tính tạo ra dựa trên thuật toán nhận dạng khuôn mặt bằng việc quét dữ liệu từ 346 bức tranh được biết đến của họa sĩ người Hà Lan Rembrandt thế kỷ XVII. Cũng trong năm 2016, cuốn tiểu thuyết ngắn được viết bởi một chương trình máy tính của Nhật Bản đã lọt vào vòng thứ hai giải thưởng văn học quốc gia. Google đã tài trợ một dự án sử dụng trí tuệ nhân tạo để viết các bài báo đưa tin tức cho một tờ báo địa phương. Công ty trí tuệ nhân tạo Deep Mind thuộc sở hữu của Google cũng đã tạo ra phần mềm có thể sáng tác nhạc bằng cách nghe các bản ghi âm[4]. Ngày nay, nhân loại đang chứng kiến trí tuệ nhân tạo thực sự trở thành các nghệ sĩ robot (Robotic artists) có khả năng viết thơ, vẽ tranh, soạn nhạc và thậm chí có thể sáng tác cả nhạc kịch.

Trước đây, máy tính chỉ có thể tạo ra những tác phẩm nghệ thuật thô sơ. Hầu hết các tác phẩm nghệ thuật do máy tính tạo ra phụ thuộc rất nhiều vào sáng tạo của các lập trình viên, còn máy móc thì chỉ là những công cụ như bút (chổi lông) hoặc vải vẽ để hỗ trợ cho công việc sáng tạo. Ngày nay, cuộc cách mạng công nghiệp 4.0 đã mang đến những đột phá về công nghệ với sự phát triển nhanh chóng của các phần mềm tự học, cho phép máy móc có khả năng học tập từ dữ liệu đầu vào, phát triển và tự đưa ra các quyết định hoàn toàn độc lập hoặc theo định hướng có trước, mà không cần được con người lập trình cụ thể[5]. Do đó, để tạo ra một tác phẩm văn học hay nghệ thuật mới, người ta chỉ cần cung cấp cho máy tính những dữ liệu cần thiết, nó sẽ học hỏi từ những dữ liệu này, tự đưa ra quyết định mang tính độc lập trong quá trình sáng tạo về việc tác phẩm sẽ được tạo ra như thế nào. Các tác phẩm loại này, thay vì

được tạo ra bởi trí óc của con người, nó được tạo ra bởi một mạng thần kinh nhân tạo (Neural Network) có quy trình gần giống với quá trình tư duy của con người.

Cho dù trí tuệ nhân tạo ngày nay có thể có "bộ óc" như con người, thậm chí thông minh hơn cả những con người bình thường và làm được những công việc mà trước đây chỉ có con người mới làm được, thì về bản chất, nó vẫn chỉ là một "cỗ máy thông minh". Do đó, vấn đề đặt ra là những sản phẩm sáng tạo về văn học, nghệ thuật do trí tuệ nhân tạo tạo ra liệu có được bảo hộ quyền tác giả - lĩnh vực pháp luật ra đời nhằm mục tiêu ghi nhận và bảo hộ các thành quả sáng tạo tinh thần của con người.

2. Điều kiện bảo hộ tác phẩm theo pháp luật quyền tác giả

Theo truyền thống, đối tượng của quyền tác giả là sản phẩm của hoạt động sáng tạo tinh thần (khác với đối tượng của thế giới vật chất là những vật thể hữu hình). Công ước Berne là công ước quốc tế về quyền tác giả lâu đời nhất và có số lượng quốc gia thành viên đông đảo nhất, tại Điều 2 đã nêu ra một danh sách không hạn chế những tác phẩm được bảo hộ, bao gồm những sản phẩm nguyên gốc trong lĩnh vực văn học, nghệ thuật và khoa học, không phân biệt hình thức và cách thức thể hiện[6]. Trên cơ sở Công ước Berne, pháp luật quyền tác giả của hầu hết các quốc gia trên thế giới chỉ bảo hộ những tác phẩm văn học, nghệ thuật, khoa học đáp ứng điều kiện là sáng tạo tinh thần mang tính nguyên gốc.

Luật Quyền tác giả của Nhật Bản, tại mục (i) khoản 1 Điều 2 quy định: "*Tác phẩm là một sản phẩm mà ở đó những suy nghĩ hoặc tình cảm được thể hiện một cách sáng tạo và nằm trong lĩnh vực văn học, khoa học, nghệ thuật hoặc âm nhạc*"[7].

Khoản 2 Điều 2 Luật Quyền tác giả của Đức quy định về tác phẩm được bảo hộ quyền tác giả như sau: "*Các tác phẩm trong ý nghĩa của bộ luật này chỉ có thể là những sáng tạo tinh thần mang tính cá nhân*"[8].

Luật về Quyền tác giả và quyền liên quan của Thụy Sĩ định nghĩa về tác phẩm như sau: "*Tác phẩm là những sáng tạo tinh thần về văn học và nghệ thuật , mang đặc trưng riêng của tác giả, không phân biệt giá trị hay mục đích của chúng"* (khoản 1 Điều 2)[9].

Khoản 7 Điều 4 Luật Sở hữu trí tuệ Việt Nam hiện hành quy định: "*Tác phẩm là sản phẩm sáng tạo trong lĩnh vực văn học, nghệ thuật, khoa học được thể hiện bằng bất kỳ phương tiện hay hình thức nào*" và tại khoản 3 Điều 14 quy định tác phẩm được bảo hộ "*phải do tác giả trực tiếp sáng tạo ra bằng lao động trí tuệ của mình mà không sao chép từ tác phẩm của người khác*".

Dù có những cách thể hiện khác nhau thì Công ước Berne và pháp luật quyền tác giả của hầu hết các quốc gia trên thế giới đều xác định những yếu tố để một tác phẩm được bảo hộ quyền tác giả[10], đó là:

(i) Được thể hiện dưới một hình thức nhất định

Các kết quả sáng tạo nếu mới chỉ là ý tưởng nằm trong đầu tác giả, chưa được thể hiện ra bên ngoài thế giới vật chất bằng một hình thức nhất định thì không thể được coi là một tác phẩm được bảo hộ quyền tác giả. Về phía tác giả, nếu tác phẩm chỉ tồn tại trong tâm trí tác giả mà người khác không thể nhận biết được thì không đặt ra vấn đề phải bảo hộ. Mặt khác, các sản phẩm sáng tạo trong lĩnh vực văn học, nghệ thuật, khoa học chỉ có ý nghĩa khi công chúng có thể tiếp cận được nó. Vì vậy, yêu cầu đầu tiên để tác phẩm được bảo hộ là nó phải được thể hiện thông qua một hình thức nhất định để người khác có thể nhận biết, xác định được tác phẩm.

(ii) Tác phẩm phải là thành quả của hoạt động sáng tạo tinh thần

Yêu cầu này đòi hỏi tác phẩm phải là kết quả của quá trình suy nghĩ, tìm tòi, sáng tạo, phản ánh tư tưởng, tình cảm của người sáng tác. Nói một cách khác, tác phẩm phải là kết quả sáng tạo có chứa đựng một nội dung tinh thần nhất định.

(iii) Mang tính nguyên gốc (hay dấu ấn cá nhân của tác giả)

Pháp luật quốc tế cũng như pháp luật quyền tác giả của hầu hết các quốc gia trên thế giới chỉ bảo hộ những tác phẩm đáp ứng điều kiện là sáng tạo tinh thần mang tính nguyên gốc. "Tính nguyên gốc" (originality) đòi hỏi tác phẩm do chính tác giả sáng tạo ra, có thể phân biệt được với những tác phẩm của người khác. Ở pháp luật một số quốc gia, yêu cầu này được nâng lên là tác phẩm phải mang "đặc trưng riêng" hay "dấu ấn cá nhân", thể hiện ở nội dung hay hình thức của tác phẩm, hoặc cả hai. Dấu ấn riêng có thể là về tư duy hay phong cách thể hiện của người sáng tác.

Kể từ khi ra đời cho đến nay, nguyên tắc của pháp luật quyền tác giả là dành sự bảo hộ cho những sáng tạo do con người tạo ra, gắn với yếu tố cá nhân của tác giả. Những sản phẩm của trí tuệ nhân tạo về bản chất do máy móc tạo ra, nó không thể hiện tư tưởng, tình cảm của một con người cụ thể, do đó, không được coi là thành quả sáng tạo tinh thần và bị loại trừ ra khỏi những đối tượng được bảo hộ quyền tác giả.

Pháp luật và thực tiễn tại nhiều quốc gia cũng đã thể hiện rõ quan điểm không chấp nhận bảo hộ quyền tác giả đối với những tác phẩm không phải do con người sáng tạo ra. Tại Hoa Kỳ, Cơ quan Bản quyền tuyên bố: "Tác phẩm có tính nguyên gốc được đăng ký quyền tác giả nếu nó được tạo ra bởi một con người". Tuyên bố này xuất phát từ án lệ *Feist Publications* v. *Rural Telephone Service Company, Inc. 499 U.S. 340 (1991)*[11], trong đó phán quyết của toà án cho rằng luật bản quyền chỉ bảo vệ những thành quả của lao động trí tuệ mà dựa trên sự sáng tạo từ tâm trí con người. Tương tự như vậy, trong một án lệ gần đây của Australia (*Acohs Pty Ltd* v. *Ucorp Pty Ltd*)[12], tòa án tuyên bố một tác phẩm được tạo ra với sự can thiệp của máy tính không được bảo vệ bản quyền vì nó không được sáng tạo bởi con người. Tại châu Âu, Tòa án Công lý

Liên minh châu Âu (CJEU) cũng đã ra nhiều phán quyết tương tự, đặc biệt là phán quyết mang tính bước ngoặt trong vụ C-5/08 *Infopaq International A/S* v. *Danske Dagbaldes Forening*[13], cho rằng bản quyền chỉ áp dụng cho các tác phẩm mang tính nguyên gốc, tức là nó phải phản ánh sự sáng tạo trí tuệ của chính tác giả. Điều này có nghĩa, điều kiện cần thiết để tác phẩm được bảo hộ bản quyền là tác giả của nó phải là con người cụ thể[14]. Việc phủ nhận bảo hộ bản quyền đối với tác phẩm do máy tính tạo ra xuất phát từ quan điểm quyền tác giả là sự khích lệ của xã hội về mặt tinh thần đối với tác giả, bên cạnh động lực về kinh tế, vật chất. Trí tuệ nhân tạo sáng tác ra tác phẩm một cách tự động, do đó, không cần có sự khuyến khích và không đặt ra nhu cầu bảo hộ quyền tác giả cho tác phẩm ấy.

3. Tại sao cần bảo hộ quyền tác giả đối với các tác phẩm được tạo ra bởi trí tuệ nhân tạo

Ngày nay, các tác phẩm được tạo ra bởi trí tuệ nhân tạo như âm nhạc, báo chí, hội hoạ, thiết kế, tác phẩm đa phương tiện như trò chơi trên máy tính (game)… có xu hướng ngày càng gia tăng và là những sản phẩm có giá trị thương mại cao trong ngành công nghiệp văn hoá, giải trí, cũng như được ứng dụng vào nhiều lĩnh vực của cuộc sống. So với những tác phẩm thông thường được sáng tạo bởi con người, những tác phẩm do trí tuệ nhân tạo tạo ra thường đòi hỏi nguồn lực đầu tư lớn của các doanh nghiệp. Nếu các tác phẩm đó không được bảo hộ bản quyền, có nghĩa bất kỳ ai đều có thể sử dụng, khai thác mà không phải xin phép và trả phí. Những doanh nghiệp đầu tư công nghệ để tạo ra các tác phẩm đó không có quyền ngăn cấm người khác sử dụng, cũng không thu được phí hay các lợi ích vật chất từ việc khai thác tác phẩm. Điều này không chỉ gây cho họ những tổn thất về kinh tế, những nỗ lực, đầu tư của họ không được đền bù xứng đáng, mà chắc chắn sẽ có tác động tiêu cực đến nền công nghiệp, điển hình là công nghiệp máy tính. Các doanh nghiệp sẽ không còn thiết tha vào việc nghiên cứu, đầu tư công sức

và tiền bạc khi thành quả đầu tư của họ không được đền đáp thoả đáng. Và như vậy, những ưu thế của trí tuệ nhân tạo về tiết kiệm thời gian và chi phí cho việc đầu tư nguồn lực con người trong hoạt động sản xuất, kinh doanh cũng không được tận dụng hiệu quả.

Không chỉ ảnh hưởng tiêu cực đến lợi ích của chủ thể đầu tư phát triển trí tuệ nhân tạo và sự phát triển của nền công nghiệp máy tính, việc không bảo hộ quyền tác giả cho các sản phẩm do trí tuệ nhân tạo tạo ra có thể ảnh hưởng đến lợi ích của xã hội trong việc khuyến khích làm giàu có, phong phú thêm đời sống văn hoá, tinh thần. Trong trường hợp này, pháp luật quyền tác giả đã không thực hiện được sứ mạng điều hoà quyền lợi của người đầu tư sáng tạo với người khai thác, sử dụng và quyền lợi của xã hội.

Những thành tựu về khoa học kỹ thuật ngày nay đang mang lại nhiều cơ hội trong cách thức sáng tạo cũng như hình thức thể hiện tác phẩm. Các nghệ sĩ ngày càng có xu hướng tận dụng máy móc, phương tiện hiện đại, kể cả việc sử dụng trí thông minh nhân tạo để hỗ trợ hoạt động sáng tác của họ. Năm 2017, nữ ca sĩ Taryn Southern, ngôi sao ca nhạc YouTube, đã phối hợp với công ty Amper - một đơn vị phát triển trí tuệ nhân tạo (AI) có khả năng sáng tạo ra những bản nhạc chỉ trong vài giây để hoàn thành album mang tên "I am AI". Điều đặc biệt của album này là toàn bộ phần âm nhạc được sáng tác bởi trí tuệ nhân tạo của công ty Amper. Khi được hoàn thành, nó được đánh giá là khá tương đồng với những ca khúc Pop đương đại ngày nay[15]. Các siêu máy tính hiện nay được cung cấp khả năng tự học, tự phân tích, giải quyết vấn đề… không khác gì con người, dẫn đến khó có thể phân biệt sản phẩm do con người và máy móc tạo ra. Sự kết hợp giữa trí tuệ con người và trí tuệ nhân tạo trong hoạt động sáng tạo có xu hướng ngày càng lan rộng, dẫn đến ranh giới phân biệt giữa tác phẩm nghệ thuật do con người và máy tính tạo ra sẽ ngày càng mờ đi. Điều này cũng đặt ra những thách thức trong thực tiễn nếu pháp luật quyền tác giả vẫn theo nguyên tắc

truyền thống là chỉ bảo hộ sản phẩm sáng tạo tinh thần của cá nhân tác giả.

Lịch sử thế giới đã cho thấy, quyền tác giả hình thành và phát triển dựa trên ý niệm về quyền sở hữu. Trong thời kỳ cổ đại và trung đại, pháp luật mới chỉ dừng lại ở việc bảo vệ quyền sở hữu đối với những "vật thể" hữu hình chứa đựng tác phẩm như cuốn sách hay bức họa, mà chưa bảo hộ quyền sở hữu đối với "sáng tạo trí tuệ"[16]. Sau này, quyền của chủ thể sáng tạo (quyền tác giả đối với sản phẩm sáng tạo) mới được ghi nhận độc lập với quyền sở hữu đối với "vật mang tác phẩm trí tuệ". Lý thuyết về quyền sở hữu cho các lao động trí óc, được gọi là *Thuyết quyền sở hữu trí tuệ* (hay *sở hữu tinh thần*) ra đời vào thế kỷ XVIII, phân biệt bản thảo của tác phẩm (vật thể hữu hình) với tác phẩm (sản phẩm vô hình) và đã chỉ ra mối quan hệ chặt chẽ giữa tác giả với tác phẩm.

Dù khách thể của quyền tác giả là thành quả sáng tạo tinh thần, quyền tác giả vẫn mang những tính chất của một loại quyền sở hữu, đó là: (1) Tính tuyệt đối: Giống như quyền sở hữu tài sản thông thường, quyền tác giả là loại quyền tuyệt đối, có hiệu lực đối với tất cả các chủ thể trong xã hội, kể cả Nhà nước, đều phải tôn trọng quyền tác giả. Bất kỳ sự hạn chế hay giới hạn quyền tác giả đều chỉ có thể thực hiện theo luật định. (2) Tính chủ động: Quyền tác giả mang tính độc quyền, chỉ có chủ sở hữu quyền tác giả được khai thác, sử dụng tác phẩm. (3) Tính bảo vệ: Chủ sở hữu có quyền ngăn cấm bất kỳ ai xâm phạm quyền tác giả của mình[17]. Như vậy, việc chúng ta ứng xử với các tác phẩm được tạo ra bởi trí tuệ nhân tạo theo hướng chỉ công nhận quyền sở hữu đối với vật thể hữu hình, mà không công nhận quyền sở hữu đối với sản phẩm sáng tạo trí tuệ là đi ngược lại xu thế phát triển của pháp luật về quyền tác giả.

Trên thế giới, hiện nay tồn tại hai hệ thống bảo hộ quyền tác giả. Nếu như đối với pháp luật về quyền tác giả của các nước châu Âu

lục địa, lợi ích kinh tế của tác giả và quan hệ của tác giả với tác phẩm của mình là trọng tâm bảo vệ, thì hệ thống pháp luật về quyền tác giả của Anh - Mỹ (Copyright) chú trọng đến quyền của người có thể khai thác tác phẩm (the right to copy)[18]. Bên cạnh việc bảo vệ các quyền nhân thân (quyền tinh thần) của tác giả, pháp luật quyền tác giả có nhiệm vụ bảo vệ các quyền tài sản (quyền kinh tế) cho tác giả, chủ sở hữu quyền tác giả. Ngày nay, với sự phát triển của nền công nghiệp văn hóa, giải trí, việc bảo hộ quyền nhân thân chỉ chiếm vai trò thứ yếu bên cạnh quyền kinh tế. Khi tác phẩm đã được đưa vào thị trường thì nó trở thành hàng hóa giao dịch và càng trở nên độc lập với tác giả. Vì vậy, đối với tác giả, nhiều khi quyền lợi kinh tế lại là mối quan tâm hàng đầu. Do đó, nếu chúng ta lập luận rằng, không thể trao những quyền nhân thân như: Quyền đứng tên trên tác phẩm, quyền đặt tên cho tác phẩm, quyền bảo vệ sự toàn vẹn tác phẩm cho máy móc - vốn được xem là vật thể vô tri vô giác, thì những quyền lợi kinh tế của chủ sở hữu quyền tác giả vẫn cần được ghi nhận và bảo vệ để khuyến khích và tặng thưởng xứng đáng đối với chủ thể đầu tư cho hoạt động sáng tạo. Đối với tác phẩm được tạo ra bởi trí tuệ nhân tạo, chủ thể đầu tư này là người tạo ra các chương trình máy tính cho trí tuệ nhân tạo hoạt động, người cung cấp dữ liệu đầu vào cho máy tính học hỏi, người vận hành sử dụng trí tuệ nhân tạo…

4. Bảo hộ quyền tác giả đối với tác phẩm do trí tuệ nhân tạo tạo ra như thế nào

4.1. Đối tượng được bảo hộ quyền tác giả

Để được bảo hộ quyền tác giả, tác phẩm do trí tuệ nhân tạo tạo ra cần phải đáp ứng yêu cầu tối thiểu về tính sáng tạo. Pháp luật quyền tác giả hiện nay bên cạnh việc bảo hộ những tác phẩm văn học nghệ thuật truyền thống thể hiện rất rõ đặc tính sáng tạo tinh thần và dấu ấn cá nhân của tác giả, thì vẫn dành ra những trường hợp ngoại lệ mà

những yêu cầu này đặt ra rất thấp, điển hình là chương trình máy tính hay bộ sưu tập dữ liệu.

Theo khoản 1 Điều 10 Hiệp định TRIPs, chương trình máy tính được bảo hộ như một tác phẩm ngôn ngữ theo định nghĩa của Công ước Berne, thời hạn bảo hộ tối thiểu là 50 năm[19]. Mặc dù được bảo hộ theo pháp luật về quyền tác giả, chương trình máy tính không mang những đặc tính thẩm mỹ hay nghệ thuật như các tác phẩm khác. Nó đơn giản chỉ là sản phẩm công nghệ tin học được tạo ra để giải quyết một vấn đề nào đó, vì vậy, khó có thể chứng minh chương trình máy tính thể hiện một nội dung tinh thần hay dấu ấn cá nhân của tác giả. Điểm khác biệt cơ bản giữa chương trình máy tính và tác phẩm truyền thống là mức độ sáng tạo thấp, chỉ cần đáp ứng điều kiện tối thiểu là "kết quả hoạt động sáng tạo" - tiêu chí quan trọng nhất để một tác phẩm được bảo hộ, còn các yêu cầu về "sáng tạo tinh thần" hay "dấu ấn cá nhân" không còn quan trọng[20]. Bộ sưu tập dữ liệu cũng là tác phẩm được bảo hộ quyền tác giả khi nó được hình thành từ việc sưu tập, tuyển chọn các tác phẩm, thông tin, dữ liệu khác để trở thành một thể thống nhất. Tương tự như chương trình máy tính, tiêu chí "sáng tạo tinh thần" của tác phẩm này đặt ra rất thấp, chỉ đòi hỏi tính sáng tạo thể hiện thông qua việc lựa chọn, sắp đặt chất liệu của người sưu tập.

Để xem xét tác phẩm do trí tuệ nhân tạo tạo ra có thể được bảo hộ quyền tác giả không, chúng ta cùng so sánh hai quy trình sáng tạo của con người và trí tuệ nhân tạo.

Thứ nhất, cả hai quy trình sáng tạo đều cần đến nguồn chất liệu cho sáng tạo. Nguồn chất liệu này có thể coi là kho tài sản chung của xã hội mà mọi người đều có thể khai thác, bao gồm: Các ý tưởng, chủ đề, các sản phẩm văn hoá của nhân loại, các chất liệu khác từ tự nhiên hay cuộc sống xã hội. Ví dụ: Đối với tác phẩm văn

học thì nguồn chất liệu đầu tiên là kho từ vựng; đối với tác phẩm âm nhạc là các hợp âm… Điểm khác là khi con người sáng tạo tác phẩm, họ tự lựa chọn chất liệu cho việc sáng tác của mình; còn đối với trí tuệ nhân tạo, các lập trình viên sẽ là người lựa chọn và cung cấp dữ liệu đầu vào cho máy tính.

Thứ hai, cả hai quy trình đều cần đến những công cụ để thể hiện tác phẩm, ví dụ: Ngôn ngữ, âm thanh, hình ảnh, màu sắc… và thường gắn với các vật liệu mang nó như: Vải để vẽ, giấy để in, đĩa chứa âm thanh; hoặc sóng âm thanh, sóng điện từ… để phát thanh, truyền hình. Khi con người sáng tạo tác phẩm, họ tự lựa chọn công cụ để thể hiện ý tưởng sáng tạo của mình. Còn đối với trí tuệ nhân tạo, người sử dụng, vận hành là người lựa chọn công cụ cho máy móc thể hiện tác phẩm.

Thứ ba, là quy trình sáng tạo tác phẩm. Trên cơ sở những chất liệu và công cụ cho việc sáng tạo, tuỳ thuộc vào khả năng, sở trường của mình, tác giả phải thêm những chất liệu riêng - chính là dấu ấn cá nhân, thể hiện tư tưởng, tình cảm, tinh thần của tác giả trong tác phẩm. Nếu như nguồn chất liệu chung và công cụ thể hiện tác phẩm có thể coi là “chất liệu tự do” mà mọi chủ thể đều có thể khai thác và không thuộc phạm vi bảo hộ quyền tác giả, thì phần chất liệu riêng mà tác giả đưa vào trong tác phẩm được bảo hộ quyền tác giả. Ví dụ: Đối với tác phẩm hội hoạ, chất liệu riêng đó là sự kết hợp của các đường nét, màu sắc, bố cục trong bức tranh… Đối với tác phẩm do con người tạo ra, tác giả thể hiện suy nghĩ, quan niệm, tư tưởng, tình cảm của mình qua cách thức thể hiện tác phẩm. Còn đối với tác phẩm do trí tuệ nhân tạo tạo ra, bộ óc của máy tính thay thế con người học hỏi, khai thác nguồn dữ liệu đầu vào, tự chọn lọc, xử lý thông tin và tự quyết định việc tạo ra tác phẩm mới như thế nào. Kết quả của hai quá trình đều cho ra đời những tác phẩm có tính sáng tạo nguyên gốc, chỉ khác nhau là sản phẩm do con người hay trí tuệ nhân tạo tạo ra.

Nếu như trước đây, máy tính chỉ tham gia vào hoạt động sáng tạo với vai trò như một công cụ (thay cho bút hay giấy vẽ) thì các thế hệ trí tuệ nhân tạo ngày nay đã có thể đưa ra nhiều quyết định liên quan đến quá trình sáng tạo mà không cần sự can thiệp của con người, nói một cách khác, trí tuệ nhân tạo là một nghệ sĩ sáng tạo thực sự. Tác phẩm được tạo ra bởi trí tuệ nhân tạo có thể đáp ứng được tiêu chí quan trọng nhất, đó là "mang tính sáng tạo", chỉ khác là nó không do con người "trực tiếp tạo ra". Do đó, tương tự như chương trình máy tính hay sưu tập dữ liệu, các nhà làm luật trên thế giới có thể cân nhắc về việc dành cho những tác phẩm được tạo ra bởi trí tuệ nhân tạo một ngoại lệ trong bảo hộ quyền tác giả, đó là không cần đòi hỏi yêu cầu về sáng tạo tinh thần hay dấu ấn cá nhân như các tác phẩm văn học, nghệ thuật truyền thống.

4.2. Chủ thể được bảo hộ quyền tác giả

Trong bối cảnh cách mạng công nghiệp lần thứ tư hiện nay, với sự tham gia rộng rãi của các "robot tự trị thông minh", đã và đang đặt ra vấn đề có công nhận robot như một chủ thể của quan hệ pháp luật hay không? Ở châu Âu, đã xuất hiện những khuyến nghị tạo ra một tư cách pháp lý riêng biệt cho robot, chẳng hạn trao tư cách "điện tử nhân" (electronic person) cho các robot tự trị thông minh nhất[21]. Năm 2017, Saudi Arabia trở thành quốc gia đầu tiên trên thế giới cấp quyền công dân cho một robot có hình dạng giống người thật và mang trí thông minh nhân tạo. Có thể thấy rằng, nếu cuộc cách mạng công nghiệp lần thứ tư dẫn tới những cải cách pháp luật quan trọng, trong đó có việc trao tư cách pháp lý "điện tử nhân" cho robot thông minh, thì điều tất yếu là pháp luật quyền tác giả bên cạnh việc trao quyền cho con người - "tự nhiên nhân", cũng cần xem xét ghi nhận quyền tác giả cho trí tuệ nhân tạo khi sáng tạo ra tác phẩm.

Một số quốc gia, điển hình là Vương quốc Anh thể hiện quan điểm ghi nhận quyền tác giả cho những người tạo ra những điều kiện

cần thiết để tác phẩm được thực hiện. Phần 9 (3) của Đạo luật bản quyền, thiết kế và bằng sáng chế Vương quốc Anh (CDPA) quy định: "*Trong trường hợp tác phẩm văn học, kịch, âm nhạc hoặc nghệ thuật do máy tính tạo ra, tác giả sẽ được coi là người đã tạo ra các điều kiện cần thiết cho tác phẩm được thực hiện*". Hơn nữa, Điều 178 của CDPA định nghĩa một tác phẩm do máy tính tạo ra "*là tác phẩm được tạo ra bởi máy tính, mà không có sự tham gia của tác giả là con người*"[22]. Ý tưởng đằng sau quy định này là tạo ra một ngoại lệ cho việc công nhận quyền tác giả đối với những tác phẩm được tạo ra bởi các chương trình có khả năng tạo ra tác phẩm, thậm chí hoàn toàn được tạo ra bởi máy móc[23]. Đây có thể xem như xu hướng mới trong việc bảo hộ quyền tác giả đối với tác phẩm do trí tuệ nhân tạo tạo ra.

Tuy nhiên, trong lúc còn tranh cãi về việc có trao quyền nhân thân (thuộc quyền tác giả) cho máy tính hay không, pháp luật quyền tác giả hoàn toàn có thể ghi nhận và bảo vệ các quyền kinh tế cho chủ sở hữu quyền tác giả. Các tác phẩm do trí tuệ nhân tạo tạo ra là sản phẩm của một quá trình có "sự tham gia" của con người như: Người cung cấp dữ liệu cho máy tính, người tạo ra phần mềm trí tuệ nhân tạo, người sử dụng, vận hành trí tuệ nhân tạo… Ví dụ, để máy tính tạo ra một bản nhạc, trước tiên, chúng ta phải cung cấp cho nó các bản nhạc theo một phong cách định sẵn để làm nguồn đào tạo cho máy tính. Cùng với đó, lập trình viên phải sử dụng những thuật toán để giúp máy tính tự điều chỉnh nhịp điệu, phong cách, tâm trạng của bản nhạc…

Trong quá trình tạo ra tác phẩm của trí tuệ nhân tạo, nguồn dữ liệu để đào tạo trí tuệ nhân tạo rất quan trọng. Nếu những dữ liệu đó được máy móc tự thu thập từ các nguồn như internet hoặc mạng xã hội, sau đó được trí tuệ nhân tạo phân tích và tự tạo ra tác phẩm, thì không cần ghi nhận quyền tác giả. Tuy nhiên, nếu nguồn dữ liệu để

đào tạo trí tuệ nhân tạo do nhà cung cấp thu thập, lựa chọn, thì phải ghi nhận quyền tác giả cho người cung cấp, tương tự như ghi nhận quyền tác giả của bộ sưu tập dữ liệu. Một điểm quan trọng là tác phẩm được tạo ra bởi máy tính là một loại tác phẩm phái sinh, vì nó được tạo ra từ những tác phẩm đã có từ nguồn dữ liệu đầu vào của máy tính. Do đó, theo quy định của pháp luật, tác giả tác phẩm phái sinh phải xin phép và thực hiện nghĩa vụ trả thù lao cho chủ sở hữu tác phẩm gốc khi muốn sử dụng tác phẩm gốc, trong trường hợp tác phẩm gốc vẫn còn trong thời hạn bảo hộ quyền tác giả. Đối với tác phẩm do trí tuệ nhân tạo tạo ra, nếu người cung cấp dữ liệu cho máy tính được công nhận là tác giả, bên cạnh việc được hưởng quyền tác giả đối với tác phẩm phái sinh, họ phải thực hiện các nghĩa vụ đối với tác giả tác phẩm gốc (tác phẩm nguồn) theo quy định của pháp luật.

Bên cạnh người cung cấp dữ liệu đầu vào cho máy tính, lập trình viên phần mềm thông qua việc tạo ra các thuật toán trí tuệ nhân tạo có khả năng tạo ra tác phẩm cũng đóng vai trò quan trọng để tác phẩm ra đời. Những lập trình viên phần mềm chính là người tạo ra các điều kiện cần thiết cho máy tính thực hiện hoạt động sáng tác. Việc ghi nhận quyền tác giả cho người tạo ra công nghệ sẽ bảo đảm và khuyến khích các công ty tiếp tục đầu tư vào công nghệ, khi khoản đầu tư của họ được bù đắp bằng việc khai thác quyền tác giả.

Như vậy, liên quan đến việc bảo hộ quyền tác giả cho các tác phẩm được tạo ra bởi trí tuệ nhân tạo, các nhà làm luật trên thế giới có thể cân nhắc đến việc ghi nhận và bảo hộ quyền tác giả cho những người đóng vai trò quan trọng và có tính quyết định đến việc tác phẩm do máy tính tạo ra như thế nào, bao gồm: Người thu thập, lựa chọn nguồn dữ liệu đầu vào để đào tạo máy tính và những lập trình viên máy tính.

Trí tuệ nhân tạo ngày nay đang tồn tại hiện hữu ở mọi lĩnh vực đời sống, ngày càng tham gia tích cực và hiệu quả vào hoạt động

sáng tạo ra những sản phẩm trí tuệ, làm phong phú và đa dạng đời sống văn hoá, tinh thần của xã hội, góp phần phát triển nền công nghiệp văn hoá cũng như công nghiệp máy tính. Thực tế đó đòi hỏi pháp luật sở hữu trí tuệ, trong đó có pháp luật quyền tác giả phải có những điều chỉnh phù hợp để khai thác tối đa những lợi ích mà trí tuệ nhân tạo mang lại cho xã hội.

---

[1] https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp, truy cập ngày 26/02/2020.

[2] Stuart J. Russell and Peter Norvig, "*Artificial Intelligence A Modern Approach*", https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwjT95Dd4ILoAhVMUd4KHbQWDQkQFjABegQIAxAB&url=https%3A%2F%2Ffaculty.psau.edu.sa%2Ffiledownload%2Fdoc-7-pdf-a154ffcec538a4161a406abf62f5b76-original.pdf &usg=AOvVaw0i7pLrlBs9LMW296xeV6b0, truy cập ngày 26/02/2020.

[3] Nguyễn Bích Thảo, trong Sách chuyên khảo "*Cách mạng công nghiệp lần thứ tư và những vấn đề đặt ra đối với cải cách pháp luật tại Việt Nam*", Đại học Quốc gia Hà Nội, Khoa Luật, Nxb. Chính trị quốc gia Sự thật, 2018, tr.126.

[4] Andres Guadamuz, University of Sussex, United Kingdom, *"Artificial intelligence and copyright*", October 2017, truy cập ngày 26/02/2020.

[5] Andres Guadamuz, University of Sussex, United Kingdom, *"Artificial intelligence and copyright*", October 2017, https://www.wipo.int/wipo_magazine/ en/2017/05/article_0003.html, truy cập ngày 26/02/2020.

[6] Tổ chức Sở hữu trí tuệ thế giới (WIPO), *"Cẩm nang sở hữu trí tuệ - Chính sách, pháp luật và áp dụng"*, 2005, tr.263.

[7] https://www.cric.or.jp/english/clj/doc/20150227_October,2014_Copyright_Law_of_Japan.pdf, truy cập ngày 26/02/2020.

[8] https://www.gesetze-im-internet.de/englisch_urhg/ englisch_urhg.html, truy cập ngày 26/02/2020.

[9] https://www.admin.ch/opc/en/classified-compilatio n/19920251/index.html, truy cập ngày 26/02/2020.

[10] Nguyễn Vân Nam, "*Quyền tác giả - Đường hội nhập không trải hoa hồng*", Nxb. Trẻ, 2016, tr.75.

[11] https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-912-introduction-to-copyrig ht-law-january-iap-2006/readings/feist.pdf, truy cập ngày 26/02/2020.

[12] https://jade.io/article/262011, truy cập ngày 26/02/2020.

[13] https://h2o.law.harvard.edu/collages/24450, truy cập ngày 26/02/2020.

[14] Andres Guadamuz, University of Sussex, United Kingdom, "*Artificial intelligence and copyright*", October 2017, https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html, truy cập ngày 26/02/2020.

[15] Dani Deahl, *"How AI - Generated music is changing the way hits are made"*, https://www.theverge.com/2018/8/31/17777008/artificial-intelligence-taryn-southern-amper-music, truy cập ngày 26/02/2020.

[16] Ploman, Edward W., and L. Clark Hamilton (1980), "*Copyright: Intellectual Property in the Information Age"*, London: Routledge & Kegan Paul, tr.5.

[17] Nguyễn Vân Nam, "*Quyền tác giả - Đường hội nhập không trải hoa hồng*", Nxb. Trẻ, 2016, tr.124.

[18] Florian Moritz & Dr. Daniela Mohr, *"What Are the Differences between European Copyright and U.S. Copyright?"*, https://www.copytrack.com/euro pean-us-copyright-difference/, truy cập ngày 26/02/2020.

[19] https://thuvienphapluat.vn/van-ban/thuong-mai/Hiep-dinh-khia-canh-lien-quan-toi-thuong-mai-cua-Quyen-so-huu-tri-tue-1994-12722.aspx, truy cập ngày 26/02/2020.

[20] Nguyễn Vân Nam, tlđd, tr.92.

[21] Ngô Huy Cương, trong Sách chuyên khảo "*Cách mạng công nghiệp lần thứ tư và những vấn đề đặt ra đối với cải cách pháp luật tại Việt Nam*", Đại học Quốc gia Hà Nội, Khoa Luật, Nxb. Chính trị quốc gia Sự thật, 2018, tr. 34.

[22] http://images.policy.mofcom.gov.cn/article/201510/1445495480384.pdf, truy cập ngày 26/02/2020.

Nguyên văn tiếng Anh:

Art 9 (3): *"In the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken".*

Art 178: *"Computer-generated, in relation to a work, means that the work is generated by computer in circumstances such that there is no human author of the work".*

[23] Andres Guadamuz, University of Sussex, United Kingdom, *"Artificial intelligence and copyright*", October 2017, https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html, truy cập ngày 26/02/2020.

Tài liệu tham khảo

1. PGS.TS. Nguyễn Thị Quế Anh, PGS.TS. Vũ Công Giao và TS. Mai Văn Thắng (đồng chủ biên), *"Trí tuệ nhân tạo với pháp luật và quyền con người"*, Nxb. Tư pháp, Hà Nội, năm 2019.

2. TS. Phan Chí Hiếu và TS. Nguyễn Văn Cương (đồng chủ biên), *"Cách mạng công nghiệp lần thứ tư và những vấn đề pháp lý đặt ra"*, Nxb. Tư pháp, Hà Nội, năm 2019.

3. Khoa Luật - Đại học Quốc gia Hà Nội, "*Cách mạng công nghiệp lần thứ tư và những vấn đề đặt ra đối với cải cách pháp luật tại Việt Nam*", Nxb. Chính trị quốc gia Sự thật, Hà Nội, 2018.

4. Tổ chức Sở hữu trí tuệ thế giới (WIPO), *"Cẩm nang sở hữu trí tuệ - Chính sách, pháp luật và áp dụng"*, năm 2005.

5. Nguyễn Vân Nam, "*Quyền tác giả - Đường hội nhập không trải hoa hồng*", Nxb. Trẻ, Thành phố Hồ Chí Minh, 2016.

6. Stuart J. Russell and Peter Norvig, "*Artificial Intelligence A Modern Approach*", https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwjT95Dd4ILoAhVMUd4KHbQWDQkQFjABegQIAxAB&url=https%3A%2F%2Ffaculty.psau.edu.sa%2Ffiledownload%2Fdoc-7-pdf-a154ffbcec538a4161a406abf62f5b76-original.pdf&usg=AOvVaw0i7pLrlBs9LMW296xeV6b0, truy cập ngày 26/02/2020.

7. Andres Guadamuz, University of Sussex, United Kingdom, *"Artificial intelligence and copyright*", October 2017, https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html, truy cập ngày 26/02/2020.

8. Dani Deahl, *"How AI - Generated music is changing the way hits are made"*, https://www.theverge.com/2018/8/31/17777008/artificial-intelligence-taryn-southern-amper-music, truy cập ngày 26/02/2020.

9. Ploman, Edward W., and L. Clark Hamilton (1980), "*Copyright: Intellectual Property in the Information Age"*, London: Routledge & Kegan Paul.

10. Florian Moritz & Dr. Daniela Mohr, *"What Are the Differences between European Copyright and U.S. Copyright?"*, https://www.co pytrack.com/euro pean-us-copyright-difference/, truy cập ngày 26/02/2020.

11. Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, Harvard Journal of Law and Technology, Vol. 29, No. 2, Spring 2016.

12. Steven J. Frank, *What AI Practitioners Should Know about the Law*, AI Magazine, Vol. 9, No. 2, 1988.

ARTIFICIAL INTELLIGENCE AND COPYRIGHT PROTECTION

Dr. Tran Van Bien
Institute of State and Law, Vietnam Academy of Social Sciences
Assoc. Prof. Dr. Vu Thi Hai Yen

1. Artificial Intelligence (AI) and works created by AI

Artificial Intelligence (AI) simply means intelligent machines that simulate human intelligence and are programmed to think like humans and imitate their actions. The term can also be used to any machine that exhibits traits related to the human mind such as learning, decision making, or problem solving[1]. If the first robots in the past could only simulate human behaviors, today, the modern generation of robots has outstanding abilities and intelligence like or even better than human, also being capable of thinking, acting… like a human. In the book "Artificial Intelligence: A Modern Approach", the authors analyzed four characteristics of AI, which are: i) Thinking humanly; ii) Thinking rationally; iii) Acting humanly; iv) Acting rationally[2]. Artificial intelligence experts also point out more basic characteristics of artificial intelligence, which are: i) Creativity; ii) Unpredictability; iii) Independence and autonomy in operation, without human intervention; iv) Reasonableness; v) Ability to self-study, constantly improve and develop through interaction with the environment; vi) Data collection and communication capabilities; vii) Efficiency and accuracy; viii) Possibility to freely use alternatives[3].

With those outstanding capabilities, today, artificial intelligence has been able to produce literary and artistic works that make us really surprised and admired. In 2016, a collaborative project between Microsoft and the Rembrandt Museum introduced a work of

a 3D portrait called "The Next Rembrandt" which is computer generated based on facial recognition algorithm by scanning data from 346 known paintings by the 17th century Dutch painter Rembrandt. Also in 2016, a short novel written by a Japanese computer program entered the second round of the national literary award. Google has funded a project that uses artificial intelligence to write news articles for a local newspaper. Google-owned artificial intelligence company Deep Mind has also created software that can compose music by listening to recordings[4]. Today, humanity is witnessing artificial intelligence actually becoming robotic artists capable of writing poetry, painting, composing music and even composing opera plays.

In the past, computers could only create rudimentary artworks. Most computer-generated artworks rely a lot on the creativity of programmers, and machines are just tools like pens (brushes) or canvas to aid in producing the work. Today, the industrial revolution 4.0 has brought technological breakthroughs with the rapid development of self-learning softwares which allow machines to learn from input data, develop and make completely independent or pre-directed decisions themselves, without the need to be specifically programmed by humans[5]. Therefore, in order to create a new literary or artistic work, one only needs to provide the computer with the necessary data, it will learn from this data, make its own independent decisions about how the work will be created in the creative process. Works of this kind, instead of being produced by the human mind, are created by an artificial neural network whose process closely resembles that of human thought.

Although, today's artificial intelligence can have a "brain" like a human, even smarter than normal humans, and can do jobs that only humans could do before, in essence it's still just an "intelligent machine". Therefore, the question is whether the creative works of

literature and art created by artificial intelligence can be protected by copyright - a field of law that was born with the aim of recognizing and protecting human intellectual creative products.

2. Requirements for protection of works under the copyright law

Traditionally, the subject of copyright is the product of creative mental activities (as opposed to the objects of the physical world, which are tangible objects). The Berne Convention is the oldest international copyright convention and has the largest number of member states. Article 2 of the Convention provides an unlimited list of protected works, including original work in the fields of literature, art and science, regardless of form or mode of expression[6]. On the basis of the Berne Convention, the copyright laws of most countries in the world only protect literary, artistic and scientific works that satisfy the requirement that they are original mental creations.

The Copyright Law of Japan, in section (i) Clause 1 Article 2 stipulates: " "Work" means a production in which thoughts or sentiments are expressed in a creative way and which falls within the literary, academic, artistic or musical domain"[7].

Clause 2, Article 2 of the German Copyright Law provides for works protected by copyright as follows: "Only the author's own intellectual creation constitute works within the meaning of this Act"[8].

The Swiss Law on Copyright and Related Rights defines works as follows: "Works are literary and artistic intellectual creations with individual character, irrespective of their value or purposes" (Clause 1, Article 2)[9].

Clause 7, Article 4 of the current Vietnam Intellectual Property Law stipulates: "A work means a creation of the mind in the literary, artistic or scientific domain, whatever may be the mode or form of its

expression" and at Clause 3, Article 14 provides that a protected work "must be crated personally by authors through their intellectual labor without copying others' works".

Although there are different expressions, the Berne Convention and the copyright laws of most countries in the world define the elements for a work to be protected by copyright[10], which are:

(i) Being expressed in a certain form

Creative results, if just being ideas in the author's mind, have not been expressed to the physical world in a certain form, cannot be considered as a copyright-protected work. On the author's side, if the work exists only in the mind of the author and cannot be recognized by others, it does not pose a problem of protection. On the other hand, creative products in the fields of literature, art, and science have meaning only when they are accessible to the public. Therefore, the first requirement for a work to be copyright-protected is that it must be expressed in a certain form so that others can recognize and identify.

(ii) Must result from mental creative activities

This requirement means that a work must be the result of a process of thinking, exploring, creating, and reflecting the thoughts and feelings of the creator. In other words, a work must be a creative result containing a certain intellectual content.

(iii) Originality (or author's personal imprint)

The international laws as well as the copyright laws of most countries in the world only protect works that meet the requirment that they are original mental creations. "Originality" requires that a work created by the author himself being distinguishable from the works of others. In some countries' laws, this requirement is stricter that the

work must bear "unique feature" or "personal imprint", reflected in the content or expresion form of the work, or both. The personal impression can be about the author's thinking or expression form.

Since its inception until now, the principle of copyright law is to give protection to human-created creations associated with the author's personal factors. The products of artificial intelligence are essentially machine-generated, they do not express the thoughts and feelings of a specific human being, therefore, are not considered creative mental outcomes and are excluded from copyright protected subjects.

Laws and practices in many countries have also clearly shown that they do not accept copyright protection for works that are not created by humans. In the United States, the Copyright Office states: “A original work is subject to copyright protection if it is created by a human being.” This statement comes from the case *Feist Publications v. Rural Telephone Service Company, Inc. 499 U.S. 340 (1991)*[11], in which the court ruled that copyright law only protects the fruits of intellectual labor that is based on creativity of a human mind. Likewise, in a recent Australian case (*Acohs Pty Ltd v. Ucorp Pty Ltd*)[12], the court declared a work created with computer intervention would not be protected by copyright because it was not created by humans. In Europe, the Court of Justice of the European Union (CJEU) also made many similar rulings, especially the landmark ruling in the case C-5/08 *Infopaq International A/S v. Danske Dagbaldes Forening*[13], argued that copyright only applies to original works, i.e. it must reflect the author's own intellectual creativity. This means, a necessary requirement for a work to be protected by copyright is that its author must be a specific human being[14]. The denial of copyright protection for computer-generated works comes from the view that copyright is the social incentive for the author, in

addition to economic and material motivation. Artificial intelligence creates works automatically, so there is no incentive and no need for copyright protection.

3. Why is the copyright protection necessary for works created by artificial intelligence

Nowadays, works created by artificial intelligence such as music, newspaper, painting, design or multimedia works such as computer games, etc. tend to increase and they are products of high commercial value in the cultural and entertainment industries, as well as being applied in many fields of life. Compared with ordinary works created by humans, works created by artificial intelligence often require large investment resources by businesses. If those works are not protected by copyright, that means anyone can use and exploit them without having to ask for permission and paying a fee. Businesses that invest in technology to create such works neither have the right to prohibit others from using them, nor be able to obtain fees or material benefits from the exploitation of the works. This not only causes them economic losses, their efforts and investments are not properly compensated, but may certainly have negative impacts on the industry, typically the computer industry. Businesses would no longer be eager to carry out researches, or invest effort and money when the return on their investment is not satisfactory. And if so, the advantages of artificial intelligence in terms of saving time and costs for human resource investment in production and business activities are also not effectively utilized.

Not only negatively affecting the interests of the entities who invest in developing artificial intelligence and the computer industry, the failure to provide copyright protection for products created by artificial intelligence may also affect the interests of society in promoting wealth and enriching cultural and spiritual life. In such case,

the copyright law would fail to fulfill its mission of harmonizing the interests of creative investors with those of exploiters, users as well as the interests of society.

Today's science and technology achievements are bringing about many opportunities in the way of creativity as well as the form of expressing works. Artists increasingly tend to take advantage of modern machines and means, including using artificial intelligence to support their composing activities. In 2017, singer Taryn Southern, YouTube music star, collaborated with AmperCompany - an artificial intelligence (AI) development organization capable of creating music in seconds to complete the album titled "I am AI". The special thing about this album is that the entire music is composed by AmperCompany AI. When completed, it is considered to be quite similar to today's contemporary Pop songs [15] . Today's supercomputers are provided with the ability to learn, analyze, and solve problems themselves like humans, making it difficult to distinguish between human-made and machine-made products. The combination of human intelligence and artificial intelligence in creative activities tends to be more and more widespread, leading to the distinction between human-generated and computer-generated artworks becoming increasingly blurred. This also poses challenges in practice if copyright law still follows the traditional principle of protecting only the individual author's creative mental work.

World history has shown that copyright is formed and developed based on the concept of ownership. In the ancient and medieval periods, the law only protected the ownership of tangible "objects" containing works such as books or paintings, but did not protect the ownership rights of "intellectual creation" [16] . Later, the rights of the creative subject (copyright to creative works) were recognized independently of the ownership rights to the "carrier of intellectual works". The theory of property rights for intellectual labor,

known as the Theory of intellectual property (or mental property) rights born in the eighteenth century, distinguishes manuscripts of works (tangible objects) from works (intangible products) and has shown a close relationship between the author and the work.

Although the subject of copyright is the result of mental creation, copyright still has the feature of a type of ownership right, which are: (1) Absoluteness: Like ordinary property rights, copyright is an absolute right, having effect to all subjects in society, including the State, everyone must respect the granted copyright. Any restriction or limitation of copyright is only possible by law. (2) Proactivity: Copyright is exclusive, only the copyright owner can exploit and use the work. (3) Protection: The owner has the right to prevent anyone from infringing his copyright[17]. Thus, our treatment to works created by artificial intelligence in the direction of only recognizing ownership rights to tangible objects, but not recognizing ownership rights to intellectual creative product is against the development trend of the copyright law.

There are currently two systems of copyright protection in the world. If with the copyright laws of European countries, the economic interests of the author and the author's relationship with his work are the focus of protection, then the Anglo-American legal system of copyright pays special attention to the rights of those who can exploit the work (the right to copy)[18]. In addition to protecting the personal rights (spiritual rights) of the author, copyright law has the task of protecting the property rights (economic rights) for the author and the copyright owner. Today, with the development of the cultural and entertainment industry, the protection of personal rights only plays a secondary role to economic rights. Once the work has entered the market, it becomes a tradable commodity and becomes more and more independent from the authorship. Then, for the author, economic interests are often the top concern. Therefore, if we argue

that, it is not possible to give personal rights such as: The right to be named on the work, the right to name the work, the right to protect the integrity of the work by machines - which are considered inanimate objects, the economic interests of copyright holders still need to be recognized and protected in order to encourage and reward investors of creative activities. For works created by artificial intelligence, this investor is the creator of the computer programs enabling the artificial intelligence to work, the person who provides input data for the computer to learn, the operator of artificial intelligence…

4. How to provide copyright protection to works created by artificial intelligence

4.1. Works to be protected by copyright

To be protected by copyright, works created by artificial intelligence need to meet the minimum requirements for creativity. In addition to protecting traditional literary and artistic works that clearly show the author's mental creation and personal imprint, the current copyright law still makes exceptions with low requirements, typically for computer programsor data collections.

As in paragraph 1 of Article 10 of the TRIPs Agreement, a computer program is protected as a linguistic work as defined by the Berne Convention for a minimum term of 50 years[19]. Although they are protected by copyright law, computer programs do not possess the same aesthetic or artistic properties as other works. It is simply a product of information technology created to solve a certain problem, so it is difficult to prove that a computer program represents a mental content or personal imprint of the author. The basic difference between a computer program and a traditional work is the low level of creativity. A computer program only needs to meet the minimum condition of "creative outcomes" - the most important criterion for a

work to be protected, while the requirements of "mental creation" or "personal imprint" are not important[20]. A data collection is also a work protected by copyright when it is resulted from the collection and selection of works, then information and other data become a unified whole. Similar to computer programs, the criteria of "mental creation" of this work are set very low, only requiring creativity expressed through the collection and arrangement of materials by the collector.

To see if works created by artificial intelligence can be copyrighted, let's compare the two creative processes of humans and artificial intelligence.

Firstly, both creative processes require a source of materials for creation. This source of materials can be considered as a common property of society that everyone can exploit, including: Ideas, themes, cultural products of mankind, and other materials from nature or social life. For example: For literary works, the first source of materials is vocabulary; for musical works it is chords… The different point is that when people create works, they choose materials for their own compositions; while for artificial intelligence, programmers will be the ones to select and provide input data to the computer.

Secondly, both processes need tools to express the work, for example: Language, sound, images, colors... and are often associated with materials that carry it such as: canvas for drawing, paper for printing, discs for containing sound; or sound waves, electromagnetic waves... for radio and television broadcasting. When people create works, they choose tools to express their creative ideas. As for artificial intelligence, the user and operator is the one who chooses the tools for the machine to display the work.

Thirdly, it is the process of producing works. Basing on materials and tools for creativity, depending on his ability and forte, the author must add his own materials - which is his personal imprint,

expressing his thoughts, feelings, and spirit in the work. If the general source of materials and tools for expressing the work can be considered "free materials" that can be exploited by all subjects and not within the scope of copyright protection, then the specific materials that the author put into the work would be protected by copyright. For example: For a painting work, its own materials are the combination of lines, colors, and composition in the picture... For works created by human, the author expresses his or her thoughts, concepts, ideas and feelings through the way of expressing the work. As for works created by artificial intelligence, the computer's brain replaces humans to learn, exploit input data sources, select, process information and decide on the creation of new works by themselves. The results of the two processes are both originally creative works, the only difference is that the product is man-made or AI-made.

If in the past, computers only involved in creative activities as a tool (instead of a pen or drawing paper), today's generations of artificial intelligence can make many decisions related to the creative process without human intervention, in other words, artificial intelligence is a real creative artist. Works created by artificial intelligence can meet the most important criteria, which is "creative", except that it is not "directly created" by humans. Therefore, similar to computer programs or data collections, lawmakers around the world can consider giving works created by artificial intelligence an exception in copyright protection, that is spiritual creativity or personal imprints should not be required like traditional literary and artistic works.

4.2. Subjects to be protected by copyright

In the current context of the fourth industrial revolution, with the widespread participation of "intelligent autonomous robots", the question has been raised is whether to recognize robots as subjects of legal relations or not? In Europe, there have been

recommendations to create a separate legal status for robots, such as giving the status of "electronic person" to the most intelligent autonomous robots[21]. In 2017, Saudi Arabia became the first country in the world to grant citizenship to a humanoid robot which has artificial intelligence. It can be seen that, if the fourth industrial revolution leads to important legal reforms, including the granting of "electronic person" legal status to intelligent robots, then it is inevitable that the copyright law should also consider to recognize copyright of artificial intelligence when creating works besides empowering people - "natural person".

Some countries, notably the UK, express the point of view of realizing copyright to those who make necessary requirements for a work to be performed. Section 9 (3) of The UK Copyright, Design and Patent Act (CDPA) states: *"In the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken".* Moreover, Section 178 of CPDA defines: *"Computer-generated, in relation to a work, means that the work is generated by computer in circumstances such that there is no human author of the work"*[22]*.* The idea behind this provision is to make an exception to the recognition of copyright for works created by programs that are capable of creating works, including entirely machine-generated ones[23]. This can be seen as a new trend in copyright protection for works created by artificial intelligence.

However, while there is controversy about whether to give personal right (belonging to copyright) to computers or not, copyright law can fully recognize and protect economic rights for copyright owners. Works created by artificial intelligence are the products of a process with the "participation" of humans such as: data providers for computers, creators of artificial intelligence software, artificial

intelligence users and operators… For example, in order for a computer to generate a piece of music, we must first feed it with tracks in a predetermined style to serve as a training source for the computer. Along with that, the programmers must use algorithms to help the computer itself adjust the rhythm, style, mood of the music ...

In the process of creating works of artificial intelligence, the source of data for training artificial intelligence is very important. If such data is automatically collected by machines from sources such as the internet or social networks, and then analyzed and created by artificial intelligence itself, then no copyright is required. However, if the data source for artificial intelligence training is collected and selected by a human supplier, then the supplier's copyright must be realized, as well as the data collection's copyright. An important point is that a computer-generated work is a type of derivative work, since it is created from existing works of the computer's input data. Therefore, according to the law, the author of a derivative work must obtain permission and perform the obligation to pay the owner of the original works when he wants to use the original works if they are still in the period of copyright protection. For works created by artificial intelligence, if the person providing the data for the computer is recognized as the author, in addition to enjoying the copyright to the derivative work, they must perform the obligations to the author of the original works (source work) in accordance with the law.

Besides the person who provides input data to the computer, the software programmer, through the creation of artificial intelligence algorithms capable of creating works, also plays an important role in the creation of the work. Software programmers are the ones who create necessary conditions for computers to perform creative activities. Recognition of copyright to technology creators will insure and incentivize companies to continue investing in technology, as their investment is offset by exploitation of copyright.

Thus, regarding the copyright protection for works created by artificial intelligence, the world's lawmakers may consider the recognition and protection of copyright for those who play an important and decisive role in the way a work is created by a computer, including: People who collect and select input data sources to train computers and computer programmers.

Today, artificial intelligence exists in all fields of life, increasingly actively and effectively participating in activities of creating intellectual products, enriching and diversifying cultural and spiritual life of society, contributing to the development of the cultural industry as well as the computer industry. That fact requires intellectual property law, including copyright law, to make appropriate adjustments to exploit the most of the benefits that artificial intelligence brings to society.

---

[1] https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp, accessed on Feb. 26, 2020.

[2] Stuart J. Russell and Peter Norvig, "*Artificial Intelligence A Modern Approach*", https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwjT95Dd4ILoAhVMUd4KHbQWDQkQFjABegQIAxAB&url=https%3A%2F%2Ffaculty.psau.edu.sa%2Ffiledownload%2Fdoc-7-pdf-154ffbcec538a4161a406abf62f5b76-original.pdf&usg=AOvVaw0i7pLrlBs9LMW296xeV6b0, accessed on Feb. 26, 2020.

[3] Nguyen BichThao, in the monograph "Cách mạng công nghiệp lần thứ tư và những vấn đề đặt ra đối với cải cách pháp luật tại Việt Nam" (The Fourth Industrial Revolution and the issues raised for legal reform in Vietnam), Vietnam National University, Hanoi, Faculty of Law, National Politics Truth Publishing House, 2018, p. 126.

[4]Andres Guadamuz,University of Sussex, United Kingdom, *"Artificial intelligence and copyright*", October 2017, accessed on Feb. 26, 2020.

[5] Andres Guadamuz,University of Sussex, United Kingdom, *"Artificial intelligence and copyright*", October 2017, https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html, accessed on Feb. 26, 2020.

[6] World Intellectual Property Organization (WIPO), "*Intellectual Property Handbook: Policy, Law and Use*", 2005, p. 263.

[7] https://www.cric.or.jp/english/clj/doc/20150227_October,2014_Copyright_Law_of_Japan.pdf, accessed on Feb. 26, 2020.

[8] https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html, accessed on Feb. 26, 2020.
[9] https://www.admin.ch/opc/en/classified-compilation/19920251/index.html, accessed on Feb. 26, 2020.
[10] Nguyen Van Nam, "*Quyền tác giả- Đường hội nhập không trải hoa hồng*" (*Copyright - The road of integration is not covered by roses)*, Youth Publishing House, 2016, p.75.
[11] https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-912-introduction -to-copyrig ht-law-january-iap-2006/readings/feist.pdf, accessed on Feb. 26, 2020.
[12] https://jade.io/article/262011, accessed on Feb. 26, 2020.
[13] https://h2o.law.harvard.edu/collages/24450, accessed on Feb. 26, 2020.
[14] Andres Guadamuz,University of Sussex, United Kingdom, "*Artificial intelligence and copyright*", October 2017, https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html, accessed on Feb. 26, 2020.
[15] Dani Deahl, *"How AI - Generated music is changing the way hits are made"*, https://www.theverge.com/2018/8/31/17777008/artificial-intelligence-taryn-southern-amper-music, accessed on Feb. 26, 2020.
[16] Ploman, Edward W., and L. Clark Hamilton (1980), "*Copyright: Intellectual Property in the Information Age"*, London: Routledge & Kegan Paul, p.5.
[17] Nguyen Van Nam, "*Quyền tác giả- Đường hội nhập không trải hoa hồng*" (*Copyright - The road of integration does not spread roses)*, Youth Publishing House, 2016, p. 124.
[18] Florian Moritz & Dr. Daniela Mohr, *"What Are the Differences between European Copyright and U.S. Copyright?"*, https://www.copytrack.com/euro pean-us-copyright-difference/, accessed on Feb. 26, 2020.
[19] https://thuvienphapluat.vn/van-ban/thuong-mai/Hiep-dinh-khia-canh-lien-quan-toi-th uong-mai-cua-Quyen-so-huu-tri-tue-1994-12722.aspx, accessed on Feb. 26, 2020.
[20] Nguyễn Vân Nam, ibid., p.92.
[21] Ngo Huy Cuong, in the monograph "*Cách mạng công nghiệp lần thứ tư và những vấn đề đặt ra đối với cải cách pháp luật tại Việt Nam*" (The fourth industrial revolution and the issues raised for legal reform in Vietnam"), Vietnam National University, Hanoi, Faculty of Law, National Politics Truth Publishing House., 2018, p.34.
[22] http://images.policy.mofcom.gov.cn/article/20151 0/1445495480384.pdf, accessed on Feb. 26, 2020.
[23] Andres Guadamuz, University of Sussex, United Kingdom, *"Artificial intelligence and copyright*", October 2017, https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html, accessed on Feb. 26, 2020.

References

1. Assoc.Prof.Dr. Nguyen Thi Que Anh, Assoc.Prof. Dr. Vu Cong Giao and Dr. Mai Văn Thang (co-editor), *"Trí tuệ nhân tạo với pháp luật và quyền con người"* (Artificial intelligence in relations with legal system and and human rights), Justice Publishing House, Hanoi, 2019.

2. Dr. Phan Chi Hieu and Dr. Nguyen Van Cuong (co-editor), *"Cách mạng công nghiệp lần thứ tư và những vấn đề pháp lý đặt ra"* (The Fourth Industrial Revolution and legal issues), Justice Publishing House, Hanoi, 2019.

3. Vietnam National University, Hanoi, Faculty of Law, "*Cách mạng công nghiệp lần thứ tư và những vấn đề đặt ra đối với cải cách pháp luật tại Việt Nam*" (The Fourth Industrial Revolution and the issues for legal reform in Vietnam), National Politics Truth Publishing House, 2018.

4. World Intellectual Property Organization (WIPO), "*Intellectual Property Handbook: Policy, Law and Use*", 2005.

5. Nguyen Van Nam, "*Quyền tác giả - Đường hội nhập không trải hoa hồng*" (*Copyright - The road of integration is not covered by roses),* Youth Publishing House, 2016.

6. Stuart J. Russell and Peter Norvig, "*Artificial Intelligence: A Modern Approach*", https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwjT95Dd4ILoAhVMUd4KHbQWDQkQFjABegQIAxAB&url=https%3A%2F%2Ffaculty.psau.edu.sa%2Ffiledownload%2Fdoc-7-pdf-a154ffbcec538a4161a406abf62f5b76-original.pdf&usg=AOvVaw0i7pLrlBs9LMW296xeV6b0, accessed on Feb. 26, 2020.

7. Andres Guadamuz*,* University of Sussex, United Kingdom, *"Artificial intelligence and copyright*", October 2017*,* https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html, accessed on Feb. 26, 2020.

8. Dani Deahl, *"How AI - Generated music is changing the way hits are made"*, https://www.theverge.com/2018/8/31/17777008/artificial-intelligence-taryn-southern-amper-music, accessed on Feb. 26, 2020.

9. Ploman, Edward W., and L. Clark Hamilton (1980), "*Copyright: Intellectual Property in the Information Age"*, London: Routledge & Kegan Paul.

10. Florian Moritz & Dr. Daniela Mohr, *"What Are the Differences between European Copyright and U.S. Copyright?"*, https://www.copytrack.com/european-us-copyright-difference/, accessed on Feb. 26, 2020.

11. Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, Harvard Journal of Law and Technology, Vol. 29, No. 2, Spring 2016.

12. Steven J. Frank, *What AI Practitioners Should Know about the Law*, AI Magazine, Vol. 9, No. 2, 1988.

Thông tin Tác giả
Authors' Information

Ông Christian Manhart mang quốc tịch Đức, là Trưởng Đại diện UNESCO tại Việt Nam từ tháng 10 năm 2021. Trước đó, từ tháng 8 năm 2014, ông là Trưởng Đại diện UNESCO tại Nepal. Trong thời gian này, ông đã hỗ trợ chính phủ nước sở tại tái xây dựng di sản văn hóa sau trận động đất năm 2015, đảm nhiệm vai trò thiết yếu trong việc thi hành chiến lược giáo dục quốc gia và hỗ trợ truyền thông. Ông có bằng Thạc sĩ chuyên ngành Lịch sử Nghệ thuật và Khảo cổ học tại Đại học Munich và Đại học Sorbonne (Paris) và bắt đầu làm việc tại UNESCO từ năm 1987 với các hoạt động trong lĩnh vực bảo tồn di tích lịch sử tại Nam Á và Trung Á như đền Taj Mahal ở Ấn Độ, Chogha Zanbil ở Iran, hoặc Bamiyan ở Afghanistan. Trong nhiều năm, ông chịu trách nhiệm thực hiện các dự án do Tổng Giám đốc khu vực Địa Trung Hải khởi xướng, nhằm nâng cao sự gắn kết xã hội trong khu vực thành thị. Năm 2005, ông được bổ nhiệm là Trưởng ban Truyền thông và Quan hệ đối tác tại Ủy ban Di sản thế giới UNESCO. Tiếp đó, ông chịu trách nhiệm các hoạt động về bảo tàng và công ước quốc tế về bảo tồn di sản văn hóa: trong bối cảnh xung đột, phòng chống lưu thông bất hợp pháp tài sản văn hóa, di sản văn hóa dưới nước và Ủy ban Liên Chính phủ về Hoàn trả và Bồi thường Tài sản Văn hóa.

Mr. Christian Manhart, a German national, took up his assignment as UNESCO representative to Viet Nam in October 2021. Before he was since August 2014 UNESCO representative to Nepal, where he assisted the government in cultural heritage rebuilding following the

2015 earthquake, took a major role in the implementation of the nation's education plan and supporting the media. He has a master's in art history and archaeology from the universities in Munich and Sorbonne in Paris, and joined UNESCO in 1987 where he was mostly working for the conservation of historical sites in South and Central Asia, such as the Taj Mahal in India, Chogha Zanbil in Iran, or Bamiyan in Afghanistan. For several years he conducted projects initiated by the Director General in the Mediterranean to enhance social cohesion in cities. In 2005, he has been appointed to the UNESCO World Heritage Centre, as chief of communication and partnerships and later in charge of museums and different international conventions for the protection of cultural heritage: in the event of armed conflict, against the illicit traffic of cultural property, the underwater cultural heritage and the Intergovernmental Committee for Return and Restitution of Cultural Property.

Giáo sư Tiến sĩ Đặng Nguyên Anh là Phó Chủ tịch Viện Hàn lâm Khoa học xã hội Việt Nam. Ông là Phó Chủ tịch Ủy ban Quốc gia UNESCO Việt Nam, Trưởng Tiểu ban UNESCO về Khoa học Xã hội, Phó Chủ tịch kiêm Tổng Thư ký Hội Xã hội học Việt Nam. Giáo sư Đặng Nguyên Anh tốt nghiệp Cử nhân Kinh tế tại Đại học Kinh tế Quốc dân Việt Nam năm 1983, Thạc sĩ năm 1994 và Tiến sĩ năm 1997 về Xã hội học tại Đại học Brown, Hoa Kỳ. Ông được phong học hàm Giáo sư về Xã hội học từ năm 2018.

Các lĩnh vực nghiên cứu chuyên môn của Giáo sư Đặng Nguyên Anh bao gồm (i) di cư, đô thị hóa và tái định cư con người; (ii) các vấn đề về môi trường, biến đổi khí hậu và xóa đói giảm nghèo; (iii) phát triển giáo dục thanh niên và vị thành niên; và (iv) dân số, sức khỏe và bảo trợ xã hội. Giáo sư đã từng chủ trì nhiều dự án nghiên cứu được tài trợ và dùng để tham vấn cho các tổ chức quốc tế, các cơ quan của chính phủ. Giáo sư Đặng Nguyên Anh đã xuất bản hàng trăm xuất bản phẩm trong nước và quốc tế.

Professor Dr. Dang Nguyen Anh is Vice President of Vietnam Academy of Social Sciences. He is also Vice Chairman of the Vietnam National Commission for UNESCO, Chairman of the UNESCO Sub-Commission for Social Sciences, and Vice President cum General Secretary of Vietnam Sociological Association. He earned his Bachelor in Economics from Vietnam National Economic University in 1983, Master degree in 1994 and Doctoral degree in 1997 both in Sociology from Brown University, the US. And he became full professor of Sociology in 2018.

Professor Dang's fields of expertise include (i) migration, urbanization, and human resettlement; (ii) environmental risks, climate change, and poverty alleviation; (iii) youth, education, and adolescent development; and (iv) population, health and social protection. He has been the principle investigator of many research projects funded by and provided consultancy services to international organizations, governmental ministries and departments. Professor Dang has also published hundreds of journal articles, book chapters, and books in both international and national publications.

Ông Mai Phan Dũng hiện là Vụ trưởng Vụ Văn hóa Ngoại giao và UNESCO, Bộ Ngoại giao Việt Nam (MOFA). Ông cũng là Tổng Thư ký Ủy ban Quốc gia UNESCO Việt Nam.

Ông tốt nghiệp Trường Chuyên Ngoại ngữ năm 1994 và Đại học Luật, Hà Nội năm 2002. Ông thông thạo cả tiếng Anh và tiếng Pháp. Trong sự nghiệp của mình, ông đã trải qua các chức vụ quan trọng như Phó Vụ trưởng, Vụ Ngoại giao Văn hóa và UNESCO (MOFA), Tham tán kiêm Chánh Văn phòng Đại sứ quán Việt Nam tại Pháp, Cục phó Cục Ngoại vụ, Phó Vụ trưởng Vụ các tổ chức quốc tế (MOFA), v.v.

Mr. Mai Phan Dung is currently the Director General of Department of Cultural Diplomacy and UNESCO, Ministry of Foreign Affairs of Vietnam (MOFA). He is also the General Secretary of Vietnam National Commission for UNESCO.

He graduated from Foreign Language Specialized School in 1994 and from University of Law, Hanoi in 2002. He is fluent in both English and French. During his career, he has gone through different important positions such as Deputy Director General of Department of Cultural Diplomacy and UNESCO (MOFA), Counselor and Head of Chancellery at the Vietnamese Embassy in France, Deputy Director General of Provincial Department, Department of International Organizations (MOFA), etc.

Thượng tướng Nguyễn Văn Thành là nguyên Thứ trưởng Bộ Công an Việt Nam. Ông hiện là Phó Chủ tịch Hội đồng Lý luận Trung Ương.

PGS. TS Nguyễn Văn Thành tốt nghiệp cử nhân Luật và Anh văn, Kỹ sư kinh tế công nghiệp; Thạc sĩ kinh tế công nghiệp; và Tiến sĩ kinh tế. Ông từng có thời gian tham dự khóa học nâng cao tại Volgograd, Nga (1986-1988).

Ông là Chủ tịch Ủy ban Nhân dân thành phố Hải Phòng (2009-2010) và Bí thư Thành ủy Hải Phòng (2010-2014), Phó Chánh Văn phòng Ban Chấp hành Trung ương Đảng (2014-2015), Thứ trưởng Bộ Công an (2015-2021).

PGS. TS. Nguyễn Văn Thành là Thành viên cao cấp của Viện Quản lý Malik, Thụy Sĩ; là thành viên của Hội Khoa học Hệ thống Hải Phòng (HPSSS), Hiệp hội khoa học hệ thống thế giới (ISSS), Hiệp hội Kinh tế và Môi trường Đại dương...; Ông cũng là thành viên sáng lập của tổ chức WCCD (Hội đồng xây dựng và quản trị bộ chỉ số dữ liệu đô thị thế giới Trụ sở tại Toronto Canada) và Cố vấn cao cấp của Ban cố vấn WCCD. Ông là chuyên gia về Tư duy hệ thống, Khoa học Hệ thống, Phát triển bền vững, Đô thị thông minh và là Giảng viên thỉnh giảng của nhiều trường đại học khác nhau trong và ngoài nước.

Ông là tác giả của một số ấn phẩm như: Tư duy hệ thống cho mọi người hành trình từ lý thuyết đến thực tiễn (xuất bản năm 2015 và tái bản lần thứ 3 năm 2017); Tác động của chính sách công nghiệp nhằm nâng cao khả năng cạnh tranh của doanh nghiệp công nghiệp (xuất bản năm 2015 và lần 3 tái bản năm 2017); Quan điểm và giải pháp cho nạn thất nghiệp (Chương 5 - Tạo việc làm cho người lao động tại Hải Phòng, Việt Nam), IntechOpen xuất bản tháng 8/2018 tại

Anh và Croatia; Xây dựng và Quản trị thành phố thông minh (xuất bản năm 2019).

Senior Lieutenant General Assoc. Prof. Dr. Nguyen Van Thanh is the former Deputy Minister of the Ministry of Public Security of Vietnam. He is now Vice Chairman of the Central Theoretical Council.

Assoc. Prof. Dr. Nguyen Van Thanh has bachelor's degrees of Laws and English Studies, Engineering and Economics; Master of Science; and PhD in Economics. He also studied the Advanced Level Study Course in Volgograd, Russia (1986-1988).

He was the Chairman of the People's Committee of Hai Phong City (2009-2010) and the Secretary of the Communist Party of Hai Phong (2010-2014), the Deputy Head of the Office of Central Committee of the Communist Party of Vietnam (2014-2015), the Deputy Minister of Public Security of Vietnam (2015-2021).

Assoc. Prof. Dr. Nguyen Van Thanh is a Senior Fellow of the Malik Management Institute, Switzerland; a member of the Hai Phong Systems Science Society (HPSSS), International Society for the System Science (ISSS), Ocean Economy and Environment Association etc.; a founding member of the World Council on Cities Data (WCCD) and a Senior Advisor of WCCD Advisory Board. He is a specialist in Systems Thinking, Systems Science, Sustainable Development, Smart City and a visiting Lecturer of various universities in and out of Vietnam.

He is the author of a number of books such as: Systems Thinking for Everyone (published in 2015 & 3rd reprint in 2017); Impact of Industrial Policies on enhancing competitive capacity of industrial enterprises (published in 2015 & 3rd reprint in 2017); Unemployment Perspective and Solutions (Chapter 5 - Job Creation in Haiphong, Vietnam), IntechOpen published August 2018 in UK and Croatia; Building and Developing Smart Cities (published in 2019).

Tiến sĩ Nguyễn Phi Lê nhận bằng kỹ sư năm 2007 và thạc sĩ năm 2010 của Đại học Tokyo, Nhật Bản. Cô nhận bằng Tiến sĩ chuyên ngành Tin học của Trường Cao học Nghiên cứu Cao cấp, Viện Tin học Quốc gia, Tokyo, Nhật Bản năm 2019. Hiện tại, Tiến sĩ Nguyễn Phi Lê là giảng viên công nghệ thông tin tại Khoa Công nghệ Thông tin và Truyền thông, Đại học Bách Khoa Hà Nội.

Cô cũng là Giám đốc điều hành của Trung tâm Nghiên cứu Quốc tế về Trí tuệ Nhân tạo (BK.AI), Đại học Bách khoa Hà Nội (HUST).

Lĩnh vực nghiên cứu của Tiến sĩ Nguyễn Phi Lê bao gồm mạng thông tin và truyền thông, ứng dụng AI mạng, mạng Internet vạn vật, mạng điện toán biên, phân tích và khai phá dữ liệu.

Dr. Nguyen Phi Le received her B.E. and M.S. degrees from the University of Tokyo in 2007 and 2010, respectively. She received her Ph.D. in Informatics from The Graduate University for Advanced Studies, National Institute of Informatics, Tokyo, Japan, in 2019. Currently, she is an assistant professor at the School of Information and Communication, Hanoi University of Science and Technology, Vietnam.

She also serves as the managing director of The International Research Center for Artificial Intelligence (BK.AI), HUST.

Her research interests include network architectures, applied AI in networking, Internet of Thing networks, mobile edge computing networks, crowdsensing, data mining.

Tiến sĩ Hoàng Vũ Linh Chi là cán bộ nghiên cứu tại Phòng Văn hóa - Giáo dục thuộc Viện Xã hội học, Viện Hàn lâm Khoa học Xã hội Việt Nam.

Nghiên cứu của cô tập trung vào vai trò của văn hóa và các mối quan hệ giữa văn hóa và phát triển kinh tế - xã hội trong bối cảnh Cách mạng Công nghiệp 4.0. Tiến sĩ Hoàng Vũ Linh Chi đặc biệt quan tâm nghiên cứu mối quan hệ nhân quả giữa các giá trị văn hóa và hành vi của hộ gia đình.

Dr. Hoang Vu Linh Chi is a researcher at Cultural and Educational Department in Institute of Sociology, Vietnam Academy of Social Sciences.

Her research focuses on the role of culture and the relations between culture and socio-economic development in the context of Industry 4.0. She is particularly interested in understanding the causal relations between cultural values and housing behavior.

Phó Giáo sư, Tiến sĩ Cao Thu Hằng hiện đang công tác tại Tạp chí Cộng sản. Trước khi chuyển sang Tạp chí Cộng sản, Phó Giáo sư đã có thời gian dài công tác tại Viện Triết học, Viện Hàn lâm Khoa học xã hội Việt Nam. Phó Giáo sư Cao Thu Hằng quan tâm nghiên cứu về các lĩnh vực: giá trị, đạo đức và văn hóa. Phó Giáo sư là tác giả của một số cuốn sách, bài báo nghiên cứu và đồng chủ biên một số bộ sách liên quan đến các chủ đề giá trị, đạo đức và văn hóa, bao gồm *Xây dựng con người Việt Nam từ góc nhìn truyền thống*; *Tính dân tộc, tính hiện đại và tính nhân văn trong quá trình xây dựng nền văn hóa Việt Nam trong thời kỳ toàn cầu hóa hiện nay*… Ngoài công việc biên tập, nghiên cứu, Phó Giáo sư Cao Thu Hằng còn tham gia giảng dạy tại một số trường đại học, cơ sở đào tạo tại Việt Nam và là Ủy viên Ban Chấp hành Hội Triết học Việt Nam, nhiệm kỳ 2020-2025.

Assoc. Prof. Dr. Cao Thu Hang is currently working at the *Communist Review*. Before joining the *Review*, she had worked for a long time at the Institute of Philosophy, Vietnam Academy of Social Sciences. Her main research interests are in the following areas: values, ethics, and culture. She is the author of several books, research papers and co-editor of several volumes related to the topics of values, ethics and cultural themes, including *Building the Vietnamese Man from the a traditional perspective; Nationality, modernity and humanity in the process of building Vietnamese culture in the current globalization era*… In addition to editing and research work, she also teaches at a number of universities and academies in Vietnam. Assoc.Prof.Dr. Cao Thu Hang is currently a member of the Executive Committee of the Philosophical Society (Vietnam), term 2020-2025.

Tiến sĩ Ngô Vĩnh Bạch Dương là chuyên gia luật cao cấp: Luật gia chuyên nghiệp. Năm 1998, Tiến sĩ Dương bắt đầu sự nghiệp giảng dạy tại Việt Nam, đồng thời từ đó đến nay tham gia giảng dạy về tổ chức và thủ tục tư pháp tại Khoa Luật, Đại học Quốc gia Hà Nội.

Trong 25 năm qua, Tiến sĩ Dương đã hành nghề luật sư với tư cách là người tư vấn về Pháp lý, Cải cách thể chế tại Việt Nam. Tiến sĩ đã hợp tác trực tiếp và tích cực với các cơ quan trung ương như Bộ Tư pháp, Tòa án Nhân dân Tối cao, Viện Kiểm sát Nhân dân Tối cao và Văn phòng Chính phủ trong khuôn khổ nhiều dự án. Tiến sĩ Dương có kiến thức chuyên sâu và nhiều kinh nghiệm làm việc với các chương trình ODA và các nhà tài trợ quốc tế.

Tiến sĩ Ngô Vinh Bạch Dương cũng là một nhà phân tích chính sách và pháp luật của Việt Nam. Bên cạnh công việc tư vấn, Tiến sĩ hiện còn làm công tác nghiên cứu tại Viện Nhà nước và Pháp luật (ISL), Viện Hàn lâm Khoa học Xã hội Việt Nam (VASS) và chịu trách nhiệm tư vấn luật và chính sách về doanh nghiệp và tiêu dùng, tham gia trình bày các báo cáo nghiên cứu và xuất bản sách báo về những lĩnh vực này.

Dr. Ngo Vinh Bach Duong is a senior legal expert: a Jurist by profession. In 1998, Dr. Duong started his teaching career in Vietnam, and has taught judicial organisation and procedures at the School of Law, Vietnam National University (VNU) in Hanoi ever since.

For the last 25 years he has practiced law as a Legal, Institutional Reform consultant in Vietnam. He has collaborated with central

governmental bodies, such as the MoJ, SPC, SPP, and the Office of Government directly and actively in his projects´ frame work. Duong has also good knowledge and rich experience in working with ODA programmes and international donors.

Dr. Ngo Vinh Bach Duong is also a Vietnamese law and policy analyst. Besides consultancy work, he now also works for the Institute of State and Law (ISL), Vietnam Academy of Social Sciences (VASS) and responsible for advising on business, consumer law and policy, presenting research reports and publishing books and articles in these areas.

TS. Trần Văn Biên tốt nghiệp cử nhân tại Khoa Luật - Đại học Quốc gia Hà Nội, thạc sỹ tại Viện Nhà nước và Pháp luật, tiến sỹ tại Khoa Luật - Học viện Khoa học xã hội.

Trong hoạt động nghề nghiệp, TS. Trần Văn Biên có 20 năm kinh nghiệm nghiên cứu về pháp luật dân sự, kinh tế, giao dịch điện tử; biên tập và xuất bản tạp chí chuyên ngành luật; đào tạo đại học và sau đại học ngành luật. TS. Trần Văn Biên đã công bố hơn 30 bài báo khoa học trên các tạp chí chuyên ngành luật; là tác giả hoặc đồng tác giả của gần 20 cuốn sách về luật học. Hiện TS. Trần Văn Biên là Tổng biên tập, Tạp chí Nhà nước và Pháp luật, Viện Nhà nước và Pháp luật.

Dr. Tran Van Bien graduated with a bachelor's degree from the Faculty of Law - Vietnam National University, Hanoi; a master's degree from the Institute of State and Law; and a doctorate from the Faculty of Law - Graduate Academy of Social Sciences.

For professional activities, Dr. Tran Van Bien has 20 years of research experience in the field of civil law, economic law, and law on electronic transactions; in editing and publishing specialized law journals; in educating undergraduate and postgraduate in law. Dr. Tran Van Bien has published more than 30 scientific articles in specialized law journals; is the author or co-author of nearly 20 law books. Currently, Dr. Tran Van Bien is the Editor in Chief of the State and Law Review, Institute of State and Law.

Dự thảo Khuyến nghị
về Đạo đức của Trí tuệ Nhân tạo

Draft text of the Recommendation
on the ethics of AI

DỰ THẢO VĂN BẢN KHUYẾN NGHỊ VỀ ĐẠO ĐỨC TRONG TRÍ TUỆ NHÂN TẠO

(Bản dịch tham khảo)

PHẦN MỞ ĐẦU

Đại hội đồng Của Tổ chức Giáo dục, Khoa học và Văn hóa Liên hợp quốc (UNESCO), tại phiên họp lần thứ 41 tại Paris từ ngày 9 đến 24 tháng 11,

Ghi nhận những tác động tích cực và tiêu cực sâu sắc và rõ rệt của trí tuệ nhân tạo (AI) đối với xã hội, môi trường, hệ sinh thái và cuộc sống của con người, bao gồm cả tinh thần, một phần vì những cách thức mới mà việc sử dụng AI ảnh hưởng đến suy nghĩ, sự tương tác và việc đưa ra quyết định của con người cũng như ảnh hưởng đến giáo dục, khoa học tự nhiên, khoa học xã hội và nhân văn, văn hóa và thông tin - truyền thông,

Nhắc lại, theo các điều khoản của Hiến chương, UNESCO tìm tòi đóng góp cho hòa bình và an ninh bằng cách thúc đẩy sự phối hợp giữa các quốc gia thông qua các hoạt động giáo dục, khoa học, văn hóa, truyền thông và thông tin, để thể hiện sự tôn trọng hơn với công lý, pháp quyền, nhân quyền và các quyền tự do cơ bản được khẳng định cho các dân tộc trên thế giới,

Tin tưởng Khuyến nghị được trình bày ở đây, với vai trò là công cụ tạo ra tiêu chuẩn được xây dựng thông qua cách tiếp cận toàn cầu, dựa trên luật pháp quốc tế, tập trung vào phẩm giá con người và nhân quyền, cũng như bình đẳng giới, công bằng và phát triển xã hội và kinh tế, sức khỏe thể chất và tinh thần, sự đa dạng, kết nối, tính bao trùm, bảo vệ môi trường và hệ sinh thái có thể định hướng các công nghệ AI phát triển theo xu thế có trách nhiệm,

Được định hướng bởi các mục đích và nguyên tắc của Hiến chương Liên hợp quốc,

Xem xét việc các công nghệ AI có thể phục vụ tốt cho con người và tất cả các quốc gia có thể được hưởng lợi từ các công nghệ AI, nhưng cũng làm gia tăng mối lo về đạo đức cơ bản, ví dụ về những thành kiến mà AI có thể liên quan và làm trầm trọng thêm, khả năng dẫn đến việc phân biệt đối xử, bất bình đẳng, các ranh giới số, loại trừ và tạo ra mối đe dọa đối với đa dạng văn hóa, xã hội và sinh học cũng như sự chia rẽ về xã hội hoặc kinh tế; nhu cầu về sự minh bạch và dễ hiểu về phương thức hoạt động của các thuật toán và dữ liệu mà chúng đã được đào tạo; và tác động tiềm ẩn lên các yếu tố như phẩm giá con người, nhân quyền và các quyền tự do cơ bản, bình đẳng giới, dân chủ, xã hội, kinh tế, chính trị và văn hóa, thực tiễn khoa học và kỹ thuật, phúc lợi cho loài vật, môi trường và hệ sinh thái,

Ghi nhận rằng các công nghệ AI có thể làm sâu sắc thêm sự chia rẽ và bất bình đẳng hiện có trên thế giới, bên trong từng quốc gia và giữa các quốc gia, công lý, niềm tin và sự công bằng phải được duy trì để không quốc gia nào và không cá nhân nào bị bỏ lại phía sau, bằng cách tiếp cận công bằng với các công nghệ AI và tận hưởng lợi ích mà chúng đem lại hoặc trong việc bảo vệ chống lại những tác động tiêu cực của chính AI, với việc ghi nhận những điều kiện khác nhau của các quốc gia khác nhau và tôn trọng mong muốn của một số cá nhân không tham gia vào tất cả những tiến bộ công nghệ,

Thừa nhận thực tế là tất cả các quốc gia đang phải đối mặt với sự gia tăng trong sử dụng các công nghệ thông tin và truyền thông cũng như công nghệ AI, nhu cầu ngày càng tăng về kiến thức thông tin và truyền thông, và nền kinh tế số đặt ra những thách thức xã hội, kinh tế và môi trường quan trọng và cơ hội chia sẻ lợi ích, đặc biệt là đối với các nước có thu nhập thấp và trung bình (LMICs), kể cả các

nước phát triển thấp nhất (LLDCs), các nước đang phát triển trong lục địa (LLDCs) và các quốc đảo nhỏ đang phát triển (SIDS), đòi hỏi sự công nhận, bảo vệ và thúc đẩy các nền văn hóa, giá trị và kiến thức nội sinh để phát triển nền kinh tế số bền vững,

Nhận thức rõ hơn rằng các công nghệ AI có tiềm năng mang lại lợi ích cho môi trường và hệ sinh thái, đông thời để có những lợi ích đó, các tác hại tiềm tàng và tác động tiêu cực đến môi trường và hệ sinh thái không nên bị bỏ qua mà phải được giải quyết,

Lưu ý việc giải quyết các rủi ro và mối quan tâm về đạo đức không nên cản trở đổi mới và phát triển mà tạo ra các cơ hội mới và khuyến khích nghiên cứu và đổi mới được tiến hành về mặt đạo đức níu giữ các công nghệ AI trong nhân quyền và các quyền tự do, các giá trị và nguyên tắc, phản ánh tính luân lý và đạo đức,

Nhớ lại tháng 11 năm 2019, Đại hội đồng UNESCO, tại phiên họp thứ 40, đã thông qua Nghị quyết 40C/Resolution 37, theo đó yêu cầu Tổng giám đốc UNESCO "xây dựng một công cụ thiết lập chuẩn quốc tế về đạo đức trong trí tuệ nhân tạo (AI) dưới dạng một khuyến nghị", để trình lên Đại hội đồng tại phiên họp thứ 41 vào năm 2021,

Công nhận sự phát triển của các công nghệ AI đòi hỏi sự gia tăng tương ứng về dữ liệu, phương tiện truyền thông và kiến thức thông tin cũng như truy cập vào các nguồn thông tin độc lập, đa chiều, đáng tin cậy, bao gồm một phần nỗ lực giảm thiểu rủi ro thông tin sai lệch, thông tin giả mạo và ngôn từ kích động thù địch, và thiệt hại gây ra thông qua việc lạm dụng dữ liệu cá nhân,

Quan sát thấy một khuôn khổ quy phạm cho các công nghệ AI và ý nghĩa xã hội của nó có cơ sở trong các khuôn khổ pháp lý quốc tế và quốc gia, các quyền con người và các quyền tự do cơ bản, đạo đức, mong muốn tiếp cận dữ liệu, thông tin và kiến thức, tự do nghiên cứu và đổi mới, sự thịnh vượng của con người, của môi trường và hệ sinh thái, kết nối các giá trị và nguyên tắc đạo đức với

những thách thức và cơ hội liên quan đến công nghệ AI, trên cơ sở các mục tiêu chung và hiểu,

Cũng công nhận các giá trị và nguyên tắc đạo đức có thể giúp phát triển và thực hiện các biện pháp chính sách dựa trên quyền và các chuẩn mực pháp lý, bằng cách đưa ra hướng dẫn với quan điểm về tốc độ phát triển công nghệ nhanh chóng,

Cũng tin tưởng rằng các tiêu chuẩn đạo đức được chấp nhận trên toàn cầu cho các công nghệ AI, với sự hoàn toàn tôn trọng luật pháp quốc tế, đặc biệt là luật nhân quyền, có thể đóng một vai trò quan trọng trong việc xây dựng các chuẩn mực liên quan đến AI trên toàn cầu,

Ghi nhớ Tuyên ngôn Quốc tế Nhân quyền (1948), các văn kiện của khuôn khổ nhân quyền quốc tế, bao gồm Công ước về vị thế của người tị nạn (1951), Công ước về phân biệt đối xử trong việc làm và nghề nghiệp (1958), Công ước quốc tế về xóa bỏ mọi hình thức phân biệt chủng tộc (1965), Công ước quốc tế về quyền dân sự và chính trị (1966), Công ước quốc tế về các quyền kinh tế, xã hội và văn hóa (1966), Công ước xoá bỏ mọi hình thức phân biệt đối xử đối với phụ nữ (1979), Công ước về quyền trẻ em (1989) và Công ước về quyền của người khuyết tật (2006), Công ước chống phân biệt đối xử trong giáo dục (1960), Công ước về bảo vệ và phát huy sự đa dạng của biểu đạt văn hóa (2005), cũng như bất kỳ văn bản, khuyến nghị và tuyên bố quốc tế có liên quan khác,

Đồng thời lưu ý Tuyên bố của Liên Hợp Quốc về Quyền phát triển (1986); Tuyên bố về trách nhiệm của thế hệ hiện tại đối với các thế hệ tương lai (1997); Tuyên ngôn quốc tế về đạo đức sinh học và nhân quyền (2005); Tuyên bố của Liên Hợp Quốc về quyền của các dân tộc bản địa (2007); Nghị quyết của Đại hội đồng Liên Hợp Quốc về Hội nghị thượng đỉnh về Xã hội thông tin (A/RES/70/125) (2015); Nghị quyết của Đại hội đồng Liên hợp quốc về chuyển đổi thế giới

của chúng ta: Chương trình nghị sự 2030 vì phát triển bền vững (A/RES/70/1) (2015); Khuyến nghị liên quan đến bảo tồn và tiếp cận di sản văn bản bao gồm hình thức kỹ thuật số (2015); Tuyên bố các nguyên tắc đạo đức liên quan đến biến đổi khí hậu (2017); Khuyến nghị về khoa học và các nhà nghiên cứu khoa học (2017); Các chỉ số phổ quát Internet (được xác nhận bởi Chương trình quốc tế của UNESCO cho sự phát triển của truyền thông năm 2018), bao gồm nguyên tắc ROAM (được Đại hội đồng UNESCO thông qua năm 2015); Nghị quyết của Hội đồng Nhân quyền về "Quyền riêng tư trong thời đại kỹ thuật số" (A/HRC/RES/42/15) (2019); và Nghị quyết của Hội đồng Nhân quyền về "Công nghệ kỹ thuật số mới và mới nổi cùng các quyền con người" (A/HRC/RES/41/11) (2019),

Nhấn mạnh cần đặc biệt chú ý đến LMICs, bao gồm cả LDCs, LLDCs và SIDS, vì các quốc gia này có năng lực riêng nhưng đã không được đại diện trong cuộc trao đổi về đạo đức AI, điều này làm dấy lên mối lo ngại về việc bỏ qua những hiểu biết về địa phương, sự đa dạng văn hóa, các hệ giá trị và nhu cầu có sự công bằng toàn cầu để đối phó với các tác động tích cực cũng như tiêu cực của các công nghệ AI,

Cũng thấy rõ hiện có nhiều chính sách quốc gia, các khuôn khổ và sáng kiến khác được xây dựng bởi các cơ quan Liên Hợp Quốc có liên quan, các tổ chức liên chính phủ, bao gồm cả các tổ chức khu vực, cũng như các chính sách của khu vực tư nhân, các tổ chức nghề nghiệp, các tổ chức phi chính phủ và cộng đồng khoa học, liên quan đến đạo đức và quy định về công nghệ AI,

Ngoài ra tin tưởng rằng các công nghệ AI có thể mang lại lợi ích quan trọng, nhưng việc đạt được chúng cũng có thể làm giảm căng thẳng xung quanh sự đổi mới, tiếp cận bất đối xứng với kiến thức và công nghệ, bao gồm thâm hụt kiến thức số của người dân làm hạn chế khả năng tham gia của công chúng vào các chủ đề liên quan đến AI, cũng như các rào cản để tiếp cận thông tin và khoảng

trống về khả năng, năng lực của con người và thể chế, rào cản đối với việc tiếp cận đổi mới công nghệ, thiếu cơ sở hạ tầng cơ học và kỹ thuật số thích hợp và các khung pháp lý, bao gồm cả những khung pháp lý liên quan đến dữ liệu, tất cả những điều này đều cần được giải quyết,

Nhấn mạnh tăng cường hợp tác và đoàn kết toàn cầu, bao gồm hành động thông qua chủ nghĩa đa phương, là cần thiết để tạo điều kiện tiếp cận công bằng với các công nghệ AI và giải quyết những thách thức mà AI mang lại cho sự đa dạng và kết nối của các nền văn hóa và hệ thống đạo đức, để giảm thiểu việc lạm dụng tiềm ẩn, để nhận ra tiềm năng đầy đủ mà AI có thể mang lại, đặc biệt là trong lĩnh vực phát triển, và để đảm bảo rằng các chiến lược AI quốc gia được định hướng bởi các nguyên tắc đạo đức,

Có cân nhắc đầy đủ sự phát triển nhanh chóng của các công nghệ AI thách thức việc thực hiện và quản trị đạo đức của họ, cũng như sự tôn trọng và bảo vệ sự đa dạng văn hóa, và có khả năng phá vỡ các tiêu chuẩn và giá trị đạo đức bản địa và khu vực,

1. ***Thông qua*** Khuyến nghị hiện tại về đạo đức trong trí tuệ nhân tạo;

2. ***Khuyến nghị các*** quốc gia thành viên áp dụng trên cơ sở tự nguyện các quy định của Khuyến nghị này bằng cách thực hiện các bước thích hợp, bao gồm biện pháp lập pháp hoặc các biện pháp cần thiết khác, phù hợp với thực tiễn của hiến pháp và các cấu trúc quản lý của từng quốc gia, để có hiệu lực trong phạm vi quyền hạn của họ đối với các nguyên tắc và chuẩn mực của Khuyến nghị phù hợp với luật pháp quốc tế, bao gồm bộ luật nhân quyền quốc tế;

3. ***Đồng thời khuyến nghị các*** quốc gia thành viên khuyến khích sự tham gia của tất cả các bên liên quan, bao gồm cả các doanh nghiệp kinh doanh, để đảm bảo rằng họ đóng vai trò tương xứng trong việc thực hiện Khuyến nghị này; và tạo được sự quan

tâm đối với Khuyến nghị của các cơ quan quản lý, các cơ quan, tổ chức nghiên cứu và học thuật, các đơn vị và tổ chức trong khu vực công, tư và các tổ chức dân sự được tham gia vào công nghệ AI, để việc phát triển và sử dụng công nghệ AI được dẫn dắt nghiên cứu khoa học lành mạnh cũng như phân tích và đánh giá về đạo đức.

I. PHẠM VI ÁP DỤNG

1. Khuyến nghị này đề cập đến các vấn đề đạo đức liên quan đến lĩnh vực Trí tuệ nhân tạo trong phạm vi quyền hạn của UNESCO. Nó tiếp cận đạo đức AI như một sự phản ánh quy chuẩn có hệ thống, dựa trên một khuôn khổ tổng thể, toàn diện, đa văn hóa và phát triển của các giá trị, nguyên tắc và hành động phụ thuộc lẫn nhau có thể định hướng các xã hội đối phó có trách nhiệm với các tác động đã biết và chưa biết của công nghệ AI đối với con người, xã hội, môi trường và hệ sinh thái, đồng thời cung cấp cho họ cơ sở để chấp thuận hoặc từ chối các công nghệ AI. Khuyến nghị coi đạo đức là một cơ sở tích cực cho việc đánh giá và hướng dẫn quy chuẩn cho các công nghệ AI, đề cập đến phẩm giá con người, sự thịnh vượng và phòng chống tác hại như một la bàn và bắt nguồn từ đạo đức của khoa học và công nghệ.

2. Khuyến nghị này không có tham vọng cung cấp một định nghĩa duy nhất về AI, vì định nghĩa này sẽ cần thay đổi theo thời gian, phù hợp với sự phát triển công nghệ. Thay vào đó, tham vọng của Khuyến nghị là giải quyết những tính năng của các hệ thống AI liên quan chủ yếu đến đạo đức. Do đó, Khuyến nghị tiếp cận các hệ thống AI như các hệ thống có khả năng xử lý dữ liệu và thông tin theo cách giống với hành vi thông minh và thường bao gồm các khía cạnh lý luận, học tập, nhận thức, dự đoán, lập kế hoạch hoặc kiểm soát. Ba yếu tố có một vị trí trung tâm trong cách tiếp cận này:

(a) Hệ thống AI là công nghệ xử lý thông tin có các mô hình và thuật toán tạo ra khả năng học hỏi và thực hiện các nhiệm vụ nhận

thức dẫn đến kết quả như dự đoán và ra quyết định trong môi trường vật chất và ảo. Các hệ thống AI được thiết kế để hoạt động với các mức độ tự chủ khác nhau bằng cách mô hình hóa và đại diện kiến thức đồng thời bằng cách khai thác dữ liệu và tính toán mối tương quan. Các hệ thống AI có thể bao gồm một số phương pháp, chẳng hạn như:

(i) học máy, bao gồm học sâu và học tăng cường;

(ii) lý giải máy, bao gồm lập kế hoạch, lập kế hoạch, đại diện kiến thức và lý luận, tìm kiếm và tối ưu hóa.

Các hệ thống AI có thể được sử dụng trong các hệ thống vật lý mạng, bao gồm Internet vạn vật, hệ thống robot, robot xã hội và giao diện người-máy tính, liên quan đến kiểm soát, nhận thức, xử lý dữ liệu được thu thập bởi các cảm biến và hoạt động của các bộ truyền động trong môi trường nơi các hệ thống AI hoạt động.

(b) Các câu hỏi đạo đức liên quan đến các hệ thống AI gắn với tất cả các giai đoạn của vòng đời hệ thống AI, được hiểu bao gồm từ nghiên cứu, thiết kế và phát triển đến triển khai và sử dụng, bao gồm bảo trì, vận hành, thương mại, tài chính, giám sát và đánh giá, xác nhận, kết thúc sử dụng, tháo rời và chấm dứt. Ngoài ra, các chủ thể AI có thể được định nghĩa là bất kỳ tác nhân nào tham gia vào ít nhất một giai đoạn của vòng đời hệ thống AI và kể cả người tự nhiên và pháp nhân, chẳng hạn như các nhà nghiên cứu, lập trình viên, kỹ sư, nhà khoa học dữ liệu, người dùng cuối, doanh nghiệp kinh doanh, trường đại học và các pháp nhân công/tư, và cả những người khác.

(c) Các hệ thống AI đặt ra các vấn đề đạo đức mới bao gồm tác động của chúng đối với việc ra quyết định, việc làm và lao động, tương tác xã hội, chăm sóc sức khỏe, giáo dục, truyền thông, truy cập thông tin, khoảng cách số, dữ liệu cá nhân và bảo vệ, môi trường, dân chủ, pháp quyền, an ninh và chính sách, sử dụng kép, nhân quyền và các quyền tự do cơ bản, bao gồm tự do ngôn luận,

quyền riêng tư và không phân biệt đối xử. Hơn nữa, những thách thức đạo đức mới được tạo ra bởi các thuật toán AI nhằm tái tạo và củng cố các khuynh hướng hiện có, và do đó làm trầm trọng hơn các hình thức phân biệt đối xử, định kiến và rập khuôn đang tồn tại. Một số vấn đề này có liên quan đến khả năng của hệ thống AI trong thực hiện các nhiệm vụ mà trước đây chỉ có các thực thể sống mới có thể làm, và trong một số trường hợp thậm chí chỉ giới hạn ở con người. Những đặc điểm này mang lại cho các hệ thống AI một vai trò sâu sắc, mới mẻ trong thực tiễn và xã hội của con người, cũng như trong mối quan hệ với môi trường và hệ sinh thái, tạo ra một bối cảnh mới cho trẻ em và thanh thiếu niên lớn lên, phát triển sự hiểu biết về thế giới và bản thân họ, hiểu nghiêm túc về phương tiện truyền thông và thông tin, và học cách đưa ra quyết định. Về lâu dài, các hệ thống AI có thể thách thức ý thức đặc biệt của con người về kinh nghiệm và đại diện, làm gia tăng mối quan tâm về khả năng tự hiểu của con người, tương tác xã hội, văn hóa và môi trường, tính tự chủ, đại diện, giá trị và phẩm giá.

3. Khuyến nghị này đặc biệt chú ý đến ý nghĩa đạo đức rộng lớn hơn của các hệ thống AI liên quan đến các lĩnh vực cốt lõi của UNESCO: giáo dục, khoa học, văn hóa và truyền thông và thông tin, như đã được khám phá trong Nghiên cứu sơ bộ năm 2019 về đạo đức trong trí tuệ nhân tạo của Ủy ban Thế giới về Đạo đức Tri thức khoa học và Công nghệ (COMEST):

(a) Giáo dục, do sống trong các xã hội số hóa đòi hỏi các thực tiễn giáo dục mới, phản ánh đạo đức, tư duy phê phán, hoạt động thiết kế có trách nhiệm và các kỹ năng mới, do những tác động đối với thị trường lao động, việc làm và sự tham gia của người dân.

(b) Khoa học, theo nghĩa rộng nhất và bao gồm tất cả các lĩnh vực học thuật từ khoa học tự nhiên và y học đến khoa học xã hội và nhân văn, vì các công nghệ AI mang lại năng lực và cách tiếp cận

nghiên cứu mới, có ý nghĩa đối với các khái niệm hiểu biết và giải thích khoa học của chúng ta, và tạo ra một cơ sở mới cho việc ra quyết định.

(c) Bản sắc văn hóa và đa dạng văn hoá, vì các công nghệ AI có thể làm phong phú thêm các ngành công nghiệp văn hóa và sáng tạo, nhưng cũng có thể dẫn đến sự tập trung ngày càng tăng nguồn cung nội dung văn hóa, dữ liệu, thị trường và thu nhập vào tay chỉ một vài nhân tố, với một số hàm ý tiêu cực tiềm ẩn đối với tính đa dạng và đa nguyên về ngôn ngữ, phương tiện truyền thông, biểu đạt văn hóa, sự tham gia và sự bình đẳng.

(d) Truyền thông và thông tin, vì công nghệ AI đóng vai trò ngày càng quan trọng trong xử lý, tổ chức và cung cấp thông tin; các vấn đề như báo chí tự động và cung cấp thuật toán về tin tức và kiểm duyệt và quản lý nội dung trên phương tiện truyền thông xã hội cũng như công cụ tìm kiếm chỉ là một vài ví dụ nêu lên các vấn đề liên quan đến tiếp cận thông tin, thông tin giả mạo, thông tin sai lệch, ngôn từ kích động thù địch, sự xuất hiện của các hình thức mới của câu chuyện xã hội, phân biệt đối xử, tự do ngôn luận, quyền riêng tư và phương tiện truyền thông và kiến thức thông tin, v.v.

4. Khuyến nghị này được gửi đến các quốc gia thành viên, cả với tư cách là các chủ thể AI đồng thời với tư cách là người chịu trách nhiệm xây dựng các khung khổ pháp lý và quy tắc trong toàn bộ vòng đời hệ thống AI và nhằm thúc đẩy trách nhiệm doanh nghiệp. Khuyến nghị cũng đưa ra chỉ dẫn đạo đức cho tất cả các chủ thể AI, bao gồm cả khu vực công và tư, bằng cách cung cấp cơ sở cho việc đánh giá tác động đạo đức của hệ thống AI trong cả vòng đời.

II. MỤC ĐÍCH VÀ MỤC TIÊU

5. Khuyến nghị này nhằm mục đích cung cấp cơ sở cho các hệ thống AI hoạt động vì lợi ích của toàn nhân loại, của các cá nhân, xã hội, môi trường và hệ sinh thái, và để ngăn ngừa tác hại. Khuyến

nghị cũng nhằm mục đích khuyến khích việc sử dụng hài hòa các hệ thống AI.

6. Ngoài các khuôn khổ đạo đức hiện có liên quan đến AI trên toàn thế giới, Khuyến nghị nhằm mục đích mang lại một công cụ quy phạm được chấp nhận trên toàn cầu không chỉ tập trung vào việc khớp nối các giá trị và nguyên tắc, mà còn về triển khai thực hiện, thông qua các khuyến nghị chính sách cụ thể, với việc nhấn mạnh vào các vấn đề như bình đẳng giới, bảo vệ môi trường và hệ sinh thái.

7. Do tính phức tạp của các vấn đề đạo đức xung quanh AI đòi hỏi sự hợp tác của nhiều bên liên quan ở các cấp và các lĩnh vực khác nhau của cộng đồng quốc tế, khu vực và quốc gia, Khuyến nghị này nhằm mục đích cho phép các bên liên quan chịu trách nhiệm chung dựa trên một cuộc đối thoại toàn cầu và liên văn hóa.

8. Mục tiêu của khuyến nghị là:

(a) cung cấp một khuôn khổ phổ quát về các giá trị, nguyên tắc và hành động để hướng dẫn các quốc gia xây dựng luật pháp, chính sách hoặc các công cụ khác liên quan đến AI, phù hợp với luật pháp quốc tế;

(b) hướng dẫn hành động của các cá nhân, các nhóm, cộng đồng, tổ chức và các công ty trong khu vực tư nhân để đảm bảo gắn đạo đức với tất cả các giai đoạn của vòng đời hệ thống AI;

(c) bảo vệ, thúc đẩy và tôn trọng các quyền con người và các quyền tự do cơ bản, phẩm giá con người và bình đẳng, bao gồm bình đẳng giới; bảo vệ lợi ích của các thế hệ hiện tại và tương lai; bảo tồn môi trường, đa dạng sinh học và hệ sinh thái; tôn trọng sự đa dạng văn hóa trong tất cả các giai đoạn của vòng đời hệ thống AI;

(d) thúc đẩy đối thoại và xây dựng đồng thuận giữa các bên liên quan, đa ngành và đa nguyên về các vấn đề đạo đức liên quan đến hệ thống AI;

(e) thúc đẩy tiếp cận công bằng với sự phát triển và kiến thức trong lĩnh vực AI và chia sẻ lợi ích, đặc biệt là sự đáp ứng nhu cầu và đóng góp của LMICs, bao gồm LDCs, LLDCs và SIDS.

III. CÁC GIÁ TRỊ VÀ NGUYÊN TẮC

9. Các giá trị và nguyên tắc bao gồm dưới đây cần được tôn trọng bởi tất cả các chủ thể trong vòng đời của hệ thống AI, ngay từ đầu và khi cần thiết cũng như ở mức độ phù hợp, được thúc đẩy thông qua những sửa đổi đối với luật hiện hành và xây dựng luật pháp, quy định và hướng dẫn kinh doanh mới. Điều này phải tuân thủ luật pháp quốc tế, bao gồm Hiến chương Liên Hợp Quốc và các nghĩa vụ về nhân quyền của các quốc gia thành viên, phải phù hợp với các mục tiêu bền vững về xã hội, chính trị, môi trường, giáo dục, khoa học và kinh tế đã được quốc tế thống nhất, chẳng hạn như Các Mục tiêu Phát triển Bền vững của Liên Hợp Quốc (SDGs).

10. Các giá trị đóng một vai trò tích cực như là động lực lý tưởng trong các biện pháp chính sách và chuẩn mực pháp lý. Trong khi các giá trị được nêu dưới đây truyền cảm hứng cho hành vi mong muốn và đại diện cho nền tảng của các nguyên tắc, thì các nguyên tắc giải nén các giá trị nền tảng một cách cụ thể hơn để các giá trị có thể được vận hành nhiều hơn trong các tuyên bố và hành động chính sách.

11. Trong khi tất cả các giá trị và nguyên tắc được nêu dưới đây bản thân là mong muốn, trong bất kỳ bối cảnh thực tế nào đều có thể có sự căng thẳng giữa các giá trị và nguyên tắc này. Trong bất kỳ tình huống nào, một sự đánh giá theo bối cảnh sẽ là cần thiết để quản lý những căng thẳng tiềm ẩn, có tính đến nguyên tắc tỷ lệ và tuân thủ các quyền con người và các quyền tự do cơ bản. Trong mọi trường hợp, bất kỳ hạn chế nào có thể xảy ra về nhân quyền và các quyền tự do cơ bản phải có cơ sở hợp pháp, hợp lý, cần thiết và tương xứng, phù hợp với nghĩa vụ của các quốc gia theo luật pháp

quốc tế. Để điều hướng các kịch bản như vậy một cách thận trọng thường sẽ đòi hỏi sự tham gia của một loạt các bên liên quan thích hợp, sử dụng đối thoại xã hội, cũng như cân nhắc về mặt đạo đức, thẩm định và đánh giá tác động.

12. Độ tin cậy và toàn vẹn trong vòng đời của các hệ thống AI là điều cần thiết để đảm bảo rằng các công nghệ AI sẽ hoạt động vì lợi ích của toàn nhân loại, của các cá nhân, xã hội, môi trường và hệ sinh thái, đồng thời thể hiện các giá trị và nguyên tắc được nêu trong Khuyến nghị này. Mọi người cần có lý do chính đáng để tin tưởng rằng các hệ thống AI có thể mang lại lợi ích cá nhân và chia sẻ, đồng thời áp dụng các biện pháp thích hợp để giảm thiểu rủi ro. Một yêu cầu thiết yếu cho độ tin cậy đó là, trong suốt vòng đời, các hệ thống AI phải chịu sự giám sát rất kỹ và phù hợp của các bên liên quan. Vì độ tin cậy là kết quả của việc tổ chức vận hành các nguyên tắc nêu trong tài liệu này, các hành động chính sách được đề xuất trong Khuyến nghị này đều nhằm mục tiêu tăng cường độ tin cậy trong tất cả các giai đoạn của vòng đời hệ thống AI.

III.1 CÁC GIÁ TRỊ

Tôn trọng, bảo vệ và thúc đẩy nhân quyền, các quyền tự do cơ bản và phẩm giá con người

13. Phẩm giá bất khả xâm phạm và vốn có của mỗi con người tạo thành nền tảng cho hệ thống phổ quát, không thể chia nhỏ, bất khả xâm phạm, phụ thuộc lẫn nhau và liên quan đến các quyền con người và các quyền tự do cơ bản. Do đó, sự tôn trọng, bảo vệ và nêu cao nhân phẩm và quyền con người được thiết lập bởi luật pháp quốc tế, bao gồm luật nhân quyền quốc tế, là điều cần thiết trong suốt vòng đời của các hệ thống AI. Phẩm giá con người liên quan đến việc công nhận giá trị nội tại và bình đẳng của mỗi con người, bất kể chủng tộc, màu da, nguồn gốc, giới tính, tuổi tác, ngôn ngữ, tôn giáo, quan điểm chính trị, nguồn gốc quốc gia, nguồn gốc dân

tộc, nguồn gốc xã hội, điều kiện kinh tế hoặc xã hội của sự sinh thành, hoặc khuyết tật và bất kỳ căn cứ nào khác.

14. Không có con người hay cộng đồng con người nào đáng bị tổn hại hoặc lệ thuộc, cho dù về thể chất, kinh tế, xã hội, chính trị, văn hóa hay tinh thần trong bất kỳ giai đoạn nào trong vòng đời của các hệ thống AI. Trong suốt vòng đời của các hệ thống AI, chất lượng cuộc sống của con người cần được củng cố, mặt khác định nghĩa về "chất lượng cuộc sống" nên được mở cho các cá nhân hoặc nhóm, miễn là không có vi phạm hoặc lạm dụng nhân quyền và các quyền tự do cơ bản, hoặc phẩm giá của con người theo định nghĩa này.

15. Mọi người có thể tương tác với các hệ thống AI trong suốt vòng đời và nhận được sự hỗ trợ từ AI, chẳng hạn như chăm sóc cho những người dễ bị tổn thương hoặc những người trong các tình huống dễ bị tổn thương, bao gồm trẻ em, người già, người khuyết tật hoặc người bệnh. Trong những tương tác như vậy, con người cần không bao giờ bị phản đối, cũng không bị làm mất đi phẩm giá, hay nhân quyền và các quyền tự do cơ bản bị vi phạm hoặc lạm dụng.

16. Nhân quyền và các quyền tự do cơ bản phải được tôn trọng, bảo vệ và thúc đẩy trong suốt vòng đời của các hệ thống AI. Các chính phủ, khu vực tư nhân, tổ chức dân sự, các tổ chức quốc tế, cộng đồng kỹ thuật và các tổ chức khoa học phải tôn trọng các công cụ và khuôn khổ nhân quyền trong những can thiệp của họ vào các quá trình xung quanh vòng đời của các hệ thống AI. Các công nghệ mới cần cung cấp phương tiện mới để ủng hộ, bảo vệ và thực hiện nhân quyền và không vi phạm chúng.

Môi trường và hệ sinh thái phát triển mạnh

17. Môi trường và hệ sinh thái đang phát triển mạnh cần được công nhận, bảo vệ và thúc đẩy trong vòng đời của các hệ thống AI. Hơn nữa, môi trường và hệ sinh thái là nhu cầu hiện hữu để nhân

loại và các sinh vật khác có thể tận hưởng những lợi ích của những tiến bộ trong AI.

18. Tất cả các chủ thể tham gia vào vòng đời của các hệ thống AI phải tuân thủ luật pháp quốc tế hiện hành và luật pháp, tiêu chuẩn và thông lệ trong nước, chẳng hạn như các biện pháp phòng ngừa, được thiết kế để bảo vệ và phục hồi môi trường và hệ sinh thái, và để phát triển bền vững. Họ cần làm giảm tác động môi trường của các hệ thống AI, như ở lượng khí thải carbon, để đảm bảo giảm thiểu tác động biến đổi khí hậu và các yếu tố rủi ro môi trường, và ngăn chặn việc khai thác, sử dụng và chuyển đổi tài nguyên thiên nhiên không bền vững góp phần làm suy thoái môi trường và hệ sinh thái.

Đảm bảo tính đa dạng và bao trùm

19. Tôn trọng, bảo vệ và thúc đẩy sự đa dạng và bao trùm cần được đảm bảo trong suốt vòng đời của các hệ thống AI, phù hợp với luật pháp quốc tế, bao gồm cả luật nhân quyền. Điều này có thể được thực hiện bằng cách thúc đẩy sự tham gia tích cực của tất cả các cá nhân hoặc các nhóm không phân biệt chủng tộc, màu da, nguồn gốc, giới tính, tuổi tác, ngôn ngữ, tôn giáo, quan điểm chính trị, nguồn gốc quốc gia, nguồn gốc dân tộc, nguồn gốc xã hội, tình trạng kinh tế hoặc xã hội của sự sinh thành, hoặc khuyết tật và bất kỳ căn cứ nào khác.

20. Phạm vi lựa chọn lối sống, niềm tin, quan điểm, biểu hiện hoặc trải nghiệm cá nhân, bao gồm việc sử dụng tùy ý các hệ thống AI và đồng thiết kế các kiến trúc này không nên bị hạn chế trong bất kỳ giai đoạn nào trong vòng đời của các hệ thống AI.

21. Hơn nữa, cần nỗ lực, bao gồm cả hợp tác quốc tế, để vượt qua và không bao giờ lợi dụng, sự thiếu cơ sở hạ tầng công nghệ, giáo dục và kỹ năng cần thiết, cũng như các khuôn khổ pháp lý, đặc biệt là tại các nước LMICs, LDCs, LLDCs và SIDS, để ảnh hưởng đến cộng đồng.

Sống trong các xã hội hòa bình, công bằng và kết nối với nhau

22. Các chủ thể AI nên đóng vai trò tham gia và tạo động lực nhằm đảm bảo xã hội hòa bình và công bằng, dựa trên một tương lai kết nối vì lợi ích của tất cả mọi người, phù hợp với nhân quyền và các quyền tự do cơ bản. Giá trị của việc sống trong các xã hội hòa bình và công bằng hướng tới tiềm năng của các hệ thống AI để đóng góp trong suốt vòng đời của chúng cho các môi trường kết nối tất cả các sinh vật sống với nhau và với môi trường tự nhiên.

23. Khái niệm con người được kết nối với nhau dựa trên tri thức mọi người thuộc về một tổng thể lớn hơn, tổng thế đó sẽ phát triển mạnh khi tất cả các bộ phận cấu thành của nó được cho phép phát triển mạnh. Sống trong các xã hội hòa bình, công bằng và liên kết với nhau đòi hỏi một mối liên kết hữu cơ, ngay lập tức, không tính toán, đặc trưng bởi một cuộc tìm kiếm không ngừng cho các mối quan hệ hòa bình, hướng tới việc chăm sóc người khác và môi trường tự nhiên theo nghĩa rộng nhất của khái niệm này.

24. Giá trị này đòi hỏi hòa bình, bao trùm và bình đẳng, công bằng và kết nối cần được thúc đẩy trong suốt vòng đời của các hệ thống AI, trong từng quá trình vòng đời của các hệ thống AI không nên tách biệt, vật hóa hoặc làm mất đi sự tự do và quyền ra quyết định chủ động cũng như sự an toàn của con người và cộng đồng, chia rẽ và khiến các cá nhân và các nhóm chống lại nhau, hoặc đe dọa sự chung sống giữa con người, các sinh vật khác và môi trường tự nhiên.

III.2 NGUYÊN TẮC

Cân xứng và không gây hại

25. Cần thừa nhận rằng các công nghệ AI không nhất thiết phải đảm bảo sự phát triển của con người cũng như môi trường và

hệ sinh thái. Hơn nữa, không có quy trình nào liên quan đến cả vòng đời của hệ thống AI sẽ vượt quá những gì cần thiết để đạt được các mục đích hoặc mục tiêu hợp pháp và cần phải phù hợp với bối cảnh. Trong trường hợp có thể xảy ra bất kỳ tổn hại nào đối với con người, nhân quyền và các quyền tự do cơ bản, cộng đồng và xã hội nói chung hoặc môi trường và hệ sinh thái, việc thực hiện các thủ tục đánh giá rủi ro và áp dụng các biện pháp để ngăn chặn sự xuất hiện của tác hại đó phải được đảm bảo.

26. Việc lựa chọn sử dụng hệ thống AI hay lựa chọn phương pháp AI nào để sử dụng cần được cân nhắc theo những cách sau: (a) phương pháp AI được chọn phải phù hợp và cân xứng để đạt được một mục đích hợp pháp nhất định; (b) phương pháp AI được chọn không được vi phạm các giá trị cơ bản được nêu trong tài liệu này, đặc biệt, việc sử dụng nó không được xâm phạm hoặc lạm dụng nhân quyền; và (c) phương pháp AI phải phù hợp với bối cảnh và phải dựa trên các cơ sở khoa học nghiêm ngặt. Trong các tình huống mà các quyết định được hiểu là có tác động không thể đảo ngược hoặc khó đảo ngược hoặc có thể liên quan đến quyết định sống còn, cần áp dụng quyết định cuối cùng của con người. Đặc biệt, các hệ thống AI không nên được sử dụng các cho mục đích chấm điểm xã hội hoặc giám sát hàng loạt.

An toàn và an ninh

27. Những tác hại không mong muốn (rủi ro an toàn), cũng như khả năng dễ bị tấn công (rủi ro bảo mật) cần phải tránh và cần được giải quyết, ngăn chặn và loại bỏ trong suốt vòng đời của các hệ thống AI nhằm đảm bảo an toàn và bảo mật của con người, môi trường và hệ sinh thái. AI an toàn và bảo mật sẽ được kích hoạt bởi sự phát triển của các khung khổ truy cập dữ liệu có thể bảo vệ quyền riêng tư, có thể thúc đẩy đào tạo tốt hơn và xác nhận các mô hình AI sử dụng dữ liệu chất lượng.

Công bằng và không phân biệt đối xử

28. Các chủ thể AI cần đẩy mạnh công bằng xã hội và bảo vệ sự công bằng và không phân biệt đối xử bằng bất cứ hình thức nào đồng thời luôn thủ luật pháp quốc tế. Điều này ngụ ý một cách tiếp cận toàn diện để đảm bảo rằng lợi ích của công nghệ AI có sẵn và có thể tiếp cận được bởi tất cả mọi người, có tính đến nhu cầu cụ thể của các nhóm tuổi khác nhau, hệ thống văn hóa, các nhóm ngôn ngữ khác nhau, người khuyết tật, trẻ em gái và phụ nữ, những người bên lề hoặc những người dễ bị tổn thương, hoặc những người trong các hoàn cảnh dễ bị tổn thương. Các quốc gia thành viên cần hành động để thúc đẩy quyền truy cập toàn diện cho tất cả mọi người, bao gồm cả cộng đồng địa phương, vào các hệ thống AI với nội dung và dịch vụ có liên quan đến địa phương, và với sự tôn trọng đối với đa ngôn ngữ và đa dạng văn hóa. Các quốc gia thành viên cần hành động để giải quyết sự khác biệt về kỹ thuật số và đảm bảo quyền truy cập toàn diện và quyền tham gia vào phát triển AI. Tại cấp độ quốc gia, các quốc gia thành viên cần thúc đẩy sự bình đẳng giữa nông thôn và thành thị, và giữa tất cả mọi người không kể chủng tộc, màu da, nguồn gốc, giới tính, tuổi tác, ngôn ngữ, tôn giáo, quan điểm chính trị, nguồn gốc quốc gia, nguồn gốc dân tộc, nguồn gốc xã hội, hoàn cảnh kinh tế hoặc xã hội của sự sinh thành, hoặc khuyết tật và bất kỳ căn cứ nào khác, về khả năng tiếp cận và tham gia vào vòng đời của hệ thống AI. Ở cấp độ quốc tế, các quốc gia có công nghệ tiên tiến nhất có trách nhiệm đoàn kết với những quốc gia kém tiên tiến nhất để đảm bảo rằng lợi ích của công nghệ AI được chia sẻ để việc truy cập và tham gia vào vòng đời của hệ thống AI của các quốc gia kém phát triển này góp phần vào một trật tự thế giới công bằng hơn về thông tin, truyền thông, văn hóa, giáo dục, nghiên cứu và ổn định kinh tế-xã hội và chính trị.

29. Các chủ thể AI cần thực hiện tất cả các nỗ lực hợp lý để giảm thiểu và tránh tăng cường hoặc duy trì các ứng dụng và kết cục

phân biệt đối xử hoặc thiên vị trong suốt vòng đời của hệ thống AI để đảm bảo sự công bằng của các hệ thống đó. Cần có sẵn biện pháp khắc phục hiệu quả để chống lại sự phân biệt đối xử và xác định thuật toán thiên vị.

30. Hơn nữa, sự khác biệt về kỹ thuật số và tri thức trong và giữa các quốc gia cần được giải quyết trong suốt vòng đời của hệ thống AI, bao gồm cả khả năng tiếp cận và chất lượng tieps cận công nghệ và dữ liệu, phù hợp với các khung pháp lý quốc gia, khu vực và quốc tế có liên quan, cũng đối với việc kết nối, kiến thức, kỹ năng và sự tham gia có ý nghĩa của các cộng đồng bị ảnh hưởng, để mọi người được đối xử công bằng.

Tính bền vững

31. Sự phát triển của các xã hội bền vững dựa trên việc đạt được một tập hợp phức tạp các mục tiêu về sự liên tục của các chiều cạnh con người, xã hội, văn hóa, kinh tế và môi trường. Sự ra đời của các công nghệ AI có thể mang lại lợi ích cho các mục tiêu bền vững hoặc cản trở việc thực hiện các mục tiêu, tùy thuộc vào cách chúng được áp dụng giữa các quốc gia có mức độ phát triển khác nhau. Do đó, việc đánh giá liên tục tác động đối với con người, xã hội, văn hóa, kinh tế và môi trường của các công nghệ AI cần được thực hiện với nhận thức đầy đủ về ý nghĩa của các công nghệ AI đối với sự bền vững như một tập hợp các mục tiêu liên tục phát triển trên một loạt các khía cạnh như hiện được xác định trong Các Mục tiêu Phát triển Bền vững (SDGs) của Liên Hợp Quốc.

Quyền riêng tư và Bảo vệ Dữ liệu

32. Quyền riêng tư, một quyền thiết yếu để bảo vệ phẩm giá con người, quyền tự chủ của con người và đại diện con người, phải được tôn trọng, bảo vệ và thúc đẩy trong suốt vòng đời của các hệ thống AI. Điều quan trọng là dữ liệu cho các hệ thống AI phải được thu thập, sử dụng, chia sẻ, lưu trữ và xóa theo những cách phù hợp

với luật pháp quốc tế và phù hợp với các giá trị và nguyên tắc được nêu trong Khuyến nghị này, đồng thời tôn trọng các khung pháp lý quốc gia, khu vực và quốc tế có liên quan.

33. Các khung khổ bảo vệ dữ liệu và các cơ chế quản trị phù hợp nên được thiết lập theo cách tiếp cận với nhiều bên liên quan ở cấp quốc gia hoặc quốc tế, được bảo vệ bởi các hệ thống tư pháp và được đảm bảo trong suốt vòng đời của các hệ thống AI. Các khung khổ bảo vệ dữ liệu và bất kỳ cơ chế liên quan nào cần tham khảo từ các nguyên tắc và tiêu chuẩn bảo vệ dữ liệu quốc tế liên quan đến việc thu thập, sử dụng và tiết lộ dữ liệu cá nhân và thực hiện các quyền của họ bởi các đối tượng dữ liệu trong khi vẫn đảm bảo mục đích hợp pháp và cơ sở pháp lý hợp lệ cho việc xử lý dữ liệu cá nhân, bao gồm cả sự đồng ý với đầy đủ thông tin.

34. Các hệ thống thuật toán yêu cầu đánh giá tác động quyền riêng tư đầy đủ, bao gồm các yếu tố xã hội và đạo đức về việc sử dụng chúng và sử dụng sáng tạo quyền riêng tư theo cách tiếp cận thiết kế. Các chủ thể AI cần đảm bảo rằng họ chịu trách nhiệm về việc thiết kế và triển khai các hệ thống AI theo cách nhằm đảm bảo thông tin cá nhân được bảo vệ trong suốt vòng đời của hệ thống AI.

Sự giám sát và quyết tâm của con người

35. Các quốc gia thành viên cần đảm bảo luôn có thể quy trách nhiệm đạo đức và pháp lý cho bất kỳ giai đoạn nào trong vòng đời của các hệ thống AI, cũng như trong các trường hợp khắc phục liên quan đến hệ thống AI, cho con người vật lý hoặc các pháp nhân hiện có. Do đó, sự giám sát của con người không chỉ đề cập đến sự giám sát cá nhân của con người, mà còn đề cập đến sự giám sát công khai toàn diện một cách thích hợp.

36. Có thể đôi khi con người sẽ chọn dựa vào hệ thống AI vì lý do hiệu quả, nhưng quyết định nhượng lại quyền kiểm soát trong bối cảnh hạn chế vẫn là của con người, vì con người có thể sử dụng các

hệ thống AI trong việc ra quyết định và hành động, nhưng một hệ thống AI không bao giờ có thể thay thế việc chịu trách nhiệm và trách nhiệm giải trình cuối cùng của con người. Về nguyên tắc, quyết định sống và chết không nên được nhượng lại cho các hệ thống AI.

Tính minh bạch và khả năng giải trình

37. Tính minh bạch và khả năng giải trình của các hệ thống AI thường là điều kiện tiên quyết cần thiết để đảm bảo sự tôn trọng, bảo vệ và thúc đẩy nhân quyền, tự do tinh thần và các nguyên tắc đạo đức. Minh bạch là điều cần thiết để các chế độ trách nhiệm quốc gia và quốc tế có liên quan có thể hoạt động hiệu quả. Sự thiếu minh bạch cũng có thể làm suy yếu khả năng các quyết định thách thức hiệu quả dựa trên kết quả được tạo ra bởi các hệ thống AI và do đó có thể vi phạm quyền được thử nghiệm công bằng và khắc phục hiệu quả, và làm hạn chế các lĩnh vực mà các hệ thống này có thể được sử dụng hợp pháp.

38. Trong khi những nỗ lực cần phải được thực hiện để tăng tính minh bạch và khả năng giải trình của các hệ thống AI, bao gồm những hệ thống có tác động ngoài lãnh thổ, trong suốt vòng đời của chúng nhằm hỗ trợ quản trị dân chủ, mức độ minh bạch và giải trình phải luôn phù hợp với bối cảnh và tác động, vì có thể cần phải cân bằng giữa tính minh bạch và khả năng giải thích cũng như các nguyên tắc khác như quyền riêng tư, an toàn và bảo mật. Mọi người cần được cung cấp thông tin đầy đủ khi một quyết định được thông tin hoặc được đưa ra trên cơ sở các thuật toán AI, bao gồm cả khi nó ảnh hưởng đến sự an toàn hoặc nhân quyền, và trong các trường hợp đó nên có cơ hội yêu cầu thông tin giải trình từ các chủ thể AI có liên quan hoặc các thể chế của khu vực công. Ngoài ra, các cá nhân sẽ có thể truy cập vào các lý do cho một quyết định ảnh hưởng đến quyền và tự do của họ, và có lựa chọn đệ trình cho một thành viên được chỉ định của công ty tư nhân hoặc thể chế công có thể xem xét

và sửa chữa quyết định. Các chủ thể AI cần thông tin cho người dùng khi một sản phẩm hoặc dịch vụ được cung cấp trực tiếp hoặc với sự hỗ trợ của các hệ thống AI một cách thích hợp và kịp thời.

39. Từ lăng kính kỹ thuật xã hội, sự minh bạch hơn góp phần cho các xã hội hòa bình, công bằng, dân chủ và toàn diện hơn. Nó cho phép công chúng xem xét kỹ có thể làm giảm tham nhũng và phân biệt đối xử, và cũng có thể giúp phát hiện và ngăn chặn các tác động tiêu cực đến nhân quyền. Sự minh bạch nhằm mục đích cung cấp thông tin thích hợp cho các địa chỉ tương ứng để cho phép sự hiểu biết và tăng cường niềm tin của họ. Cụ thể đối với hệ thống AI, tính minh bạch có thể cho phép mọi người hiểu cách mỗi giai đoạn của hệ thống AI được tạo ra, phù hợp với bối cảnh và độ nhạy của hệ thống AI. Nó cũng có thể bao gồm cái nhìn sâu sắc về các yếu tố ảnh hưởng đến một dự đoán hoặc quyết định cụ thể và liệu các cam đoan phù hợp (như các biện pháp an toàn hoặc công bằng) có được đưa ra hay không. Trong trường hợp các mối đe dọa nghiêm trọng về tác động bất lợi của nhân quyền, tính minh bạch cũng có thể yêu cầu chia sẻ mã hoặc bộ dữ liệu.

40. Khả năng giải trình đề cập đến việc làm cho dễ hiểu và cung cấp cái nhìn sâu sắc về kết quả của các hệ thống AI. Khả năng giải trình của các hệ thống AI cũng đề cập đến tính hiểu biết của đầu vào, đầu ra và hoạt động của từng khối xây dựng thuật toán và cách nó đóng góp vào kết quả của các hệ thống. Do đó, khả năng giải trình được chậm lại chặt chẽ để minh bạch, vì kết quả và quy trình phụ dẫn đến kết quả cần nhằm mục đích dễ hiểu và truy xuất nguồn gốc, phù hợp với bối cảnh. Các chủ thể AI nên cam kết đảm bảo rằng các thuật toán được phát triển có thể giải thích được. Trong trường hợp các ứng dụng AI ảnh hưởng đến người dùng cuối theo cách không phải là tạm thời, không dễ dàng đảo ngược hoặc có nguy cơ thấp, cần đảm bảo rằng lời giải trình có ý nghĩa được cung cấp với bất kỳ quyết định nào dẫn đến hành động được thực hiện để kết quả được minh bạch.

41. Tính minh bạch và khả năng giải trình liên quan chặt chẽ đến các biện pháp trách nhiệm và trách nhiệm giải trình đầy đủ, cũng như độ tin cậy của các hệ thống AI.

Trách nhiệm và trách nhiệm giải trình

42. Các chủ thể AI và các quốc gia thành cần nên tôn trọng, bảo vệ và đẩy mạnh nhân quyền cũng như các quyền tự do cơ bản, đồng thời cũng cần thúc đẩy việc bảo vệ môi trường và hệ sinh thái, chịu trách nhiệm đạo đức và pháp lý tương ứng của họ, phù hợp với luật pháp quốc gia và quốc tế, đặc biệt là các nghĩa vụ về nhân quyền của các quốc gia thành viên và các chỉ dẫn về đạo đức trong suốt vòng đời của các hệ thống AI, bao gồm các chủ thể AI trong lãnh thổ và quyền kiểm soát của họ. Trách nhiệm đạo đức và nghĩa vụ đối với các quyết định và hành động dựa vào hệ thống AI theo bất kỳ cách nào cuối cùng cũng luôn thuộc về các chủ thể AI tương ứng với vai trò của họ trong vòng đời của hệ thống AI.

43. Các cơ chế giám sát, đánh giá tác động, kiểm toán và thẩm định thích hợp, bao gồm cả sự bảo vệ của người tố cáo (whislteblowers), nên được xây dựng để đảm bảo trách nhiệm giải trình cho các hệ thống AI và tác động của chúng trong suốt vòng đời. Cả thiết kế kỹ thuật và thể chế phải đảm bảo khả năng kiểm toán và truy xuất nguồn gốc của (hoạt động của) các hệ thống AI đặc biệt để giải quyết xung đột với các chuẩn mực và tiêu chuẩn về nhân quyền và các mối đe dọa đối với môi trường và hệ sinh thái.

Nhận thức và kiến thức

44. Nhận thức và hiểu biết của công chúng về công nghệ AI và giá trị của dữ liệu cần được thúc đẩy thông qua giáo dục mở và có thể tiếp cận, sự tham gia của người dân, các kỹ năng số và đào tạo về đạo đức AI, truyền thông cũng như đào tạo thông tin và đào tạo do các chính phủ, các tổ chức liên chính phủ, tổ chức dân sự, tổ chức khoa học, phương tiện truyền thông, các nhà lãnh đạo cộng

đồng và khu vực tư nhân, đồng thời chú ý đến sự đa dạng ngôn ngữ, xã hội và văn hóa hiện có, nhằm đảm bảo sự tham gia hiệu quả của công chúng để tất cả các thành viên trong xã hội có thể đưa ra quyết định sáng suốt về việc sử dụng các hệ thống AI và được bảo vệ khỏi ảnh hưởng thái quá.

45. Tìm hiểu về tác động của các hệ thống AI cần bao gồm việc tìm hiểu, thông qua và đối với các quyền con người và các quyền tự do cơ bản, có nghĩa là cách tiếp cận và hiểu biết về các hệ thống AI nên dựa trên tác động của chúng đối với nhân quyền và khả năng tiếp cận các quyền đó, cũng như đối với môi trường và hệ sinh thái.

Quản trị và hợp tác thích ứng và với các bên liên quan

46. Luật pháp quốc tế và chủ quyền quốc gia phải được tôn trọng trong việc sử dụng dữ liệu. Điều đó có nghĩa là các quốc gia, bằng việc tuân thủ luật pháp quốc tế, có thể điều chỉnh dữ liệu được tạo ra hoặc đi qua lãnh thổ của họ và thực hiện các biện pháp hướng tới quản lý dữ liệu hiệu quả, bao gồm việc bảo vệ dữ liệu, dựa trên sự tôn trọng quyền riêng tư phù hợp với luật pháp quốc tế và các chuẩn mực và tiêu chuẩn nhân quyền khác.

47. Sự tham gia của các bên liên quan khác nhau trong suốt vòng đời của hệ thống AI là cần thiết cho các phương pháp tiếp cận toàn diện về quản trị AI, cho phép tất cả mọi người chia sẻ lợi ích và đóng góp cho sự phát triển bền vững. Các bên liên quan bao gồm các chính phủ, các tổ chức liên chính phủ, cộng đồng kỹ thuật, các tổ chức dân sự, các nhà nghiên cứu và các tổ chức khoa học, truyền thông, giáo dục, các nhà hoạch định chính sách, các công ty tư nhân, các thể chế về nhân quyền và các cơ quan về quyền bình đẳng, các cơ quan giám sát chống lại sự phân biệt đối xử, và các tổ chức dành cho thanh thiếu niên và trẻ em. Cần áp dụng các tiêu chuẩn mở và khả năng tương tác để tạo điều kiện hợp tác. Cần áp dụng các biện pháp tính đến sự thay đổi trong công nghệ, sự xuất hiện của các

nhóm lợi ích liên quan mới và cho phép sự tham gia có ý nghĩa của các nhóm, các cộng đồng và cá nhân bị thiệt thòi và, khi có liên quan, trong trường hợp của người bản địa, đồng thời tôn trọng sự tự quản lý dữ liệu của họ.

III. CÁC LĨNH VỰC HÀNH ĐỘNG CHÍNH SÁCH

48. Các hành động chính sách mô tả in các lĩnh vực chính sách dưới đây vận hành các giá trị và nguyên tắc được nêu trong Khuyến nghị này. Hành động chủ yếu là các quốc gia thành viên đưa ra các biện pháp hiệu quả, ví dụ như các khuôn khổ hoặc cơ chế chính sách, và để đảm bảo rằng các bên liên quan khác, chẳng hạn như các công ty tư nhân, các tổ chức học thuật và nghiên cứu, và các tổ chức dân sự cùng với các hành động khác tuân thủ bằng cách khuyến khích tất cả các bên liên quan đẩy mạnh nhân quyền, pháp quyền, dân chủ, và đánh giá tác động đạo đức và các công cụ thẩm định phù hợp với hướng dẫn bao gồm các Nguyên tắc hướng dẫn của Liên Hợp Quốc về Kinh doanh và Nhân quyền. Quá trình xây dựng các chính sách hoặc cơ chế như vậy phải bao gồm tất cả các bên liên quan và cần tính đến các điều kiện và ưu tiên của mỗi quốc gia thành viên. UNESCO có thể là đối tác và hỗ trợ các quốc gia thành viên trong việc xây dựng cũng như giám sát và đánh giá các cơ chế chính sách.

49. UNESCO công nhận rằng để triển khai thực hiện Khuyến nghị này, các quốc gia thành viên ở các mức độ sẵn sàng khác nhau về mặt khoa học, công nghệ, kinh tế, giáo dục, pháp lý, quy định, cơ sở hạ tầng, xã hội, văn hóa và các khía cạnh khác. Cần lưu ý rằng "sự sẵn sàng" ở đây là một trạng thái động. Do đó, để có thể đẩy mạnh hiệu quả Khuyến nghị này, UNESCO sẽ: (1) xây dựng phương pháp đánh giá mức độ sẵn sàng để hỗ trợ các quốc gia thành viên mong muốn xác định tình trạng của họ tại các thời điểm cụ thể của

quỹ đạo sẵn sàng theo một chiều liên tục; và (2) đảm bảo hỗ trợ cho các quốc gia thành viên quan tâm về việc xây dựng phương pháp đánh giá tác động đạo đức của UNESCO (EIA) về công nghệ AI, chia sẻ các kinh nghiệm thực tiễn tốt nhất, hướng dẫn đánh giá và các cơ chế cũng như hoạt động phân tích khác.

LĨNH VỰC CHÍNH SÁCH 1: ĐÁNH GIÁ TÁC ĐỘNG ĐẠO ĐỨC

50. Các quốc gia thành viên cần đưa ra các khuôn khổ đánh giá tác động, chẳng hạn như đánh giá tác động đạo đức, để xác định và đánh giá lợi ích, mối quan tâm và rủi ro của hệ thống AI, cũng như các biện pháp phòng ngừa, giảm thiểu và giám sát rủi ro thích hợp, trong số các cơ chế đảm bảo khác. Những đánh giá tác động như vậy cần xác định các tác động đối với quyền con người và các quyền tự do cơ bản, đặc biệt là quyền của những người bị thiệt thòi và dễ bị tổn thương hoặc những người trong các điều kiện dễ bị tổn thương, quyền lao động, môi trường và hệ sinh thái cũng như ý nghĩa đạo đức và xã hội, đồng thời tạo điều kiện cho sự tham gia của công dân phù hợp với các giá trị và nguyên tắc được nêu trong Khuyến nghị này.

51. Các quốc gia thành viên và các công ty tư nhân cần xây dựng các cơ chế thẩm tra và giám sát để xác định, ngăn chặn, giảm thiểu và tính đến cách họ giải quyết tác động của các hệ thống AI đối với sự tôn trọng nhân quyền, pháp quyền và xã hội bao trùm. Các quốc gia thành viên cũng cần có khả năng đánh giá tác động kinh tế xã hội của các hệ thống AI về nghèo đói và đảm bảo rằng khoảng cách giữa người giàu và người nghèo, cũng như sự khác biệt về kỹ thuật số trong nước và giữa các nước, sẽ không tăng lên với việc áp dụng rộng rãi các công nghệ AI hiện tại và trong tương lai. Đặc biệt, để làm được điều này cần triển khai các tiêu chuẩn minh bạch có thể thực thi, tương ứng với việc truy cập thông tin, bao gồm thông tin về

lợi ích công cộng do các pháp nhân tư nhân nắm giữ. Các quốc gia thành viên, các công ty tư nhân và các tổ chức dân sự cần điều tra tác động xã hội học và tâm lý học của các khuyến nghị dựa trên AI đối với con người trong quyền tự chủ ra quyết định. Các hệ thống AI được xác định là có những rủi ro tiềm ẩn đối với nhân quyền nên được thử nghiệm rộng rãi bởi các chủ thể AI, bao gồm cả trong điều kiện thực tế nếu cần thiết, như một phần của việc Đánh giá Tác động Đạo đức, trước khi đưa chúng ra thị trường.

52. Các quốc gia thành viên và doanh nghiệp kinh doanh cần thực hiện các biện pháp thích hợp để theo dõi tất cả các giai đoạn trong vòng đời của hệ thống AI, bao gồm chức năng của các thuật toán được sử dụng để ra quyết định,dữ iệu, cũng như các chủ thể AI tham gia vào quá trình này, đặc biệt là trong các dịch vụ công và những nơi cần tương tác trực tiếp với người dùng cuối, như một phần của đánh giá tác động đạo đức. Nghĩa vụ của các quốc gia thành viên theo luật nhân quyền cần phải là một phần của các khía cạnh đạo đức trong đánh giá hệ thống AI.

53. Các chính phủ cần áp dụng một khung pháp lý để đề ra thủ tục, đặc biệt là đối với các cơ quan công quyền, nhằm thực hiện đánh giá tác động đạo đức trên các hệ thống AI để dự đoán hậu quả, giảm thiểu rủi ro, tránh hậu quả có hại, tạo điều kiện cho sự tham gia của công dân và giải quyết các thách thức xã hội. Việc đánh giá cũng cần thiết lập các cơ chế giám sát thích hợp, bao gồm khả năng kiểm toán, truy xuất nguồn gốc và khả năng giải trình, cho phép đánh giá các thuật toán, dữ liệu và quy trình thiết kế, cũng như bao gồm đánh giá bên ngoài các hệ thống AI. Đánh giá tác động đạo đức phải minh bạch và mở một cách thích hợp cho công chúng. Những đánh giá như vậy cũng nên được thực hiện đa ngành, đa bên liên quan, đa văn hóa, đa nguyên và bao trùm. Các cơ quan công quyền cần được yêu cầu giám sát các hệ thống AI được thực hiện và / hoặc triển khai bởi các cơ quan đó bằng cách giới thiệu các cơ chế và công cụ thích hợp.

LĨNH VỰC CHÍNH SÁCH 2: QUẢN TRỊ VÀ QUẢN LÝ ĐẠO ĐỨC

54. Các quốc gia thành viên cần đảm bảo rằng các cơ chế quản trị AI bao gồm, minh bạch, đa ngành, đa phương (bao gồm khả năng giảm thiểu và khắc phục thiệt hại qua biên giới) và nhiều bên liên quan. Đặc biệt, quản trị cần bao gồm các khía cạnh của dự đoán và bảo vệ hiệu quả, giám sát tác động, thực thi và khắc phục.

55. Các quốc gia thành viên cần đảm bảo rằng những tác hại gây ra thông qua các hệ thống AI phải được điều tra và khắc phục, bằng cách ban hành các cơ chế thực thi mạnh mẽ và các hành động khắc phục hậu quả, để đảm bảo rằng nhân quyền và các quyền tự do cơ bản và pháp quyền được tôn trọng trong thế giới kỹ thuật số và trong thế giới vật chất. Các cơ chế và hành động như vậy cần bao gồm các cơ chế khắc phục được cung cấp bởi các doanh nghiệp công và tư. Khả năng kiểm toán và truy xuất nguồn gốc của các hệ thống AI cần được thúc đẩy hướng đến mục đích này. Ngoài ra, các quốc gia thành viên cần tăng cường năng lực thể chế để thực hiện cam kết này và cần hợp tác với các nhà nghiên cứu và các bên liên quan khác để điều tra, ngăn chặn và giảm thiểu bất kỳ việc sử dụng có hại tiềm ẩn nào của các hệ thống AI.

56. Các quốc gia thành viên được khuyến khích phát triển các chiến lược AI quốc gia và khu vực đồng thời xem xét các hình thức quản trị mềm như cơ chế chứng nhận cho các hệ thống AI và sự công nhận lẫn nhau về chứng nhận của quốc gia mình, theo độ nhạy cảm của lĩnh vực ứng dụng và tác động dự kiến đối với nhân quyền, môi trường và hệ sinh thái, cũng như những cân nhắc đạo đức khác được nêu trong Khuyến nghị này. Một cơ chế như vậy có thể bao gồm các mức độ kiểm toán khác nhau về hệ thống, dữ liệu và tuân thủ các hướng dẫn đạo đức cũng như các yêu cầu thủ tục theo quan điểm về các khía cạnh đạo đức. Đồng thời, một cơ chế như vậy

không nên cản trở sự đổi mới hoặc gây bất lợi cho các doanh nghiệp vừa và nhỏ hay các doanh nghiệp khởi nghiệp, các tổ chức dân sự cũng như các tổ chức nghiên cứu và khoa học, do gánh nặng hành chính thái quá. Các cơ chế này cũng nên bao gồm cấu phần giám sát thường xuyên để đảm bảo tính toàn vẹn của hệ thống và tuân thủ các hướng dẫn đạo đức trong toàn bộ vòng đời của hệ thống AI, nếu cần thì yêu cầu có chứng nhận.

57. Các quốc gia thành viên và các cơ quan công quyền cần thực hiện tự đánh giá tính minh bạch về các hệ thống AI hiện có và đặc biệt được đề xuất cần bao gồm đánh giá việc áp dụng AI có phù hợp hay không và, nếu có, cần đánh giá thêm để xác định phương pháp thích hợp là gì, cũng như đánh giá liệu việc áp dụng đó có dẫn đến vi phạm hoặc lạm dụng nghĩa vụ của các quốc gia thành viên theo luật nhân quyền hay không, và nếu đúng như vậy thì cần cấm sử dụng.

58. Các quốc gia thành viên cần khuyến khích các pháp nhân của khu vực công, các công ty tư nhân và các tổ chức dân khuyến khích sự tham gia của các bên liên quan khác nhau vào quản trị AI và xem xét bổ sung vai trò của một nhân viên độc lập Đạo đức AI hoặc một số cơ chế khác để giám sát đánh giá tác động đạo đức, kiểm toán và các nỗ lực giám sát liên tục đồng thời đảm bảo những hướng dẫn đạo đức của các hệ thống AI. Các quốc gia thành viên, các công ty tư nhân và các tổ chức dân sự, với sự hỗ trợ của UNESCO, được khuyến khích tạo ra một mạng lưới các nhân viên độc lập về Đạo đức AI để hỗ trợ trong quá trình này ở cấp quốc gia, khu vực và quốc tế.

59. Các quốc gia thành viên cần tăng cường sự phát triển và khả năng tiếp cận, một hệ sinh thái số để phát triển đạo đức và toàn diện của các hệ thống AI ở cấp quốc gia, bao gồm giải quyết những lỗ hổng trong việc truy cập vào vòng đời hệ thống AI, đồng thời đóng

góp vào hợp tác quốc tế. Một hệ sinh thái như vậy bao gồm, đặc biệt là các công nghệ kỹ thuật số và cơ sở hạ tầng, các cơ chế để chia sẻ một cách thích hợp kiến thức AI.

60. Các quốc gia thành viên cần thiết lập các cơ chế, phối hợp với các tổ chức quốc tế, các tập đoàn xuyên quốc gia, các tổ chức khoa học và tổ chức dân sự, để đảm bảo sự tham gia tích cực của tất cả các quốc gia thành viên, đặc biệt là các LMICs, đặc biệt là LDCs, LLDC và SIDS, trong các trao đổi quốc tế liên quan đến quản trị AI. Điều này có thể thông qua việc cung cấp các quỹ, sự tham gia bình đẳng của khu vực hoặc bất kỳ cơ chế nào khác. Hơn nữa, để đảm bảo sự bao trùm của các diễn đàn về AI, các quốc gia thành viên cần tạo điều kiện cho việc đi lại của các chủ thể AI trong và ngoài lãnh thổ của họ, đặc biệt là từ các LMICs, đặc biệt là LDCs, LLDC và SIDS, với mục đích tham gia vào các hoạt động này.

61. Sửa đổi luật hiện hành hoặc xây dựng luật pháp quốc gia mới quản lý các hệ thống AI phải tuân thủ các nghĩa vụ của các quốc gia thành viên theo luật nhân quyền và thúc đẩy nhân quyền và các quyền tự do cơ bản rong suốt vòng đời của các hệ thống AI. Việc quảng bá cũng nên dưới hình thức sáng kiến quản trị, các ví dụ điển hình về các thực tiễn hợp tác liên quan đến hệ thống AI và các hướng dẫn về kỹ thuật cũng như phương pháp luận của quốc gia và quốc tế khi các công nghệ AI ngày càng phát triển. Các khu vực khác nhau, bao gồm cả khu vực tư nhân, trong thực tế hoạt động của họ liên quan đến các hệ thống AI phải tôn trọng, bảo vệ và thúc đẩy nhân quyền và các quyền tự do cơ bản bằng cách sử dụng các công cụ hiện có và các công cụ mới kết hợp với Khuyến nghị này.

62. Các quốc gia thành viên sử dụng các hệ thống Al cho các trường hợp nhạy cảm về nhân quyền, chẳng hạn như thực thi pháp luật, phúc lợi, việc làm, phương tiện truyền thông và nhà cung cấp thông tin, chăm sóc sức khỏe và hệ thống tư pháp độc lập cần có

các cơ chế để giám sát tác động kinh tế và xã hội của các hệ thống đó bởi các cơ quan giám sát thích hợp, bao gồm các cơ quan bảo vệ dữ liệu độc lập, giám sát ngành và các cơ quan công quyền chịu trách nhiệm giám sát.

63. Các quốc gia thành viên cần nâng cao năng lực của ngành tư pháp để đưa ra các quyết định liên quan đến hệ thống AI theo quy định của pháp luật đồng thời phù hợp với luật pháp và tiêu chuẩn quốc tế, bao gồm cả việc sử dụng các hệ thống AI trong các cuộc thảo luận của họ, đồng thời đảm bảo rằng các nguyên tắc giám sát của con người được duy trì. Trong trường hợp các hệ thống AI được sử dụng bởi tư pháp điện tử, cần có đủ các biện pháp bảo vệ để đảm bảo bảo vệ được các quyền cơ bản của con người, pháp quyền, độc lập tư pháp cũng như nguyên tắc giám sát của con người đồng thời đảm bảo sự phát triển và sử dụng các hệ thống AI đáng tin cậy, định hướng lợi ích chung và lấy con người làm trung tâm trong tư pháp.

64. Các quốc gia thành viên cần đảm bảo rằng các chính phủ và các tổ chức đa phương đóng vai trò hàng đầu trong việc đảm bảo an toàn và bảo mật của các hệ thống AI, với sự tham gia của nhiều bên liên quan. Cụ thể, các quốc gia thành viên, các tổ chức quốc tế và các cơ quan liên quan khác cần xây dựng các tiêu chuẩn quốc tế mô tả mức độ an toàn và minh bạch có thể đo lường, có thể kiểm tra được để các hệ thống có thể được đánh giá khách quan và xác định được các trường hợp tuân thủ. Hơn nữa, các quốc gia thành viên và doanh nghiệp kinh doanh cần liên tục hỗ trợ nghiên cứu chiến lược về rủi ro an toàn và bảo mật tiềm ẩn của công nghệ AI đồng thời cần khuyến khích nghiên cứu về tính minh bạch và khả năng giải trình, tính bao trùm và tri thức bằng cách tài trợ thêm cho các lĩnh vực đó với các ngành và cấp độ khác nhau, chẳng hạn như ngôn ngữ kỹ thuật và ngôn ngữ tự nhiên.

65. Các quốc gia thành viên cần thực hiện các chính sách để đảm bảo rằng các hành động của các chủ thể AI phù hợp với luật,

tiêu chuẩn và nguyên tắc nhân quyền quốc tế trong suốt vòng đời của các hệ thống AI, đồng thời quan tâm đầy đủ đến sự đa dạng văn hóa và xã hội hiện tại, bao gồm phong tục địa phương và truyền thống tôn giáo liên quan đến khía cạnh đi trước và phổ quát của nhân quyền.

66. Các quốc gia thành viên cần đưa ra các cơ chế để yêu cầu các chủ thể AI tiết lộ và chống lại bất kỳ sự rập khuôn nào trong kết quả của các hệ thống và dữ liệu AI, cho dù do thiết kế hay sơ suất, và để đảm bảo rằng các bộ dữ liệu cho các hệ thống AI không thúc đẩy sự bất bình đẳng về văn hóa, kinh tế hoặc xã hội, định kiến, sự lan truyền thông tin sai lệch và thông tin giả mạo hay gián đoạn quyền tự do ngôn luận và tiếp cận thông tin. Đặc biệt, cần chú ý đến những nơi khan hiếm dữ liệu.

67. Các quốc gia thành viên cần thực hiện các chính sách để thúc đẩy và làm gia tăng sự đa dạng và bao trùm, phản ánh các yêu tố dân số trong các nhóm phát triển AI và các bộ dữ liệu đào tạo, đồng thời để đảm bảo quyền truy cập bình đẳng vào các công nghệ AI cũng các lợi ích, đặc biệt là đối với các nhóm bị thiệt thòi, cả từ khu vực nông thôn và thành thị.

68. Các quốc gia thành viên cần xây dựng, xem xét và điều chỉnh, nếu thích hợp, các khung pháp lý để đạt được trách nhiệm giải trình và trách nhiệm đối với nội dung và kết quả của hệ thống AI bắt nguồn từ các giai đoạn khác nhau trong vòng đời. Các quốc gia thành viên khi cần thiết nên đưa ra các khung trách nhiệm pháp lý hoặc làm rõ việc giải trình các khuôn khổ hiện có để đảm bảo có thể quy trách nhiệm giải trình cho kết quả và hoạt động của hệ thống AI. Hơn nữa, khi xây dựng các khung pháp lý, các quốc gia thành viên, đặc biệt, cần tính đến việc chịu trách nhiệm và trách nhiệm giải trình cuối cùng phải luôn thuộc về con người tự nhiên hoặc pháp nhân và các hệ thống AI không nên được trao nhân cách pháp lý. Để đảm bảo điều này, các khung pháp lý như vậy phải phù hợp với nguyên

tắc giám sát của con người và thiết lập một cách tiếp cận toàn diện tập trung vào các chủ thể AI và các quy trình công nghệ liên quan đến các giai ddaonj khác nhau trong vòng đời của hệ thống AI.

69. Để thiết lập các quy tắc trong trường hợp chưa có, hoặc để điều chỉnh các khung pháp lý hiện có, các quốc gia thành viên cần có sự tham gia của tất cả các chủ thể AI (bao gồm các nhà nghiên cứu, đại diện của tổ chức dân sự và cơ quan thực thi pháp luật, công ty bảo hiểm, nhà đầu tư, nhà sản xuất, kỹ sư, luật sư và người dùng). Các quy tắc có thể phát triển thành các thực tiễn, luật pháp và quy định tốt nhất. Các quốc gia thành viên được khuyến khích sử dụng các cơ chế như nguyên mẫu chính sách và sandbox quy định để đẩy nhanh sự phát triển của luật pháp, quy định và chính sách, bao gồm đánh giá thường xuyên, phù hợp với sự phát triển nhanh chóng của các công nghệ mới và đảm bảo rằng luật pháp và quy định có thể được thử nghiệm trong một môi trường an toàn trước khi được chính thức thông qua. Các quốc gia thành viên cần hỗ trợ chính quyền địa phương xây dựng các chính sách, quy định và pháp luật của địa phương phù hợp với khung pháp lý quốc gia và quốc tế.

70. Các quốc gia thành viên cần đặt ra các yêu cầu rõ ràng về tính minh bạch và khả năng giải trình của hệ thống AI để giúp đảm bảo độ tin cậy của toàn bộ vòng đời hệ thống AI. Các yêu cầu như vậy phải bao gồm việc thiết kế và thực hiện các cơ chế tác động có tính đến bản chất của miền ứng dụng, mục đích sử dụng, đối tượng mục tiêu và tính khả thi của từng hệ thống AI cụ thể.

LĨNH VỰC CHÍNH SÁCH 3: CHÍNH SÁCH DỮ LIỆU

71. Các quốc gia thành viên cần hành động để xây dựng các chiến lược quản trị dữ liệu đảm bảo đánh giá liên tục chất lượng dữ liệu đào tạo cho các hệ thống AI bao gồm tính đầy đủ của các quy trình và lựa chọn cắt dán dữ liệu, các biện pháp bảo mật và bảo vệ dữ liệu thích hợp, cũng như các cơ chế phản hồi để học hỏi từ

những sai lầm và chia sẻ các thực tiễn tốt nhất giữa tất cả các chủ thể AI.

72. Các quốc gia thành viên cần đưa ra các biện pháp bảo vệ thích hợp để bảo vệ quyền riêng tư phù hợp với luật pháp quốc tế, bao gồm giải quyết các mối quan tâm như giám sát. Các quốc gia thành viên, cùng với các quốc gia khác, cần thông qua hoặc thực thi các khuôn khổ pháp lý cung cấp sự bảo vệ thích hợp, tuân thủ luật pháp quốc tế. Các quốc gia thành viên cần khuyến khích mạnh mẽ tất cả các chủ thể AI, bao gồm cả các doanh nghiệp kinh doanh, tuân theo các tiêu chuẩn quốc tế hiện có và đặc biệt là thực hiện các đánh giá tác động quyền riêng tư một cách đầy đủ đầy đủ, như một phần của đánh giá tác động đạo đức, có tính đến tác động kinh tế xã hội một cách rộng hơn của việc xử lý dữ liệu dự định và áp dụng quyền riêng tư theo thiết kế trong hệ thống của họ. Quyền riêng tư cần được tôn trọng, bảo vệ và đẩy mạnh trong suốt vòng đời của các hệ thống AI.

73. Các quốc gia thành viên cần đảm bảo rằng các cá nhân được duy trì quyền đối với dữ liệu cá nhân của họ và được bảo vệ bởi một khuôn khổ quy định: tính minh bạch; các biện pháp bảo vệ thích hợp để xử lý dữ liệu nhạy cảm; mức độ bảo vệ dữ liệu thích hợp; các chương trình và cơ chế trách nhiệm giải trình hiệu quả và có ý nghĩa; hưởng đầy đủ các quyền của chủ thể dữ liệu và khả năng truy cập và xóa dữ liệu cá nhân của họ trong các hệ thống AI, trừ một số trường hợp phù hợp với luật pháp quốc tế; mức độ bảo vệ thích hợp vớ luật bảo vệ dữ liệu, nơi dữ liệu đang được sử dụng cho các mục đích thương mại như cho phép quảng cáo nhắm mục tiêu vi mô, chuyển giữ liệu xuyên biên giới và giám sát độc lập hiệu quả như một phần của cơ chế quản trị dữ liệu giữ cho các cá nhân được kiểm soát dữ liệu cá nhân của họ và thúc đẩy lợi ích của một luồng thông tin tự do quốc tế, bao gồm quyền truy cập vào dữ liệu.

74. Các quốc gia thành viên cần thiết lập các chính sách dữ liệu hoặc các khuôn khổ tương đương, hoặc củng cố các chính sách hiện có, để đảm bảo an ninh đầy đủ cho dữ liệu cá nhân và dữ liệu nhạy cảm, nếu được tiết lộ, có thể gây ra thiệt hại, thương tích hoặc khó khăn khác thường cho các cá nhân. Ví dụ bao gồm dữ liệu liên quan đến tội phạm, tố tụng hình sự và kết án, và các biện pháp an ninh liên quan; dữ liệu sinh trắc học, di truyền và sức khỏe; và dữ liệu cá nhân như liên quan đến chủng tộc, màu da, nguồn gốc, giới tính, tuổi tác, ngôn ngữ, tôn giáo, quan điểm chính trị, nguồn gốc quốc gia, nguồn gốc dân tộc, nguồn gốc xã hội, tình trạng khi sinh ra hoặc khuyết tật kinh tế hoặc xã hội và bất kỳ đặc điểm nào.

75. Các quốc gia thành viên cần đẩy mạnh dữ liệu mở. Về vấn đề này, các quốc gia thành viên cần xem xét các chính sách và khung pháp lý của họ, bao gồm cả việc tiếp cận thông tin và chính phủ mở để phản ánh các yêu cầu cụ thể của AI và các cơ chế thúc đẩy, chẳng hạn như các kho lưu trữ mở cho dữ liệu được tài trợ hoặc lưu giữ công khai và mã nguồn và các quỹ dữ liệu, để hỗ trợ chia sẻ dữ liệu an toàn, công bằng, hợp pháp và phù hợp về đạo đức, cùng với các yêu tố khác.

76. Các quốc gia thành viên cần thúc đẩy và tạo điều kiện cho việc sử dụng các bộ dữ liệu chất lượng và mạnh để đào tạo, phát triển và sử dụng các hệ thống AI, đồng thời cảnh giác trong việc giám sát thu thập và sử dụng dữ liệu. Điều này, nếu có thể và khả thi, bao gồm đầu tư vào việc tạo ra các bộ dữ liệu tiêu chuẩn vàng, như các bộ dữ liệu mở và đáng tin cậy, các bộ dữ liệu này rất đa dạng, được xây dựng trên cơ sở pháp lý hợp lệ, bao gồm cả sự đồng ý của các chủ thể dữ liệu, khi pháp luật yêu cầu. Các tiêu chuẩn để chú thích các bộ dữ liệu cần được khuyến khích, bao gồm phân tách dữ liệu về giới tính và các cơ sở khác, để có thể được xác định cách thu thập dữ liệu và các thuộc tính của dữ liệu.

77. Các quốc gia thành viên, cũng như được đề xuất trong báo cáo của Ủy ban cấp cao về hợp tác kỹ thuật số của Tổng thư ký Liên Hợp Quốc, với sự hỗ trợ của Liên Hợp Quốc và UNESCO, cầnáp dụng cách tiếp cận kỹ thuật số chung đối với dữ liệu khi thích hợp, tăng khả năng tương tác của các công cụ và bộ dữ liệu và giao diện của dữ liệu lưu trữ hệ thống và khuyến khích các công ty tư nhân chia sẻ dữ liệu mà họ thu thập với tất cả các bên liên quan, với các điều kiện thích hợp, cho nghiên cứu, đổi mới hoặc lợi ích chung. Các quốc gia thành viên cũng cần thúc đẩy những nỗ lực công và tư nhằm tạo ra các nền tảng hợp tác để chia sẻ dữ liệu chất lượng trong không gian dữ liệu đáng tin cậy và được bảo mật.

LĨNH VỰC CHÍNH SÁCH 4: PHÁT TRIỂN VÀ HỢP TÁC QUỐC TẾ

78. Các quốc gia thành viên và các tập đoàn xuyên quốc gia nên ưu tiên đạo đức AI bằng cách đưa các cuộc thảo luận về vấn đề đạo đức liên quan đến AI vào các diễn đàn quốc tế, liên chính phủ và đa phương phù hợp.

79. Các quốc gia thành viên cần đảm bảo rằng việc sử dụng AI trong các lĩnh vực phát triển như giáo dục, khoa học, văn hóa, truyền thông và thông tin, chăm sóc sức khỏe, nông nghiệp và cung ứng thực phẩm, môi trường, quản lý tài nguyên thiên nhiên và cơ sở hạ tầng, lập kế hoạch và tăng trưởng kinh tế tuân thủ các giá trị và nguyên tắc được nêu trong Khuyến nghị này.

80. Các quốc gia thành viên cần thông qua các tổ chức quốc tế để cung cấp các nền tảng hợp tác quốc tế về AI để phát triển, bao gồm đóng góp chuyên môn, tài trợ, dữ liệu, kiến thức miền, hạ tầng và hỗ trợ phối hợp giữa các bên liên quan để giải quyết các vấn đề phát triển đầy thách thức, đặc biệt là đối với các LMIC, như LDC, LLDC và SIDS.

81. Các quốc gia thành viên cần hành động để thúc đẩy hợp tác quốc tế về nghiên cứu và đổi mới AI, bao gồm các trung tâm nghiên cứu và đổi mới và mạng lưới thúc đẩy sự tham gia và lãnh đạo lớn hơn của các nhà nghiên cứu từ LMICs và các quốc gia khác, bao gồm LDCs, LLDCs và SIDS.

82. Các quốc gia thành viên cần đẩy mạnh nghiên cứu đạo đức AI bằng cách tham gia vào các viện nghiên cứu và tổ chức quốc tế, cũng như các tập đoàn xuyên quốc gia, có thể là cơ sở cho việc sử dụng các hệ thống AI có đạo đức của các thực thể công cộng và tư nhân, bao gồm nghiên cứu về khả năng áp dụng các khuôn khổ đạo đức cụ thể trong các bối cảnh và nền văn hóa cụ thể và khả năng phát triển các giải pháp khả thi về mặt công nghệ phù hợp với các khuôn khổ này.

83. Các quốc gia thành viên cần khuyến khích hợp tác quốc tế và phối hợp trong lĩnh vực AI để kết nối các đường dây địa công nghệ. Trao đổi và tham vấn kỹ thuật nên diễn ra giữa các quốc gia thành viên và người dân của họ, giữa khu vực công và tư nhân, giữa các quốc gia có công nghệ tiên tiến nhất và chậm phát triển nhất trên cơ sở tôn trọng luật pháp quốc tế.

LĨNH VỰC CHÍNH SÁCH 5: MÔI TRƯỜNG VÀ HỆ SINH THÁI

84. Các quốc gia thành viên và doanh nghiệp kinh doanh nên đánh giá tác động môi trường trực tiếp và gián tiếp trong suốt vòng đời của hệ thống AI, bao gồm, nhưng không giới hạn, lượng khí thải carbon, mức tiêu thụ năng lượng và tác động môi trường của việc khai thác nguyên liệu thô để hỗ trợ sản xuất công nghệ AI, giảm tác động môi trường của các hệ thống AI và cơ sở hạ tầng dữ liệu. Các quốc gia thành viên cần đảm bảo để tất cả các chủ thể AI tuân thủ pháp luật, các chính sách và thực tiễn về môi trường.

85. Các quốc gia thành viên nên đưa ra các ưu đãi, khi cần thiết và phù hợp, để đảm bảo phát triển và áp dụng các giải pháp

dựa trên quyền và đạo đức của AI cho khả năng chống chịu rủi ro thiên tai; giám sát, bảo vệ và tái tạo môi trường và hệ sinh thái; bảo tồn hành tinh. Các hệ thống AI này cần có sự tham gia của các cộng đồng địa phương và bản địa trong suốt vòng đời của các hệ thống AI và nên hỗ trợ các cách tiếp cận kinh tế tuần hoàn và các mô hình tiêu dùng và sản xuất bền vững. Một số ví dụ bao gồm sử dụng các hệ thống AI, khi cần thiết và thích hợp, để:

(a) Hỗ trợ bảo vệ, giám sát và quản lý tài nguyên thiên nhiên.

(b) Hỗ trợ dự đoán, phòng ngừa, kiểm soát và giảm bớt các vấn đề liên quan đến khí hậu.

(c) Hỗ trợ một hệ sinh thái thực phẩm hiệu quả và bền vững hơn.

(d) Hỗ trợ tăng tốc tiếp cận và sử dụng hàng loạt năng lượng bền vững.

(e) Cho phép và thúc đẩy lồng ghép cơ sở hạ tầng bền vững, mô hình kinh doanh và tài chính bền vững cho phát triển bền vững.

(f) Phát hiện các chất ô nhiễm hoặc dự đoán mức độ ô nhiễm và qua đó giúp các bên liên quan xác định, lập kế hoạch và đưa ra các biện pháp can thiệp có mục tiêu để ngăn ngừa và giảm ô nhiễm và phơi nhiễm.

86. Khi lựa chọn các phương pháp AI, với đặc tính tiềm năng về việc sử dụng nhiều dữ liệu hoặc nhiều tài nguyên của một số trong số chúng và tác động tương ứng đối với môi trường, các quốc gia thành viên nên đảm bảo rằng các chủ thể AI, phù hợp với nguyên tắc tỷ lệ, ủng hộ các phương pháp AI hiệu quả về dữ liệu, năng lượng và tài nguyên. Cần xây dựng các yêu cầu để đảm bảo có sẵn bằng chứng thích hợp cho thấy một ứng dụng AI sẽ có hiệu quả dự kiến hoặc các biện pháp bảo vệ đi kèm với ứng dụng AI có thể giải thích cho việc sử dụng nó. Nếu điều này không thể được thực hiện,

nguyên tắc phòng ngừa phải được ưu tiên và trong trường hợp có tác động tiêu cực không cân xứng đối với môi trường, AI không nên được sử dụng.

LĨNH VỰC CHÍNH SÁCH 6: GIỚI TÍNH

87. Các quốc gia thành viên cần đảm bảo rằng tiềm năng của các công nghệ kỹ thuật số và trí tuệ nhân tạo để góp phần đạt được bình đẳng giới được tối đa hóa hoàn toàn và phải đảm bảo rằng các quyền con người và quyền tự do cơ bản của trẻ em gái và phụ nữ, và sự an toàn và toàn vẹn của họ không bị xâm phạm ở bất kỳ giai đoạn nào của vòng đời hệ thống AI. Hơn nữa, Đánh giá tác động đạo đức cần bao gồm khía cạnh chuyển giới.

88. Các quốc gia thành viên nên có các quỹ dành riêng từ ngân sách công liên quan đến tài trợ cho các chương trình đáp ứng giới, đảm bảo các chính sách số quốc gia bao gồm kế hoạch hành động giới và phát triển các chính sách liên quan, ví dụ, về giáo dục lao động, nhằm hỗ trợ trẻ em gái và phụ nữ để đảm bảo họ không bị loại khỏi nền kinh tế số được hỗ trợ bởi AI. Đầu tư đặc biệt cung cấp các chương trình mục tiêu và ngôn ngữ dành riêng cho giới tính, để tăng cơ hội tham gia của trẻ em gái và phụ nữ vào khoa học, công nghệ, kỹ thuật và toán học (STEM), bao gồm các môn học về công nghệ thông tin và truyền thông (ICT), khả năng sẵn sàng, việc làm, phát triển chuyên môn và nghề nghiệp bình đẳng cho trẻ em gái và phụ nữ, nên được xem xét và thực hiện.

89. Các quốc gia thành viên cần đảm bảo phát huy tiềm năng của các hệ thống AI để thúc đẩy việc đạt được bình đẳng giới. Họ nên đảm bảo rằng các công nghệ này không làm trầm trọng thêm khoảng cách giới tính vốn đã rộng ở một số lĩnh vực trong thế giới tương tự, và loại bỏ những khoảng trống đó. Những khoảng trống này bao gồm: khoảng cách giới tính về tiền lương; sự đại diện bất bình đẳng trong một số ngành nghề và hoạt động nhất định; thiếu đại

diện trong các vị trí quản lý cấp cao, hội đồng quản trị hoặc nhóm nghiên cứu trong lĩnh vực AI; khoảng cách giáo dục; tiếp cận kỹ thuật số và AI, khoảng cách trong việc áp dụng, sử dụng và khả năng chi trả; và sự bất bình đẳng trong phân bổ các công việc không được trả lương và trách nhiệm chăm sóc trong xã hội của chúng ta.

90. Các quốc gia thành viên cần đảm bảo rằng sự rập khuôn giới tính và phân biệt đối xử không được đưa vào các hệ thống AI, và thay vào đó cần xác định và chủ động khắc phục những điều này. Cần nỗ lực để tránh tác động tiêu cực phức tạp của khoảng cách công nghệ trong việc đạt được bình đẳng giới và tránh bạo lực như quấy rối, bắt nạt hoặc buôn bán trẻ em gái và phụ nữ và các nhóm thiếu tính đại diện, bao gồm cả trong lĩnh vực trực tuyến.

91. Các quốc gia thành viên cần khuyến khích nữ doanh nhân tham gia vào tất cả các giai đoạn trong vòng đời hệ thống AI bằng cách cung cấp và thúc đẩy các ưu đãi kinh tế, quy định, trong số các chương trình khuyến khích và hỗ trợ khác, cũng như các chính sách nhằm có sự bình đẳng giới tính trong việc tham gia nghiên cứu học thuật về AI, đại diện giới tính trong các vị trí quản lý hàng đầu, trong ban giám đốc và nhóm nghiên cứu của các công ty kỹ thuật số và AI. Các quốc gia thành viên nên đảm bảo rằng các quỹ công (cho đổi mới, nghiên cứu và công nghệ) được dành cho các chương trình và công ty với đại diện giới tính rõ ràng và tương tự, các quỹ tư nhân cũng được khuyến khích thông qua các nguyên tắc hành động khẳng định. Các chính sách về môi trường không quấy rối cần được phát triển và thực thi, cùng với việc khuyến khích chuyển giao các tập quán tốt nhất về cách thúc đẩy sự đa dạng trong suốt vòng đời hệ thống AI.

92. Các quốc gia thành viên nên thúc đẩy sự đa dạng giới tính trong nghiên cứu học thuật AI và ngành công nghiệp bằng cách đưa ra ưu đãi cho trẻ em gái và phụ nữ tham gia vào lĩnh vực này, đưa ra các cơ chế để chống lại sự rập khuôn và quấy rối giới tính trong

cộng đồng nghiên cứu AI, và khuyến khích các thực thể học thuật và tư nhân chia sẻ các tập quán tốt nhất về cách tăng cường đa dạng giới tính.

93. UNESCO có thể giúp xây dựng một kho lưu trữ các thực tiễn tốt nhất để khuyến khích sự tham gia của trẻ em gái, phụ nữ và các nhóm thiểu tính đại diện trong tất cả các giai đoạn của vòng đời hệ thống AI.

LĨNH VỰC CHÍNH SÁCH 7: VĂN HÓA

94. Các quốc gia thành viên được khuyến khích kết hợp các hệ thống AI, khi thích hợp, trong việc bảo tồn, làm giàu, hiểu biết, thúc đẩy, quản lý và khả năng tiếp cận di sản văn hóa hữu hình, tài liệu và phi vật thể, bao gồm các ngôn ngữ có nguy cơ tuyệt chủng cũng như ngôn ngữ và kiến thức bản địa, ví dụ bằng cách triển khai hoặc cập nhật các chương trình giáo dục liên quan đến việc áp dụng các hệ thống AI trong các lĩnh vực này khi thích hợp và bằng cách đảm bảo cách tiếp cận có sự tham gia, nhắm vào các tổ chức và công chúng.

95. Các quốc gia thành viên được khuyến khích kiểm tra và giải quyết tác động văn hóa của các hệ thống AI, đặc biệt là các ứng dụng xử lý ngôn ngữ tự nhiên (NLP) như dịch tự động và trợ lý giọng nói, về các sắc thái của ngôn ngữ và biểu hiện của con người. Những đánh giá như vậy sẽ cung cấp đầu vào cho việc thiết kế và triển khai các chiến lược tối đa hóa lợi ích từ các hệ thống này bằng cách thu hẹp khoảng cách văn hóa và tăng sự hiểu biết của con người, cũng như giải quyết các tác động tiêu cực như giảm sử dụng, có thể dẫn đến sự biến mất của các ngôn ngữ có nguy cơ tuyệt chủng, các thổ ngữ địa phương và các biến thể âm điệu và văn hóa liên quan đến ngôn ngữ và biểu hiện của con người.

96. Các quốc gia thành viên cần đẩy mạnh giáo dục AI và đào tạo kỹ thuật số cho các nghệ sĩ và chuyên gia sáng tạo để đánh giá sự phù hợp của các công nghệ AI để sử dụng trong nghề nghiệp của

họ, và góp phần thiết kế và triển khai các công nghệ AI phù hợp, vì các công nghệ AI đang được sử dụng để chế tạo, sản xuất, phân phối, phát sóng và tiêu thụ nhiều loại hàng hóa và dịch vụ văn hóa, lưu ý tầm quan trọng của việc bảo tồn di sản văn hóa, sự đa dạng và tự do nghệ thuật.

97. Các quốc gia thành viên cần thúc đẩy nhận thức và đánh giá các công cụ AI trong các ngành công nghiệp văn hóa địa phương và các doanh nghiệp vừa và nhỏ hoạt động trong lĩnh vực văn hóa, để tránh nguy cơ tập trung vào thị trường văn hóa.

98. Các quốc gia thành viên nên thu hút các công ty công nghệ và các bên liên quan khác để thúc đẩy nguồn cung cấp đa dạng và quyền truy cập đa dạng vào các biểu hiện văn hóa, và đặc biệt là để đảm bảo rằng thuật toán khuyến nghị tăng cường khả năng hiển thị và khả năng khám phá nội dung địa phương.

99. Các quốc gia thành viên nên thúc đẩy nghiên cứu mới về mối liên hệ giữa AI và sở hữu trí tuệ (IP), ví dụ để xác định xem có nên và làm cách nào để bảo vệ các tác phẩm được tạo ra bằng công nghệ AI thông qua quyền sở hữu trí tuệ. Các quốc gia thành viên cũng nên đánh giá ảnh hưởng của các công nghệ AI đến quyền hoặc lợi ích của chủ sở hữu IP như thế nào, những công trình được sử dụng để nghiên cứu, phát triển, đào tạo hoặc triển khai các ứng dụng AI.

100. Các quốc gia thành viên nên khuyến khích các bảo tàng, phòng trưng bày, thư viện và kho lưu trữ ở cấp quốc gia sử dụng các hệ thống AI để làm nổi bật bộ sưu tập của họ và tăng cường thư viện, cơ sở dữ liệu và cơ sở kiến thức của họ, đồng thời cung cấp quyền truy cập cho người dùng của họ.

LĨNH VỰC CHÍNH SÁCH 8: GIÁO DỤC VÀ NGHIÊN CỨU

101. Các quốc gia thành viên cần phối hợp với các tổ chức quốc tế, các tổ chức giáo dục và các tổ chức tư nhân và phi chính

phủ để cung cấp đầy đủ giáo dục AI cho công chúng ở các cấp độ tại tất cả các quốc gia để trao quyền cho mọi người và giảm bớt khoảng cách kỹ thuật số và sự bất bình đẳng truy cập số từ việc áp dụng rộng rãi các hệ thống AI.

102. Các quốc gia thành viên nên thúc đẩy việc tiếp thu "các kỹ năng tiên quyết" cho giáo dục AI, chẳng hạn như kỹ năng đọc viết cơ bản, tính toán, mã hóa và kỹ năng số, truyền thông và thông tin, cũng như tư duy phê phán và sáng tạo, làm việc theo nhóm, giao tiếp, kỹ năng đạo đức AI và cảm xúc xã hội, đặc biệt là ở các quốc gia, khu vực hoặc các vùng trong các quốc gia có khoảng cách đáng kể trong việc giáo dục các kỹ năng này.

103. Các quốc gia thành viên nên thúc đẩy các chương trình nhận thức chung về phát triển AI, bao gồm dữ liệu và các cơ hội và thách thức do công nghệ AI mang lại, tác động của các hệ thống AI đối với nhân quyền và ý nghĩa của chúng, bao gồm quyền trẻ em. Cần cấp quyền truy cập các chương trình này cho cả các nhóm phi kỹ thuật và kỹ thuật.

104. Các quốc gia thành viên cần khuyến khích các sáng kiến nghiên cứu về việc sử dụng có trách nhiệm và đạo đức các công nghệ AI trong giảng dạy, đào tạo giáo viên và học tập trực tuyến để tăng cường cơ hội và giảm thiểu những thách thức và rủi ro liên quan đến lĩnh vực này. Các sáng kiến cần đi kèm với đánh giá đầy đủ về chất lượng giáo dục và tác động đến học sinh và giáo viên về việc sử dụng các công nghệ AI. Các quốc gia thành viên cũng cần đảm bảo rằng các công nghệ AI trao quyền cho học sinh và giáo viên và nâng cao kinh nghiệm của họ, ghi nhớ rằng các khía cạnh quan hệ và xã hội và giá trị của các hình thức giáo dục truyền thống là rất quan trọng trong mối quan hệ giữa giáo viên và học sinh và giữa học sinh với nhau khi thảo luận về việc áp dụng các công nghệ AI trong giáo dục. Các hệ thống AI được sử dụng trong học tập phải tuân theo các yêu cầu nghiêm ngặt khi theo dõi, đánh giá khả năng hoặc

dự đoán hành vi của người học. AI cần hỗ trợ quá trình học tập mà không làm giảm khả năng nhận thức và không trích xuất thông tin nhạy cảm, tuân thủ các tiêu chuẩn bảo vệ dữ liệu cá nhân liên quan. Dữ liệu được bàn giao để có kiến thức thu thập được trong quá trình tương tác của người học với hệ thống AI không được sử dụng sai mục đích, chiếm dụng hoặc khai thác hình sự, kể cả cho mục đích thương mại.

105. Các quốc gia thành viên cần thúc đẩy sự tham gia và lãnh đạo của trẻ em gái và phụ nữ, sự đa dạng sắc tộc và văn hóa, người khuyết tật, người bị thiệt thòi và dễ bị tổn thương hoặc những người trong các tình huống dễ bị tổn thương, dân tộc thiểu số và tất cả những người không được hưởng đầy đủ lợi ích của việc đưa kỹ thuật số vào các chương trình giáo dục AI ở tất cả các cấp, cũng như giám sát và chia sẻ các tập quán tốt nhất về vấn đề này với các quốc gia thành viên khác.

106. Các quốc gia thành viên nên xây dựng giáo trình giảng dạy đạo đức AI cho tất cả các cấp phù hợp với các chương trình và truyền thống giáo dục quốc gia của họ, và thúc đẩy sự hợp tác chéo giữa giáo dục kỹ năng, kỹ thuật AI với các khía cạnh nhân văn, đạo đức và xã hội của giáo dục AI. Các khóa học trực tuyến và tài nguyên số của giáo dục đạo đức AI nên được phát triển bằng các ngôn ngữ địa phương, bao gồm cả ngôn ngữ bản địa và tính đến sự đa dạng của môi trường, đặc biệt là đảm bảo khả năng tiếp cận các định dạng cho người khuyết tật.

107. Các quốc gia thành viên nên thúc đẩy và hỗ trợ nghiên cứu AI, đặc biệt là nghiên cứu đạo đức AI, bao gồm cả việc đầu tư vào nghiên cứu hoặc bằng cách tạo động lực cho khu vực công và tư nhân đầu tư vào lĩnh vực này, với nhận thức rằng nghiên cứu góp phần đáng kể vào sự phát triển và cải tiến hơn nữa công nghệ AI nhằm thúc đẩy luật pháp quốc tế và các giá trị và nguyên tắc được nêu trong Khuyến nghị này. Các quốc gia thành viên cũng nên công

khai thúc đẩy các tập quán tốt nhất và hợp tác với các nhà nghiên cứu và các công ty phát triển AI một cách có đạo đức.

108. Các quốc gia thành viên cần đảm bảo rằng các nhà nghiên cứu AI được đào tạo về đạo đức nghiên cứu và yêu cầu họ cân nhắc về đạo đức trong các thiết kế, sản phẩm và ấn phẩm của họ, đặc biệt là trong các phân tích về bộ dữ liệu họ sử dụng, cách chúng được chú thích, chất lượng, phạm vi của kết quả và các ứng dụng.

109. Các quốc gia thành viên nên khuyến khích các công ty ở khu vực tư nhân hỗ trợ cộng đồng khoa học tiếp cận dữ liệu nghiên cứu, đặc biệt là trong LMICs, cụ thể là LDCs, LLDCs và SIDS. Quyền truy cập này phải phù hợp với các tiêu chuẩn bảo vệ dữ liệu và quyền riêng tư có liên quan.

110. Để có một đánh giá quan trọng về nghiên cứu AI và sự giám sát thích hợp đối với việc lạm dụng hoặc các tác động bất lợi tiềm ẩn, các quốc gia thành viên nên đảm bảo rằng bất kỳ sự phát triển nào trong tương lai liên quan đến công nghệ AI phải dựa trên nghiên cứu khoa học nghiêm ngặt và độc lập, thúc đẩy nghiên cứu AI liên ngành bằng cách đưa vào các ngành ngoài khoa học, công nghệ, kỹ thuật và toán học (STEM) như nghiên cứu văn hóa, giáo dục, đạo đức, quan hệ quốc tế, luật, ngôn ngữ học, triết học, khoa học chính trị, xã hội học và tâm lý học.

111. Thừa nhận rằng các công nghệ AI mang lại cơ hội lớn để giúp nâng cao kiến thức và thực hành khoa học, đặc biệt là trong các ngành theo mô hình truyền thống, các quốc gia thành viên nên khuyến khích các cộng đồng khoa học nhận thức được lợi ích, giới hạn và rủi ro khi sử dụng chúng; điều này bao gồm cố gắng đảm bảo rằng các kết luận được rút ra từ các phương pháp tiếp cận, mô hình và phương pháp xử lý dựa trên dữ liệu là mạnh mẽ và hợp lý. Hơn nữa, các quốc gia thành viên cần hoan nghênh và hỗ trợ vai trò của cộng đồng khoa học trong việc đóng góp cho chính sách và trau dồi nhận thức về những điểm mạnh và điểm yếu của công nghệ AI.

LĨNH VỰC CHÍNH SÁCH 9: TRUYỀN THÔNG VÀ THÔNG TIN

112. Các quốc gia thành viên nên sử dụng các hệ thống AI để cải thiện khả năng tiếp cận thông tin và kiến thức. Điều này có thể bao gồm hỗ trợ cho các nhà nghiên cứu, học thuật, nhà báo, công chúng và các nhà phát triển, để tăng cường tự do ngôn luận, tự do học thuật và khoa học, tiếp cận thông tin và chủ động công bố các dữ liệu và thông tin chính thức.

113. Các quốc gia thành viên cần đảm bảo rằng các chủ thể AI tôn trọng và thúc đẩy quyền tự do ngôn luận cũng như tiếp cận thông tin liên quan đến việc tạo nội dung tự động, kiểm duyệt và lựa chọn nội dung. Các khuôn khổ thích hợp, bao gồm quy định, sẽ đảm bảo sự minh bạch cho truyền thông trực tuyến và các nhà khai thác thông tin, đảm bảo người dùng có quyền truy cập các quan điểm đa dạng, cũng như các quy trình thông báo kịp thời cho người dùng về lý do xóa bỏ hoặc xử lý nội dung khác cũng như cơ chế khiếu nại cho phép người dùng tìm cách khắc phục.

114. Các quốc gia thành viên nên đầu tư và thúc đẩy các kỹ năng số, truyền thông và thông tin để tăng cường tư duy phê phán và năng lực cần thiết để hiểu cách sử dụng và ý nghĩa của các hệ thống AI, để giảm thiểu và chống lại thông tin sai lệch và ngôn từ kích động thù địch. Một sự hiểu biết và đánh giá tốt hơn về cả các tác động tích cực cũng như khả năng gây hại của các hệ thống khuyến nghị nên là một phần của những nỗ lực đó.

115. Các quốc gia thành viên nên tạo ra các môi trường cho phép các phương tiện truyền thông có quyền và nguồn lực để báo cáo hiệu quả về lợi ích và tác hại của các hệ thống AI, đồng thời khuyến khích các phương tiện truyền thông sử dụng các hệ thống AI có đạo đức trong hoạt động của họ.

LĨNH VỰC CHÍNH SÁCH 10: KINH TẾ VÀ LAO ĐỘNG

116. Các quốc gia thành viên cần đánh giá và giải quyết tác động của các hệ thống AI đối với thị trường lao động và ý nghĩa của nó đối với các yêu cầu giáo dục ở tất cả các quốc gia và đặc biệt chú trọng các quốc gia có nền kinh tế thâm dụng lao động. Điều này có thể bao gồm việc giới thiệu một loạt các kỹ năng "cốt lõi" và liên ngành ở tất cả các cấp học để cung cấp cho người lao động hiện tại và các thế hệ mới một cơ hội công bằng nhằm tìm kiếm việc làm trong một thị trường đang thay đổi nhanh chóng và để đảm bảo nhận thức của họ về các khía cạnh đạo đức của các hệ thống AI. Các kỹ năng như "học phương pháp học", giao tiếp, tư duy phê phán, làm việc theo nhóm, đồng cảm và khả năng chuyển giao kiến thức của một người trên các lĩnh vực, nên được dạy cùng với các kỹ năng kỹ thuật và đặc biệt, cũng như các công việc có tay nghề thấp. Quan trọng là cần có sự minh bạch về những kỹ năng cần thiết và cập nhật chương trình giảng dạy xung quanh những kỹ năng này.

117. Các quốc gia thành viên nên hỗ trợ các thỏa thuận hợp tác giữa các chính phủ, các tổ chức học thuật, các tổ chức giáo dục và đào tạo nghề, ngành công nghiệp, tổ chức của người lao động và tổ chức dân sự để kết nối các yêu cầu về kỹ năng phù hợp với các chương trình và chiến lược đào tạo đáp ứng công việc tương lai và nhu cầu của ngành công nghiệp, bao gồm cả các doanh nghiệp vừa và nhỏ. Các phương pháp giảng dạy và học tập dựa trên dự án cho AI nên được thúc đẩy, cho phép hợp tác giữa các tổ chức công cộng, các công ty khu vực tư nhân, các trường đại học và trung tâm nghiên cứu.

118. Các quốc gia thành viên nên làm việc với các công ty ở khu vực tư nhân, các tổ chức dân sự và các bên liên quan khác, bao gồm cả người lao động và công đoàn để đảm bảo một quá trình chuyển đổi công bằng cho nhân viên có khả năng bị ảnh hưởng. Điều này bao gồm việc áp dụng các chương trình nâng cao kỹ năng

và đào tạo lại, tìm kiếm các cơ chế hiệu quả để giữ chân nhân viên trong những giai đoạn chuyển tiếp đó và khám phá các chương trình "mạng lưới an toàn" cho những người không thể được đào tạo lại. Các quốc gia thành viên nên xây dựng và thực hiện các chương trình để nghiên cứu và giải quyết các thách thức được xác định, có thể bao gồm việc nâng cao kỹ năng và đào tạo lại, tăng cường bảo trợ xã hội, các chính sách công nghiệp và can thiệp chủ động, tiền thuế hay các hình thức thuế mới. Các quốc gia thành viên nên đảm bảo rằng có đủ ngân sách công để hỗ trợ các chương trình này. Các quy định liên quan, chẳng hạn như hệ thống thuế, nên được kiểm tra cẩn thận và thay đổi nếu cần thiết để chống lại hệ quả của thất nghiệp do tự động hóa dựa trên AI gây ra.

119. Các quốc gia thành viên nên khuyến khích và hỗ trợ các nhà nghiên cứu phân tích tác động của các hệ thống AI đối với môi trường lao động địa phương để dự đoán các xu hướng và thách thức trong tương lai. Những nghiên cứu này nên có cách tiếp cận liên ngành và điều tra tác động của các hệ thống AI đối với các lĩnh vực kinh tế, lxã hội và địa lý, cũng như đối với tương tác giữa con người và robot và các mối quan hệ của con người, để tư vấn về việc đào tạo lại và tái triển khai các tập quán tốt nhất.

120. Các quốc gia thành viên nên thực hiện các bước thích hợp để đảm bảo thị trường cạnh tranh và bảo vệ người tiêu dùng, xem xét các biện pháp và cơ chế có thể ở cấp quốc gia, khu vực và quốc tế, để ngăn chặn việc lạm dụng các vị trí thị trường thống trị, bao gồm cả độc quyền, liên quan đến các hệ thống AI trong suốt vòng đời của chúng, cho dù đó là dữ liệu, nghiên cứu, công nghệ hoặc thị trường. Các quốc gia thành viên nên ngăn chặn việc dẫn đến sự bất bình đẳng, đánh giá các thị trường liên quan và thúc đẩy thị trường cạnh tranh. Cần xem xét kỹ lưỡng đối với LMICs, đặc biệt là LDCs, LLDCs và SIDS, đối tượng dễ bị phơi bày và dễ bị tổn thương trước khả năng lạm dụng sự thống trị thị trường do thiếu hụt

cơ sở hạ tầng, năng lực con người và các quy định. Các chủ thể AI phát triển các hệ thống AI ở các quốc gia đã thiết lập hoặc áp dụng các tiêu chuẩn đạo đức về AI nên tôn trọng các tiêu chuẩn này khi xuất khẩu các sản phẩm, phát triển hoặc áp dụng hệ thống AI của họ ở các quốc gia nơi các tiêu chuẩn đó có thể không tồn tại, đồng thời tôn trọng luật pháp quốc tế và trong nước phù hợp, các tiêu chuẩn và tập quán của các quốc gia này.

LĨNH VỰC CHÍNH SÁCH 11: SỨC KHỎE VÀ PHÚC LỢI XÃ HỘI

121. Các quốc gia thành viên cần nỗ lực sử dụng các hệ thống AI hiệu quả để cải thiện sức khỏe con người và bảo vệ quyền sống, bao gồm giảm thiểu sự bùng phát dịch bệnh, đồng thời xây dựng và duy trì tình đoàn kết quốc tế để giải quyết các rủi ro và bất ổn về sức khỏe toàn cầu, và đảm bảo rằng việc triển khai các hệ thống AI trong chăm sóc sức khỏe phù hợp với luật pháp quốc tế và nghĩa vụ pháp lý về nhân quyền. Các quốc gia thành viên nên đảm bảo rằng các tác nhân tham gia vào các hệ thống AI chăm sóc sức khỏe cân nhắc tầm quan trọng của mối quan hệ giữa bệnh nhân với gia đình của họ và với nhân viên chăm sóc sức khỏe.

122. Các quốc gia thành viên cần đảm bảo rằng việc xây dựng và triển khai các hệ thống AI liên quan đến sức khỏe nói chung và sức khỏe tâm thần nói riêng, chú ý đúng mức đến trẻ em và thanh thiếu niên, phải đảm bảo quy định về an toàn, hiệu quả, được chứng minh bởi khoa học và y tế, tạo sự tiến bộ về y tế và đổi mới dựa trên bằng chứng. Hơn nữa, trong lĩnh vực liên quan đến can thiệp y tế kỹ thuật số, các quốc gia thành viên được khuyến khích tích cực tham gia với bệnh nhân và đại diện của họ trong tất cả các bước liên quan của sự phát triển hệ thống.

123. Các quốc gia thành viên cần đặc biệt chú ý trong việc quy định các giải pháp dự đoán, phát hiện và điều trị chăm sóc sức khỏe

trong các ứng dụng AI bằng cách:

(a) đảm bảo giám sát để giảm thiểu sự thiên vị;

(b) đảm bảo rằng chuyên gia, bệnh nhân, người chăm sóc hoặc sử dụng dịch vụ được bao gồm như một "chuyên gia lĩnh vực" trong nhóm ở tất cả các bước liên quan khi xây dựng các thuật toán;

(c) chú ý đúng mức đến quyền riêng tư vì tiềm năng nhu cầu được theo dõi y tế và đảm bảo rằng tất cả các yêu cầu về bảo vệ dữ liệu quốc gia và quốc tế liên quan được đáp ứng;

(d) đảm bảo các cơ chế hiệu quả để những người có dữ liệu cá nhân được phân tích có nhận thức và đồng ý về việc sử dụng và phân tích dữ liệu của họ, không ngăn cản quyền tiếp cận chăm sóc sức khỏe;

(e) đảm bảo sự chăm sóc của con người và quyết định chẩn đoán và điều trị cuối cùng luôn được con người thực hiện trong khi thừa nhận rằng các hệ thống AI cũng có thể hỗ trợ công việc của họ;

(f) đảm bảo, khi cần thiết, việc đánh giá các hệ thống AI bởi một ủy ban nghiên cứu đạo đức trước khi sử dụng lâm sàng.

124. Các quốc gia thành viên nên tiến hành nghiên cứu về tác động và quy định các tác hại tiềm ẩn đối với sức khỏe tâm thần liên quan đến các hệ thống AI, chẳng hạn như mức độ trầm cảm, lo lắng, cô lập xã hội cao hơn, khả năng gây nghiện, buôn bán, cực đoan hóa và thông tin sai lệch.

125. Các quốc gia thành viên nên xây dựng các hướng dẫn về tương tác giữa người và robot và tác động của chúng đối với các mối quan hệ giữa con người với nhau, dựa trên nghiên cứu và hướng đến sự phát triển trong tương lai của robot, và đặc biệt chú ý đến sức khỏe tinh thần và thể chất của con người. Cần đặc biệt chú ý đến việc sử dụng robot trong chăm sóc sức khỏe và chăm sóc người già, người khuyết tật, trong giáo dục và robot để sử dụng cho trẻ em, đồ chơi

robot, phần mềm trả lời tự động và robot đồng hành cho trẻ em và người lớn. Hơn nữa, cần sự hỗ trợ của công nghệ AI để tăng tính an toàn và sử dụng công thái học của robot, bao gồm cả trong môi trường làm việc giữa người và robot. Đặc biệt chú ý khả năng sử dụng AI để thao túng và lạm dụng các thành kiến nhận thức của con người.

126. Các quốc gia thành viên cần đảm bảo rằng những tương tác giữa con người và robot tuân thủ các giá trị và nguyên tắc tương tự như áp dụng cho bất kỳ hệ thống AI nào khác, bao gồm nhân quyền và các quyền tự do cơ bản, thúc đẩy sự đa dạng và bảo vệ những người dễ bị tổn thương hoặc trong các tình huống dễ bị tổn thương. Các câu hỏi đạo đức liên quan đến các hệ thống hỗ trợ AI cho các công nghệ thần kinh và giao diện não-máy tính nên được cân nhắc để giữ gìn phẩm giá và quyền tự chủ của con người.

127. Các quốc gia thành viên cần đảm bảo rằng người dùng có thể dễ dàng xác định xem họ đang tương tác với một sinh vật sống hay với một hệ thống AI bắt chước các đặc điểm của con người hoặc động vật, và có thể từ chối sự tương tác đó một cách hiệu quả và yêu cầu sự can thiệp của con người.

128. Các quốc gia thành viên cần thực hiện các chính sách để nâng cao nhận thức về nhân hóa các công nghệ AI và công nghệ nhận dạng và bắt chước cảm xúc của con người, bao gồm cả trong ngôn ngữ được sử dụng để đề cập đến chúng và đánh giá các biểu hiện, ý nghĩa đạo đức và những hạn chế có thể có của sự nhân hóa đó, đặc biệt là trong bối cảnh tương tác giữa robot với con người và khi có trẻ em tham gia.

129. Các quốc gia thành viên nên khuyến khích và thúc đẩy hợp tác nghiên cứu về tác động của sự tương tác lâu dài giữa con người với hệ thống AI, đặc biệt chú ý đến tác động tâm lý và nhận thức mà các hệ thống này có thể có đối với trẻ em và thanh thiếu niên. Điều này nên được thực hiện bằng cách sử dụng nhiều chuẩn

mực, nguyên tắc, giao thức, cách tiếp cận kỷ luật và đánh giá về việc sửa đổi hành vi và thói quen, cũng như đánh giá cẩn thận các tác động văn hóa và xã hội. Hơn nữa, các quốc gia thành viên nên khuyến khích nghiên cứu về tác động của công nghệ AI đối với kết quả hoạt động của hệ thống y tế và sức khỏe.

130. Các quốc gia thành viên, cũng như tất cả các bên liên quan, nên áp dụng các cơ chế để thu hút trẻ em và thanh thiếu niên tham gia một cách có ý nghĩa vào các cuộc trò chuyện, tranh luận và ra quyết định liên quan đến tác động của AI đối với cuộc sống và tương lai của họ.

IV. GIÁM SÁT VÀ ĐÁNH GIÁ

131. Các quốc gia thành viên, tùy theo các điều kiện cụ thể, bộ máy quản lý và các quy định hiến pháp, cần giám sát và đánh giá các chính sách, chương trình và cơ chế liên quan đến đạo đức của AI một cách đáng tin cậy và minh bạch, sử dụng kết hợp các phương pháp định lượng và định tính. Để hỗ trợ các quốc gia thành viên, UNESCO có thể đóng góp bằng cách:

(a) xây dựng phương pháp đánh giá tác động đạo đức công nghệ AI (EIA) của UNESCO dựa trên nghiên cứu khoa học chặt chẽ và luật nhân quyền quốc tế, hướng dẫn thực hiện trong tất cả các giai đoạn trong vòng đời hệ thống AI, và các tài liệu xây dựng năng lực để hỗ trợ các quốc gia thành viên trong việc đào tạo các quan chức chính phủ, các nhà hoạch định chính sách và các chủ thể AI phù hợp khác về phương pháp EIA;

(b) xây dựng phương pháp đánh giá sự sẵn sàng của UNESCO để hỗ trợ các quốc gia thành viên xác định khả năng sẵn sàng của họ tại các thời điểm cụ thể trên nhiều khía cạnh;

(c) xây dựng một phương pháp của UNESCO để đánh giá về hiệu lực và hiệu quả trước và sau khi ban hành các chính sách về đạo đức và khuyến khích AI chống lại các mục tiêu đã xác định;

(d) tăng cường phân tích và báo cáo dựa trên nghiên cứu và bằng chứng về các chính sách liên quan đến đạo đức AI;

(e) thu thập và phổ biến về những tiến bộ, đổi mới, báo cáo nghiên cứu, ấn phẩm khoa học, dữ liệu và thống kê liên quan đến các chính sách đạo đức AI, bao gồm thông qua các sáng kiến hiện có, để hỗ trợ chia sẻ các tập quán tốt nhất và học hỏi lẫn nhau, và thúc đẩy việc triển khai Khuyến nghị này.

132. Các quy trình giám sát và đánh giá phải đảm bảo sự tham gia rộng rãi của tất cả các bên liên quan, bao gồm, nhưng không giới hạn, những người dễ bị tổn thương hoặc trong các tình huống dễ bị tổn thương. Sự đa dạng về xã hội, văn hóa và giới tính cần được đảm bảo nhằm cải thiện quá trình học tập và tăng cường kết nối giữa các phát hiện, việc ra quyết định, sự minh bạch và trách nhiệm giải trình về kết quả.

133. Vì lợi ích của việc thúc đẩy các chính sách và tập quán tốt nhất liên quan đến đạo đức trong AI, cần xây dựng các công cụ và chỉ số phù hợp để đánh giá hiệu lực và hiệu quả của nó so với các tiêu chuẩn, ưu tiên và mục tiêu đã nhất trí, bao gồm các mục tiêu cụ thể cho những người thuộc nhóm dân số bị khó khăn, thiệt thòi, những người dễ bị tổn thương hoặc trong các tình huống dễ bị tổn thương, cũng như tác động của các hệ thống AI ở cấp độ cá nhân và xã hội. Việc giám sát và đánh giá tác động của các hệ thống AI, các tập quán và chính sách đạo đức AI liên quan nên được thực hiện liên tục một cách có hệ thống tương ứng với các rủi ro liên quan. Điều này nên dựa trên các khuôn khổ được quốc tế nhất trí và liên quan đến việc đánh giá các tổ chức, nhà cung cấp và chương trình tư nhân và công cộng, bao gồm tự đánh giá, cũng như nghiên cứu dấu vết và phát triển các bộ chỉ số. Việc thu thập và xử lý dữ liệu phải được tiến hành theo luật pháp quốc tế và quốc gia về bảo vệ dữ liệu và quyền riêng tư dữ liệu, và các giá trị và nguyên tắc được nêu trong Khuyến nghị này.

134. Đặc biệt, các quốc gia thành viên có thể muốn xem xét các cơ chế có thể để giám sát và đánh giá, chẳng hạn như ủy ban đạo đức, quan sát đạo đức AI, kho lưu trữ bao gồm sự phát triển tuân thủ nhân quyền và đạo đức của các hệ thống AI hoặc đóng góp cho các sáng kiến hiện có bằng cách giải quyết việc tuân thủ các nguyên tắc đạo đức trên các lĩnh vực năng lực của UNESCO, cơ chế chia sẻ kinh nghiệm, hộp quy định AI và hướng dẫn đánh giá cho tất cả các chủ thể AI để đánh giá sự tuân thủ đối với các khuyến nghị chính sách được đề cập trong tài liệu này.

V. SỬ DỤNG VÀ KHAI THÁC KHUYẾN NGHỊ

135. Các quốc gia thành viên và tất cả các bên liên quan khác được xác định trong Khuyến nghị này cần tôn trọng, thúc đẩy và bảo vệ các giá trị đạo đức, nguyên tắc và tiêu chuẩn liên quan đến AI được xác định trong Khuyến nghị này và cần thực hiện tất cả các bước khả thi để các khuyến nghị chính sách có hiệu lực.

136. Các quốc gia thành viên cần cố gắng mở rộng và bổ sung hành động của riêng đối với Khuyến nghị này bằng cách hợp tác với tất cả các tổ chức chính phủ và phi chính phủ có liên quan trong nước và quốc tế, cũng như các tập đoàn và tổ chức khoa học xuyên quốc gia, có hoạt động nằm trong phạm vi và mục tiêu của Khuyến nghị này. Việc phát triển phương pháp Đánh giá Tác động Đạo đức của UNESCO và thành lập các ủy ban quốc gia về đạo đức AI có thể là công cụ quan trọng cho việc này.

VI. THÚC ĐẨY KHUYẾN NGHỊ HIỆN TẠI

137. UNESCO có khả năng trở thành cơ quan chủ chốt của Liên Hợp quốc để thúc đẩy và phổ biến Khuyến nghị này, và theo đó sẽ hợp tác với các tổ chức liên quan khác của Liên Hợp Quốc, đồng thời tôn trọng các nhiệm vụ của họ và tránh trùng lặp trong hoạt động.

138. UNESCO, bao gồm các cơ quan của mình, như Ủy ban Đạo đức khoa học và Công nghệ Thế giới (COMEST), Ủy ban Đạo

đức Sinh học Quốc tế (IBC) và Ủy ban Đạo đức Sinh học liên chính phủ (IGBC), cũng sẽ hợp tác với các tổ chức chính phủ và phi chính phủ quốc tế, khu vực và tiểu vùng khác.

139. Mặc dù trong UNESCO, nhiệm vụ thúc đẩy và bảo vệ thuộc thẩm quyền của các chính phủ và các cơ quan liên chính phủ, tổ chức dân sự sẽ là một tác nhân quan trọng để ủng hộ lợi ích của khu vực công và do đó UNESCO cần đảm bảo và thúc đẩy tính hợp pháp của nó.

VII. ĐIỀU KHOẢN CUỐI CÙNG

140. Khuyến nghị này cần được hiểu như một tổng thể, các giá trị và nguyên tắc nền tảng phải được hiểu là có sự bổ sung và tương quan lẫn nhau.

141. Khuyến nghị này không bao gồm bất kỳ điều gì có thể hiểu là thay thế, sửa đổi hoặc làm tổn hại đến nghĩa vụ hoặc quyền của các quốc gia theo luật pháp quốc tế, hoặc là sự chấp thuận cho bất kỳ quốc gia, tổ chức chính trị, kinh tế hay xã hội nào khác, hay nhóm hoặc người nào tham gia vào bất kỳ hoạt động hoặc thực hiện bất kỳ hành động nào trái với nhân quyền, các quyền tự do cơ bản, phẩm giá con người và mối quan tâm đối với môi trường và hệ sinh thái cả sống và không còn sống.

ANNEX

DRAFT TEXT OF THE RECOMMENDATION ON THE ETHICS OF ARTIFICIAL INTELLIGENCE PREAMBLE

The General Conference of the United Nations Educational, Scientific and Cultural Organization (UNESCO), meeting in Paris from 9 to 24, at its 41st session,

Recognizing the profound and dynamic positive and negative impacts of artificial intelligence (AI) on societies, environment, ecosystems and human lives, including the human mind, in part because of the new ways in which its use influences human thinking, interaction and decision-making and affects education, human, social and natural sciences, culture, and communication and information,

Recalling that, by the terms of its Constitution, UNESCO seeks to contribute to peace and security by promoting collaboration among nations through education, the sciences, culture, and communication and information, in order to further universal respect for justice, for the rule of law and for the human rights and fundamental freedoms which are affirmed for the peoples of the world,

Convinced that the Recommendation presented here, as a standard-setting instrument developed through a global approach, based on international law, focusing on human dignity and human rights, as well as gender equality, social and economic justice and development, physical and mental well-being, diversity, interconnectedness, inclusiveness, and environmental and ecosystem protection can guide AI technologies in a responsible direction,

Guided by the purposes and principles of the Charter of the United Nations,

Considering that AI technologies can be of great service to humanity and all countries can benefit from them, but also raise fundamental ethical concerns, for instance regarding the biases they can embed and exacerbate, potentially resulting in discrimination, inequality, digital divides, exclusion and a threat to cultural, social and biological diversity and social or economic divides; the need for transparency and understandability of the workings of algorithms and the data with which they have been trained; and their potential impact on, including but not limited to, human dignity, human rights and fundamental freedoms, gender equality, democracy, social, economic, political and cultural processes, scientific and engineering practices, animal welfare, and the environment and ecosystems,

Also recognizing that AI technologies can deepen existing divides and inequalities in the world, within and between countries, and that justice, trust and fairness must be upheld so that no country and no one should be left behind, either by having fair access to AI technologies and enjoying their benefits or in the protection against their negative implications, while recognizing the different circumstances of different countries and respecting the desire of some people not to take part in all technological developments,

Conscious of the fact that all countries are facing an acceleration in the use of information and communication technologies and AI technologies, as well as an increasing need for media and information literacy, and that the digital economy presents important societal, economic and environmental challenges and opportunities of benefit-sharing, especially for low- and middle-income countries (LMICs), including but not limited to least developed countries (LDCs), landlocked developing countries (LLDCs) and small island developing States (SIDS), requiring the recognition, protection and promotion of endogenous cultures, values and knowledge in order to develop sustainable digital economies,

Further recognizing that AI technologies have the potential to be beneficial to the environment and ecosystems, and in order for those benefits to be realized, potential harms to and negative impacts on the environment and ecosystems should not be ignored but instead addressed,

Noting that addressing risks and ethical concerns should not hamper innovation and development but rather provide new opportunities and stimulate ethically-conducted research and innovation that anchor AI technologies in human rights and fundamental freedoms, values and principles, and moral and ethical reflection,

Also recalling that in November 2019, the General Conference of UNESCO, at its 40th session, adopted 40 C/Resolution 37, by which it mandated the Director-General "to prepare an international standard-setting instrument on the ethics of artificial intelligence (AI) in the form of a recommendation", which is to be submitted to the General Conference at its 41st session in 2021,

Recognizing that the development of AI technologies necessitates a commensurate increase in data, media and information literacy as well as access to independent, pluralistic, trusted sources of information, including as part of efforts to mitigate risks of misinformation, disinformation and hate speech, and harm caused through the misuse of personal data,

Observing that a normative framework for AI technologies and its social implications finds its basis in international and national legal frameworks, human rights and fundamental freedoms, ethics, need for access to data, information and knowledge, the freedom of research and innovation, human and environmental and ecosystem well-being, and connects ethical values and principles to the challenges and opportunities linked to AI technologies, based on common understanding and shared aims,

Also recognizing that ethical values and principles can help develop and implement rights-based policy measures and legal norms, by providing guidance with a view to the fast pace of technological development,

Also convinced that globally accepted ethical standards for AI technologies, in full respect of international law, in particular human rights law, can play a key role in developing AI-related norms across the globe,

Bearing in mind the Universal Declaration of Human Rights (1948), the instruments of the international human rights framework, including the Convention Relating to the Status of Refugees (1951), the Discrimination (Employment and Occupation) Convention (1958), the International Convention on the Elimination of All Forms of Racial Discrimination (1965), the International Covenant on Civil and Political Rights (1966), the International Covenant on Economic, Social and Cultural Rights (1966), the Convention on the Elimination of All Forms of Discrimination against Women (1979), the Convention on the Rights of the Child (1989), and the Convention on the Rights of Persons with Disabilities (2006), the Convention against Discrimination in Education (1960), the Convention on the Protection and Promotion of the Diversity of Cultural Expressions (2005), as well as any other relevant international instruments, recommendations and declarations,

Also noting the United Nations Declaration on the Right to Development (1986); the Declaration on the Responsibilities of the Present Generations Towards Future Generations (1997); the Universal Declaration on Bioethics and Human Rights (2005); the United Nations Declaration on the Rights of Indigenous Peoples (2007); the United Nations General Assembly resolution on the review of the World Summit on the Information Society

(A/RES/70/125) (2015); the United Nations General Assembly Resolution on Transforming our world: the 2030 Agenda for Sustainable Development (A/RES/70/1) (2015); the Recommendation Concerning the Preservation of, and Access to, Documentary Heritage Including in Digital Form (2015); the Declaration of Ethical Principles in relation to Climate Change (2017); the Recommendation on Science and Scientific Researchers (2017); the Internet Universality Indicators (endorsed by UNESCO's International Programme for the Development of Communication in 2018), including the ROAM principles (endorsed by UNESCO's General Conference in 2015); the Human Rights Council's resolution on "The right to privacy in the digital age" (A/HRC/RES/42/15) (2019); and the Human Rights Council's resolution on "New and emerging digital technologies and human rights" (A/HRC/RES/41/11) (2019),

Emphasizing that specific attention must be paid to LMICs, including but not limited to LDCs, LLDCs and SIDS, as they have their own capacity but have been underrepresented in the AI ethics debate, which raises concerns about neglecting local knowledge, cultural pluralism, value systems and the demands of global fairness to deal with the positive and negative impacts of AI technologies,

Also conscious of the many existing national policies, other frameworks and initiatives elaborated by relevant United Nations entities, intergovernmental organizations, including regional organizations, as well as those by the private sector, professional organizations, non-governmental organizations, and the scientific community, related to the ethics and regulation of AI technologies,

Further convinced that AI technologies can bring important benefits, but that achieving them can also amplify tension around innovation, asymmetric access to knowledge and technologies, including the digital and civic literacy deficit that limits the public's

ability to engage in topics related to AI, as well as barriers to access to information and gaps in capacity, human and institutional capacities, barriers to access to technological innovation, and a lack of adequate physical and digital infrastructure and regulatory frameworks, including those related to data, all of which need to be addressed,

Underlining that the strengthening of global cooperation and solidarity, including through multilateralism, is needed to facilitate fair access to AI technologies and address the challenges that they bring to diversity and interconnectivity of cultures and ethical systems, to mitigate potential misuse, to realize the full potential that AI can bring, especially in the area of development, and to ensure that national AI strategies are guided by ethical principles,

Taking fully into account that the rapid development of AI technologies challenges their ethical implementation and governance, as well as the respect for and protection of cultural diversity, and has the potential to disrupt local and regional ethical standards and values,

1. ***Adopts*** the present Recommendation on the Ethics of Artificial Intelligence;

2. ***Recommends*** that Member States apply on a voluntary basis the provisions of this Recommendation by taking appropriate steps, including whatever legislative or other measures may be required, in conformity with the constitutional practice and governing structures of each State, to give effect within their jurisdictions to the principles and norms of the Recommendation in conformity with international law, including international human rights law;

3. ***Also recommends*** that Member States engage all stakeholders, including business enterprises, to ensure that they play their respective roles in the implementation of this Recommendation;

and bring the Recommendation to the attention of the authorities, bodies, research and academic organizations, institutions and organizations in public, private and civil society sectors involved in AI technologies, so that the development and use of AI technologies are guided by both sound scientific research as well as ethical analysis and evaluation.

I. SCOPE OF APPLICATION

1. This Recommendation addresses ethical issues related to the domain of Artificial Intelligence to the extent that they are within UNESCO's mandate. It approaches AI ethics as a systematic normative reflection, based on a holistic, comprehensive, multicultural and evolving framework of interdependent values, principles and actions that can guide societies in dealing responsibly with the known and unknown impacts of AI technologies on human beings, societies and the environment and ecosystems, and offers them a basis to accept or reject AI technologies. It considers ethics as a dynamic basis for the normative evaluation and guidance of AI technologies, referring to human dignity, well-being and the prevention of harm as a compass and as rooted in the ethics of science and technology.

2. This Recommendation does not have the ambition to provide one single definition of AI, since such a definition would need to change over time, in accordance with technological developments. Rather, its ambition is to address those features of AI systems that are of central ethical relevance. Therefore, this Recommendation approaches AI systems as systems which have the capacity to process data and information in a way that resembles intelligent behaviour, and typically includes aspects of reasoning, learning, perception, prediction, planning or control. Three elements have a central place in this approach:

(a) AI systems are information-processing technologies that integrate models and algorithms that produce a capacity to learn and to perform cognitive tasks leading to outcomes such as prediction and decision-making in material and virtual environments. AI systems are designed to operate with varying degrees of autonomy by means of knowledge modelling and representation and by exploiting data and calculating correlations. AI systems may include several methods, such as but not limited to:

(i) machine learning, including deep learning and reinforcement learning;

(ii) machine reasoning, including planning, scheduling, knowledge representation and reasoning, search, and optimization.

AI systems can be used in cyber-physical systems, including the Internet of things, robotic systems, social robotics, and human-computer interfaces, which involve control, perception, the processing of data collected by sensors, and the operation of actuators in the environment in which AI systems work.

(b) Ethical questions regarding AI systems pertain to all stages of the AI system life cycle, understood here to range from research, design and development to deployment and use, including maintenance, operation, trade, financing, monitoring and evaluation, validation, end-of-use, disassembly and termination. In addition, AI actors can be defined as any actor involved in at least one stage of the AI system life cycle, and can refer both to natural and legal persons, such as researchers, programmers, engineers, data scientists, end-users, business enterprises, universities and public and private entities, among others.

(c) AI systems raise new types of ethical issues that include, but are not limited to, their impact on decision-making, employment

and labour, social interaction, health care, education, media, access to information, digital divide, personal data and consumer protection, environment, democracy, rule of law, security and policing, dual use, and human rights and fundamental freedoms, including freedom of expression, privacy and non-discrimination. Furthermore, new ethical challenges are created by the potential of AI algorithms to reproduce and reinforce existing biases, and thus to exacerbate already existing forms of discrimination, prejudice and stereotyping. Some of these issues are related to the capacity of AI systems to perform tasks which previously only living beings could do, and which were in some cases even limited to human beings only. These characteristics give AI systems a profound, new role in human practices and society, as well as in their relationship with the environment and ecosystems, creating a new context for children and young people to grow up in, develop an understanding of the world and themselves, critically understand media and information, and learn to make decisions. In the long term, AI systems could challenge humans' special sense of experience and agency, raising additional concerns about, inter alia, human self-understanding, social, cultural and environmental interaction, autonomy, agency, worth and dignity.

3. This Recommendation pays specific attention to the broader ethical implications of AI systems in relation to the central domains of UNESCO: education, science, culture, and communication and information, as explored in the 2019 Preliminary Study on the Ethics of Artificial

Intelligence by the UNESCO World Commission on Ethics of Scientific Knowledge and Technology (COMEST):

(a) Education, because living in digitalizing societies requires new educational practices, ethical reflection, critical thinking, responsible design practices and new skills, given the implications for the labour market, employability and civic participation.

(b) Science, in the broadest sense and including all academic fields from the natural sciences and medical sciences to the social sciences and humanities, as AI technologies bring new research capacities and approaches, have implications for our concepts of scientific understanding and explanation, and create a new basis for decision-making.

(c) Cultural identity and diversity, as AI technologies can enrich cultural and creative industries, but can also lead to an increased concentration of supply of cultural content, data, markets and income in the hands of only a few actors, with potential negative implications for the diversity and pluralism of languages, media, cultural expressions, participation and equality.

(d) Communication and information, as AI technologies play an increasingly important role in the processing, structuring and provision of information; the issues of automated journalism and the algorithmic provision of news and moderation and curation of content on social media and search engines are just a few examples raising issues related to access to information, disinformation, misinformation, hate speech, the emergence of new forms of societal narratives, discrimination, freedom of expression, privacy and media and information literacy, among others.

4. This Recommendation is addressed to Member States, both as AI actors and as authorities responsible for developing legal and regulatory frameworks throughout the entire AI system life cycle, and for promoting business responsibility. It also provides ethical guidance to all AI actors, including the public and private sectors, by providing a basis for an ethical impact assessment of AI systems throughout their life cycle.

II. AIMS AND OBJECTIVES

5. This Recommendation aims to provide a basis to make AI systems work for the good of humanity, individuals, societies and the

environment and ecosystems, and to prevent harm. It also aims at stimulating the peaceful use of AI systems.

6. In addition to the existing ethical frameworks regarding AI around the world, this Recommendation aims to bring a globally accepted normative instrument that focuses not only on the articulation of values and principles, but also on their practical realization, via concrete policy recommendations, with a strong emphasis on inclusion issues of gender equality and protection of the environment and ecosystems.

7. Because the complexity of the ethical issues surrounding AI necessitates the cooperation of multiple stakeholders across the various levels and sectors of international, regional and national communities, this Recommendation aims to enable stakeholders to take shared responsibility based on a global and intercultural dialogue.

8. The objectives of this Recommendation are:

(a) to provide a universal framework of values, principles and actions to guide States in the formulation of their legislation, policies or other instruments regarding AI, consistent with international law;

(b) to guide the actions of individuals, groups, communities, institutions and private sector companies to ensure the embedding of ethics in all stages of the AI system life cycle;

(c) to protect, promote and respect human rights and fundamental freedoms, human dignity and equality, including gender equality; to safeguard the interests of present and future generations; to preserve the environment, biodiversity and ecosystems; and to respect cultural diversity in all stages of the AI system life cycle;

(d) to foster multi-stakeholder, multidisciplinary and pluralistic dialogue and consensus building about ethical issues relating to AI systems;

(e) to promote equitable access to developments and knowledge in the field of AI and the sharing of benefits, with particular attention to the needs and contributions of LMICs, including LDCs, LLDCs and SIDS.

III. VALUES AND PRINCIPLES

9. The values and principles included below should be respected by all actors in the AI system life cycle, in the first place and, where needed and appropriate, be promoted through amendments to the existing and elaboration of new legislation, regulations and business guidelines. This must comply with international law, including the United Nations Charter and Member States' human rights obligations, and should be in line with internationally agreed social, political, environmental, educational, scientific and economic sustainability objectives, such as the United Nations Sustainable Development Goals (SDGs).

10. Values play a powerful role as motivating ideals in shaping policy measures and legal norms. While the set of values outlined below thus inspires desirable behaviour and represents the foundations of principles, the principles unpack the values underlying them more concretely so that the values can be more easily operationalized in policy statements and actions.

11. While all the values and principles outlined below are desirable per se, in any practical contexts, there may be tensions between these values and principles. In any given situation, a contextual assessment will be necessary to manage potential tensions, taking into account the principle of proportionality and in compliance with human rights and fundamental freedoms. In all cases, any possible limitations on human rights and fundamental freedoms must have a lawful basis, and be reasonable, necessary and proportionate, and consistent with States' obligations under international law. To navigate such scenarios judiciously will typically

require engagement with a broad range of appropriate stakeholders, making use of social dialogue, as well as ethical deliberation, due diligence and impact assessment.

12. The trustworthiness and integrity of the life cycle of AI systems is essential to ensure that AI technologies will work for the good of humanity, individuals, societies and the environment and ecosystems, and embody the values and principles set out in this Recommendation. People should have good reason to trust that AI systems can bring individual and shared benefits, while adequate measures are taken to mitigate risks. An essential requirement for trustworthiness is that, throughout their life cycle, AI systems are subject to thorough monitoring by the relevant stakeholders as appropriate. As trustworthiness is an outcome of the operationalization of the principles in this document, the policy actions proposed in this Recommendation are all directed at promoting trustworthiness in all stages of the AI system life cycle.

III.1 VALUES

Respect, protection and promotion of human rights and fundamental freedoms and human dignity

13. The inviolable and inherent dignity of every human constitutes the foundation for the universal, indivisible, inalienable, interdependent and interrelated system of human rights and fundamental freedoms. Therefore, respect, protection and promotion of human dignity and rights as established by international law, including international human rights law, is essential throughout the life cycle of AI systems. Human dignity relates to the recognition of the intrinsic and equal worth of each individual human being, regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds.

14. No human being or human community should be harmed or subordinated, whether physically, economically, socially, politically, culturally or mentally during any phase of the life cycle of AI systems. Throughout the life cycle of AI systems, the quality of life of human beings should be enhanced, while the definition of "quality of life" should be left open to individuals or groups, as long as there is no violation or abuse of human rights and fundamental freedoms, or the dignity of humans in terms of this definition.

15. Persons may interact with AI systems throughout their life cycle and receive assistance from them, such as care for vulnerable people or people in vulnerable situations, including but not limited to children, older persons, persons with disabilities or the ill. Within such interactions, persons should never be objectified, nor should their dignity be otherwise undermined, or human rights and fundamental freedoms violated or abused.

16. Human rights and fundamental freedoms must be respected, protected and promoted throughout the life cycle of AI systems. Governments, private sector, civil society, international organizations, technical communities and academia must respect human rights instruments and frameworks in their interventions in the processes surrounding the life cycle of AI systems. New technologies need to provide new means to advocate, defend and exercise human rights and not to infringe them.

Environment and ecosystem flourishing

17. Environmental and ecosystem flourishing should be recognized, protected and promoted through the life cycle of AI systems. Furthermore, environment and ecosystems are the existential necessity for humanity and other living beings to be able to enjoy the benefits of advances in AI.

18. All actors involved in the life cycle of AI systems must comply with applicable international law and domestic legislation,

standards and practices, such as precaution, designed for environmental and ecosystem protection and restoration, and sustainable development. They should reduce the environmental impact of AI systems, including but not limited to its carbon footprint, to ensure the minimization of climate change and environmental risk factors, and prevent the unsustainable exploitation, use and transformation of natural resources contributing to the deterioration of the environment and the degradation of ecosystems.

Ensuring diversity and inclusiveness

19. Respect, protection and promotion of diversity and inclusiveness should be ensured throughout the life cycle of AI systems, consistent with international law, including human rights law. This may be done by promoting active participation of all individuals or groups regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds.

20. The scope of lifestyle choices, beliefs, opinions, expressions or personal experiences, including the optional use of AI systems and the co-design of these architectures should not be restricted during any phase of the life cycle of AI systems.

21. Furthermore, efforts, including international cooperation, should be made to overcome, and never take advantage of, the lack of necessary technological infrastructure, education and skills, as well as legal frameworks, particularly in LMICs, LDCs, LLDCs and SIDS, affecting communities.

Living in peaceful, just and interconnected societies

22. AI actors should play a participative and enabling role to ensure peaceful and just societies, which is based on an interconnected future for the benefit of all, consistent with human

rights and fundamental freedoms. The value of living in peaceful and just societies points to the potential of AI systems to contribute throughout their life cycle to the interconnectedness of all living creatures with each other and with the natural environment.

23. The notion of humans being interconnected is based on the knowledge that every human belongs to a greater whole, which thrives when all its constituent parts are enabled to thrive. Living in peaceful, just and interconnected societies requires an organic, immediate, uncalculated bond of solidarity, characterized by a permanent search for peaceful relations, tending towards care for others and the natural environment in the broadest sense of the term.

24. This value demands that peace, inclusiveness and justice, equity and interconnectedness should be promoted throughout the life cycle of AI systems, in so far as the processes of the life cycle of AI systems should not segregate, objectify or undermine freedom and autonomous decision-making as well as the safety of human beings and communities, divide and turn individuals and groups against each other, or threaten the coexistence between humans, other living beings and the natural environment.

III.2 PRINCIPLES Proportionality and Do No Harm

25. It should be recognized that AI technologies do not necessarily, per se, ensure human and environmental and ecosystem flourishing. Furthermore, none of the processes related to the AI system life cycle shall exceed what is necessary to achieve legitimate aims or objectives and should be appropriate to the context. In the event of possible occurrence of any harm to human beings, human rights and fundamental freedoms, communities and society at large or the environment and ecosystems, the implementation of procedures for risk assessment and the adoption of measures in order to preclude the occurrence of such harm should be ensured.

26. The choice to use AI systems and which AI method to use should be justified in the following ways: (a) the AI method chosen should be appropriate and proportional to achieve a given legitimate aim; (b) the AI method chosen should not infringe upon the foundational values captured in this document, in particular, its use must not violate or abuse human rights; and (c) the AI method should be appropriate to the context and should be based on rigorous scientific foundations. In scenarios where decisions are understood to have an impact that is irreversible or difficult to reverse or may involve life and death decisions, final human determination should apply. In particular, AI systems should not be used for social scoring or mass surveillance purposes.

Safety and security

27. Unwanted harms (safety risks), as well as vulnerabilities to attack (security risks) should be avoided and should be addressed, prevented and eliminated throughout the life cycle of AI systems to ensure human, environmental and ecosystem safety and security. Safe and secure AI will be enabled by the development of sustainable, privacy-protective data access frameworks that foster better training and validation of AI models utilizing quality data.

Fairness and non-discrimination

28. AI actors should promote social justice and safeguard fairness and non-discrimination of any kind in compliance with international law. This implies an inclusive approach to ensuring that the benefits of AI technologies are available and accessible to all, taking into consideration the specific needs of different age groups, cultural systems, different language groups, persons with disabilities, girls and women, and disadvantaged, marginalized and vulnerable people or people in vulnerable situations. Member States should work to promote inclusive access for all, including local communities,

to AI systems with locally relevant content and services, and with respect for multilingualism and cultural diversity. Member States should work to tackle digital divides and ensure inclusive access to and participation in the development of AI. At the national level, Member States should promote equity between rural and urban areas, and among all persons regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds, in terms of access to and participation in the AI system life cycle. At the international level, the most technologically advanced countries have a responsibility of solidarity with the least advanced to ensure that the benefits of AI technologies are shared such that access to and participation in the AI system life cycle for the latter contributes to a fairer world order with regard to information, communication, culture, education, research and socio-economic and political stability.

29. AI actors should make all reasonable efforts to minimize and avoid reinforcing or perpetuating discriminatory or biased applications and outcomes throughout the life cycle of the AI system to ensure fairness of such systems. Effective remedy should be available against discrimination and biased algorithmic determination.

30. Furthermore, digital and knowledge divides within and between countries need to be addressed throughout an AI system life cycle, including in terms of access and quality of access to technology and data, in accordance with relevant national, regional and international legal frameworks, as well as in terms of connectivity, knowledge and skills and meaningful participation of the affected communities, such that every person is treated equitably.

Sustainability

31. The development of sustainable societies relies on the achievement of a complex set of objectives on a continuum of human,

social, cultural, economic and environmental dimensions. The advent of AI technologies can either benefit sustainability objectives or hinder their realization, depending on how they are applied across countries with varying levels of development. The continuous assessment of the human, social, cultural, economic and environmental impact of AI technologies should therefore be carried out with full cognizance of the implications of AI technologies for sustainability as a set of constantly evolving goals across a range of dimensions, such as currently identified in the Sustainable Development Goals (SDGs) of the United Nations.

Right to Privacy, and Data Protection

32. Privacy, a right essential to the protection of human dignity, human autonomy and human agency, must be respected, protected and promoted throughout the life cycle of AI systems. It is important that data for AI systems be collected, used, shared, archived and deleted in ways that are consistent with international law and in line with the values and principles set forth in this Recommendation, while respecting relevant national, regional and international legal frameworks.

33. Adequate data protection frameworks and governance mechanisms should be established in a multi-stakeholder approach at the national or international level, protected by judicial systems, and ensured throughout the life cycle of AI systems. Data protection frameworks and any related mechanisms should take reference from international data protection principles and standards concerning the collection, use and disclosure of personal data and exercise of their rights by data subjects while ensuring a legitimate aim and a valid legal basis for the processing of personal data, including informed consent.

34. Algorithmic systems require adequate privacy impact assessments, which also include societal and ethical considerations of

their use and an innovative use of the privacy by design approach. AI actors need to ensure that they are accountable for the design and implementation of AI systems in such a way as to ensure that personal information is protected throughout the life cycle of the AI system.

Human oversight and determination

35. Member States should ensure that it is always possible to attribute ethical and legal responsibility for any stage of the life cycle of AI systems, as well as in cases of remedy related to AI systems, to physical persons or to existing legal entities. Human oversight refers thus not only to individual human oversight, but to inclusive public oversight, as appropriate.

36. It may be the case that sometimes humans would choose to rely on AI systems for reasons of efficacy, but the decision to cede control in limited contexts remains that of humans, as humans can resort to AI systems in decision-making and acting, but an AI system can never replace ultimate human responsibility and accountability. As a rule, life and death decisions should not be ceded to AI systems.

Transparency and explainability

37. The transparency and explainability of AI systems are often essential preconditions to ensure the respect, protection and promotion of human rights, fundamental freedoms and ethical principles. Transparency is necessary for relevant national and international liability regimes to work effectively. A lack of transparency could also undermine the possibility of effectively challenging decisions based on outcomes produced by AI systems and may thereby infringe the right to a fair trial and effective remedy, and limits the areas in which these systems can be legally used.

38. While efforts need to be made to increase transparency and explainability of AI systems, including those with extra-territorial

impact, throughout their life cycle to support democratic governance, the level of transparency and explainability should always be appropriate to the context and impact, as there may be a need to balance between transparency and explainability and other principles such as privacy, safety and security. People should be fully informed when a decision is informed by or is made on the basis of AI algorithms, including when it affects their safety or human rights, and in those circumstances should have the opportunity to request explanatory information from the relevant AI actor or public sector institutions. In addition, individuals should be able to access the reasons for a decision affecting their rights and freedoms, and have the option of making submissions to a designated staff member of the private sector company or public sector institution able to review and correct the decision. AI actors should inform users when a product or service is provided directly or with the assistance of AI systems in a proper and timely manner.

39. From a socio-technical lens, greater transparency contributes to more peaceful, just, democratic and inclusive societies. It allows for public scrutiny that can decrease corruption and discrimination, and can also help detect and prevent negative impacts on human rights. Transparency aims at providing appropriate information to the respective addressees to enable their understanding and foster trust. Specific to the AI system, transparency can enable people to understand how each stage of an AI system is put in place, appropriate to the context and sensitivity of the AI system. It may also include insight into factors that affect a specific prediction or decision, and whether or not appropriate assurances (such as safety or fairness measures) are in place. In cases of serious threats of adverse human rights impacts, transparency may also require the sharing of code or datasets.

40. Explainability refers to making intelligible and providing insight into the outcome of AI systems. The explainability of AI systems also refers to the understandability of the input, output and the functioning of each algorithmic building block and how it contributes to the outcome of the systems. Thus, explainability is closely related to transparency, as outcomes and sub- processes leading to outcomes should aim to be understandable and traceable, appropriate to the context. AI actors should commit to ensuring that the algorithms developed are explainable. In the case of AI applications that impact the end user in a way that is not temporary, easily reversible or otherwise low risk, it should be ensured that the meaningful explanation is provided with any decision that resulted in the action taken in order for the outcome to be considered transparent.

41. Transparency and explainability relate closely to adequate responsibility and accountability measures, as well as to the trustworthiness of AI systems.

Responsibility and accountability

42. AI actors and Member States should respect, protect and promote human rights and fundamental freedoms, and should also promote the protection of the environment and ecosystems, assuming their respective ethical and legal responsibility, in accordance with national and international law, in particular Member States' human rights obligations, and ethical guidance throughout the life cycle of AI systems, including with respect to AI actors within their effective territory and control. The ethical responsibility and liability for the decisions and actions based in any way on an AI system should always ultimately be attributable to AI actors corresponding to their role in the life cycle of the AI system.

43. Appropriate oversight, impact assessment, audit and due diligence mechanisms, including whistle-blowers' protection, should be developed to ensure accountability for AI systems and their impact throughout their life cycle. Both technical and institutional designs should ensure auditability and traceability of (the working of) AI systems in particular to address any conflicts with human rights norms and standards and threats to environmental and ecosystem well- being.

Awareness and literacy

44. Public awareness and understanding of AI technologies and the value of data should be promoted through open and accessible education, civic engagement, digital skills and AI ethics training, media and information literacy and training led jointly by governments, intergovernmental organizations, civil society, academia, the media, community leaders and the private sector, and considering the existing linguistic, social and cultural diversity, to ensure effective public participation so that all members of society can take informed decisions about their use of AI systems and be protected from undue influence.

45. Learning about the impact of AI systems should include learning about, through and for human rights and fundamental freedoms, meaning that the approach and understanding of AI systems should be grounded by their impact on human rights and access to rights, as well as on the environment and ecosystems.

Multi-stakeholder and adaptive governance and collaboration

46. International law and national sovereignty must be respected in the use of data. That means that States, complying with international law, can regulate the data generated within or passing through their territories, and take measures towards effective

regulation of data, including data protection, based on respect for the right to privacy in accordance with international law and other human rights norms and standards.

47. Participation of different stakeholders throughout the AI system life cycle is necessary for inclusive approaches to AI governance, enabling the benefits to be shared by all, and to contribute to sustainable development. Stakeholders include but are not limited to governments, intergovernmental organizations, the technical community, civil society, researchers and academia, media, education, policy-makers, private sector companies, human rights institutions and equality bodies, anti-discrimination monitoring bodies, and groups for youth and children. The adoption of open standards and interoperability to facilitate collaboration should be in place. Measures should be adopted to take into account shifts in technologies, the emergence of new groups of stakeholders, and to allow for meaningful participation by marginalized groups, communities and individuals and, where relevant, in the case of Indigenous Peoples, respect for the self-governance of their data.

IV. AREAS OF POLICY ACTION

48. The policy actions described in the following policy areas operationalize the values and principles set out in this Recommendation. The main action is for Member States to put in place effective measures, including, for example, policy frameworks or mechanisms, and to ensure that other stakeholders, such as private sector companies, academic and research institutions, and civil society adhere to them by, among other actions, encouraging all stakeholders to develop human rights, rule of law, democracy, and ethical impact assessment and due diligence tools in line with guidance including the United Nations Guiding Principles on Business and Human Rights. The process for developing such

policies or mechanisms should be inclusive of all stakeholders and should take into account the circumstances and priorities of each Member State. UNESCO can be a partner and support Member States in the development as well as monitoring and evaluation of policy mechanisms.

49. UNESCO recognizes that Member States will be at different stages of readiness to implement this Recommendation, in terms of scientific, technological, economic, educational, legal, regulatory, infrastructural, societal, cultural and other dimensions. It is noted that "readiness" here is a dynamic status. In order to enable the effective implementation of this Recommendation, UNESCO will therefore: (1) develop a readiness assessment methodology to assist interested Member States in identifying their status at specific moments of their readiness trajectory along a continuum of dimensions; and (2) ensure support for interested Member States in terms of developing a UNESCO methodology for Ethical Impact Assessment (EIA) of AI technologies, sharing of best practices, assessment guidelines and other mechanisms and analytical work.

POLICY AREA 1: ETHICAL IMPACT ASSESSMENT

50. Member States should introduce frameworks for impact assessments, such as ethical impact assessment, to identify and assess benefits, concerns and risks of AI systems, as well as appropriate risk prevention, mitigation and monitoring measures, among other assurance mechanisms. Such impact assessments should identify impacts on human rights and fundamental freedoms, in particular but not limited to the rights of marginalized and vulnerable people or people in vulnerable situations, labour rights, the environment and ecosystems and ethical and social implications, and facilitate citizen participation in line with the values and principles set forth in this Recommendation.

51. Member States and private sector companies should develop due diligence and oversight mechanisms to identify, prevent, mitigate and account for how they address the impact of AI systems on the respect for human rights, rule of law and inclusive societies. Member States should also be able to assess the socio-economic impact of AI systems on poverty and ensure that the gap between people living in wealth and poverty, as well as the digital divide among and within countries, are not increased with the massive adoption of AI technologies at present and in the future. In order to do this, in particular, enforceable transparency protocols should be implemented, corresponding to the access to information, including information of public interest held by private entities. Member States, private sector companies and civil society should investigate the sociological and psychological effects of AI-based recommendations on humans in their decision-making autonomy. AI systems identified as potential risks to human rights should be broadly tested by AI actors, including in real-world conditions if needed, as part of the Ethical Impact Assessment, before releasing them in the market.

52. Member States and business enterprises should implement appropriate measures to monitor all phases of an AI system life cycle, including the functioning of algorithms used for decision- making, the data, as well as AI actors involved in the process, especially in public services and where direct end-user interaction is needed, as part of ethical impact assessment. Member States' human rights law obligations should form part of the ethical aspects of AI system assessments.

53. Governments should adopt a regulatory framework that sets out a procedure, particularly for public authorities, to carry out ethical impact assessments on AI systems to predict consequences, mitigate risks, avoid harmful consequences, facilitate citizen participation and address societal challenges. The assessment

should also establish appropriate oversight mechanisms, including auditability, traceability and explainability, which enable the assessment of algorithms, data and design processes, as well as include external review of AI systems. Ethical impact assessments should be transparent and open to the public, where appropriate. Such assessments should also be multidisciplinary, multi-stakeholder, multicultural, pluralistic and inclusive. The public authorities should be required to monitor the AI systems implemented and/or deployed by those authorities by introducing appropriate mechanisms and tools.

POLICY AREA 2: ETHICAL GOVERNANCE AND STEWARDSHIP

54. Member States should ensure that AI governance mechanisms are inclusive, transparent, multidisciplinary, multilateral (this includes the possibility of mitigation and redress of harm across borders) and multi-stakeholder. In particular, governance should include aspects of anticipation, and effective protection, monitoring of impact, enforcement and redress.

55. Member States should ensure that harms caused through AI systems are investigated and redressed, by enacting strong enforcement mechanisms and remedial actions, to make certain that human rights and fundamental freedoms and the rule of law are respected in the digital world and in the physical world. Such mechanisms and actions should include remediation mechanisms provided by private and public sector companies. The auditability and traceability of AI systems should be promoted to this end. In addition, Member States should strengthen their institutional capacities to deliver on this commitment and should collaborate with researchers and other stakeholders to investigate, prevent and mitigate any potentially malicious uses of AI systems.

56. Member States are encouraged to develop national and regional AI strategies and to consider forms of soft governance such

as a certification mechanism for AI systems and the mutual recognition of their certification, according to the sensitivity of the application domain and expected impact on human rights, the environment and ecosystems, and other ethical considerations set forth in this Recommendation. Such a mechanism might include different levels of audit of systems, data, and adherence to ethical guidelines and to procedural requirements in view of ethical aspects. At the same time, such a mechanism should not hinder innovation or disadvantage small and medium enterprises or start-ups, civil society as well as research and science organizations, as a result of an excessive administrative burden. These mechanisms should also include a regular monitoring component to ensure system robustness and continued integrity and adherence to ethical guidelines over the entire life cycle of the AI system, requiring re-certification if necessary.

57. Member States and public authorities should carry out transparent self-assessment of existing and proposed AI systems, which, in particular, should include the assessment of whether the adoption of AI is appropriate and, if so, should include further assessment to determine what the appropriate method is, as well as assessment as to whether such adoption would result in violations or abuses of Member States' human rights law obligations, and if that is the case, prohibit its use.

58. Member States should encourage public entities, private sector companies and civil society organizations to involve different stakeholders in their AI governance and to consider adding the role of an independent AI Ethics Officer or some other mechanism to oversee ethical impact assessment, auditing and continuous monitoring efforts and ensure ethical guidance of AI systems. Member States, private sector companies and civil society organizations, with the support of UNESCO, are encouraged to create a network of independent AI Ethics Officers to give support to this process at national, regional and international levels.

59. Member States should foster the development of, and access to, a digital ecosystem for ethical and inclusive development of AI systems at the national level, including to address gaps in access to the AI system life cycle, while contributing to international collaboration. Such an ecosystem includes, in particular, digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate.

60. Member States should establish mechanisms, in collaboration with international organizations, transnational corporations, academic institutions and civil society, to ensure the active participation of all Member States, especially LMICs, in particular LDCs, LLDCs and SIDS, in international discussions concerning AI governance. This can be through the provision of funds, ensuring equal regional participation, or any other mechanisms. Furthermore, in order to ensure the inclusiveness of AI fora, Member States should facilitate the travel of AI actors in and out of their territory, especially from LMICs, in particular LDCs, LLDCs and SIDS, for the purpose of participating in these fora.

61. Amendments to the existing or elaboration of new national legislation addressing AI systems must comply with Member States' human rights law obligations and promote human rights and fundamental freedoms throughout the AI system life cycle. Promotion thereof should also take the form of governance initiatives, good exemplars of collaborative practices regarding AI systems, and national and international technical and methodological guidelines as AI technologies advance. Diverse sectors, including the private sector, in their practices regarding AI systems must respect, protect and promote human rights and fundamental freedoms using existing and new instruments in combination with this Recommendation.

62. Member States that acquire AI systems for human rights-sensitive use cases, such as law enforcement, welfare, employment,

media and information providers, health care and the independent judiciary system should provide mechanisms to monitor the social and economic impact of such systems by appropriate oversight authorities, including independent data protection authorities, sectoral oversight and public bodies responsible for oversight.

63. Member States should enhance the capacity of the judiciary to make decisions related to AI systems as per the rule of law and in line with international law and standards, including in the use of AI systems in their deliberations, while ensuring that the principle of human oversight is upheld. In case AI systems are used by the judiciary, sufficient safeguards are needed to guarantee inter alia the protection of fundamental human rights, the rule of law, judicial independence as well as the principle of human oversight, and to ensure a trustworthy, public interest-oriented and human-centric development and use of AI systems in the judiciary.

64. Member States should ensure that governments and multilateral organizations play a leading role in ensuring the safety and security of AI systems, with multi-stakeholder participation. Specifically, Member States, international organizations and other relevant bodies should develop international standards that describe measurable, testable levels of safety and transparency, so that systems can be objectively assessed and levels of compliance determined. Furthermore, Member States and business enterprises should continuously support strategic research on potential safety and security risks of AI technologies and should encourage research into transparency and explainability, inclusion and literacy by putting additional funding into those areas for different domains and at different levels, such as technical and natural language.

65. Member States should implement policies to ensure that the actions of AI actors are consistent with international human rights law, standards and principles throughout the life cycle of AI systems, while

taking into full consideration the current cultural and social diversities, including local customs and religious traditions, with due regard to the precedence and universality of human rights.

66. Member States should put in place mechanisms to require AI actors to disclose and combat any kind of stereotyping in the outcomes of AI systems and data, whether by design or by negligence, and to ensure that training data sets for AI systems do not foster cultural, economic or social inequalities, prejudice, the spreading of disinformation and misinformation, and disruption of freedom of expression and access to information. Particular attention should be given to regions where the data are scarce.

67. Member States should implement policies to promote and increase diversity and inclusiveness that reflect their populations in AI development teams and training datasets, and to ensure equal access to AI technologies and their benefits, particularly for marginalized groups, both from rural and urban zones.

68. Member States should develop, review and adapt, as appropriate, regulatory frameworks to achieve accountability and responsibility for the content and outcomes of AI systems at the different phases of their life cycle. Member States should, where necessary, introduce liability frameworks or clarify the interpretation of existing frameworks to ensure the attribution of accountability for the outcomes and the functioning of AI systems. Furthermore, when developing regulatory frameworks, Member States should, in particular, take into account that ultimate responsibility and accountability must always lie with natural or legal persons and that AI systems should not be given legal personality themselves. To ensure this, such regulatory frameworks should be consistent with the principle of human oversight and establish a comprehensive approach focused on AI actors and the technological processes involved across the different stages of the AI system life cycle.

69. In order to establish norms where these do not exist, or to adapt the existing legal frameworks, Member States should involve all AI actors (including, but not limited to, researchers, representatives of civil society and law enforcement, insurers, investors, manufacturers, engineers, lawyers and users). The norms can mature into best practices, laws and regulations. Member States are further encouraged to use mechanisms such as policy prototypes and regulatory sandboxes to accelerate the development of laws, regulations and policies, including regular reviews thereof, in line with the rapid development of new technologies and ensure that laws and regulations can be tested in a safe environment before being officially adopted. Member States should support local governments in the development of local policies, regulations and laws in line with national and international legal frameworks.

70. Member States should set clear requirements for AI system transparency and explainability so as to help ensure the trustworthiness of the full AI system life cycle. Such requirements should involve the design and implementation of impact mechanisms that take into consideration the nature of application domain, intended use, target audience and feasibility of each particular AI system.

POLICY AREA 3: DATA POLICY

71. Member States should work to develop data governance strategies that ensure the continual evaluation of the quality of training data for AI systems including the adequacy of the data collection and selection processes, proper data security and protection measures, as well as feedback mechanisms to learn from mistakes and share best practices among all AI actors.

72. Member States should put in place appropriate safeguards to protect the right to privacy in accordance with international law, including addressing concerns such as surveillance. Member States

should, among others, adopt or enforce legislative frameworks that provide appropriate protection, compliant with international law. Member States should strongly encourage all AI actors, including business enterprises, to follow existing international standards and, in particular, to carry out adequate privacy impact assessments, as part of ethical impact assessments, which take into account the wider socio-economic impact of the intended data processing, and to apply privacy by design in their systems. Privacy should be respected, protected and promoted throughout the life cycle of AI systems.

73. Member States should ensure that individuals retain rights over their personal data and are protected by a framework, which notably foresees: transparency; appropriate safeguards for the processing of sensitive data; an appropriate level of data protection; effective and meaningful accountability schemes and mechanisms; the full enjoyment of the data subjects' rights and the ability to access and erase their personal data in AI systems, except for certain circumstances in compliance with international law; an appropriate level of protection in full compliance with data protection legislation where data are being used for commercial purposes such as enabling micro-targeted advertising, transferred cross-border; and an effective independent oversight as part of a data governance mechanism which keeps individuals in control of their personal data and fosters the benefits of a free flow of information internationally, including access to data.

74. Member States should establish their data policies or equivalent frameworks, or reinforce existing ones, to ensure full security for personal data and sensitive data, which, if disclosed, may cause exceptional damage, injury or hardship to individuals. Examples include data relating to offences, criminal proceedings and convictions, and related security measures; biometric, genetic and health data; and -personal data such as that relating to race, colour,

descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other characteristics.

75. Member States should promote open data. In this regard, Member States should consider reviewing their policies and regulatory frameworks, including on access to information and open government to reflect AI-specific requirements and promoting mechanisms, such as open repositories for publicly funded or publicly held data and source code and data trusts, to support the safe, fair, legal and ethical sharing of data, among others.

76. Member States should promote and facilitate the use of quality and robust datasets for training, development and use of AI systems, and exercise vigilance in overseeing their collection and use. This could, if possible and feasible, include investing in the creation of gold standard datasets, including open and trustworthy datasets, which are diverse, constructed on a valid legal basis, including consent of data subjects, when required by law. Standards for annotating datasets should be encouraged, including disaggregating data on gender and other bases, so it can easily be determined how a dataset is gathered and what properties it has.

77. Member States, as also suggested in the report of the United Nations Secretary-General's High-level Panel on Digital Cooperation, with the support of the United Nations and UNESCO, should adopt a digital commons approach to data where appropriate, increase interoperability of tools and datasets and interfaces of systems hosting data, and encourage private sector companies to share the data they collect with all stakeholders, as appropriate, for research, innovation or public benefits. They should also promote public and private efforts to create collaborative platforms to share quality data in trusted and secured data spaces.

POLICY AREA 4: DEVELOPMENT AND INTERNATIONAL COOPERATION

78. Member States and transnational corporations should prioritize AI ethics by including discussions of AI-related ethical issues into relevant international, intergovernmental and multi-stakeholder fora.

79. Member States should ensure that the use of AI in areas of development such as education, science, culture, communication and information, health care, agriculture and food supply, environment, natural resource and infrastructure management, economic planning and growth, among others, adheres to the values and principles set forth in this Recommendation.

80. Member States should work through international organizations to provide platforms for international cooperation on AI for development, including by contributing expertise, funding, data, domain knowledge, infrastructure, and facilitating multi-stakeholder collaboration to tackle challenging development problems, especially for LMICs, in particular LDCs, LLDCs and SIDS.

81. Member States should work to promote international collaboration on AI research and innovation, including research and innovation centres and networks that promote greater participation and leadership of researchers from LMICs and other countries, including LDCs, LLDCs and SIDS.

82. Member States should promote AI ethics research by engaging international organizations and research institutions, as well as transnational corporations, that can be a basis for the ethical use of AI systems by public and private entities, including research into the applicability of specific ethical frameworks in specific cultures and contexts, and the possibilities to develop technologically feasible solutions in line with these frameworks.

83. Member States should encourage international cooperation and collaboration in the field of AI to bridge geo-technological lines. Technological exchanges and consultations should take place between Member States and their populations, between the public and private sectors, and between and among the most and least technologically advanced countries in full respect of international law.

POLICY AREA 5: ENVIRONMENT AND ECOSYSTEMS

84. Member States and business enterprises should assess the direct and indirect environmental impact throughout the AI system life cycle, including, but not limited to, its carbon footprint, energy consumption and the environmental impact of raw material extraction for supporting the manufacturing of AI technologies, and reduce the environmental impact of AI systems and data infrastructures. Member States should ensure compliance of all AI actors with environmental law, policies and practices.

85. Member States should introduce incentives, when needed and appropriate, to ensure the development and adoption of rights-based and ethical AI-powered solutions for disaster risk resilience; the monitoring, protection and regeneration of the environment and ecosystems; and the preservation of the planet. These AI systems should involve the participation of local and indigenous communities throughout the life cycle of AI systems and should support circular economy type approaches and sustainable consumption and production patterns. Some examples include using AI systems, when needed and appropriate, to:

(a) Support the protection, monitoring and management of natural resources.

(b) Support the prediction, prevention, control and mitigation of climate-related problems.

(c) Support a more efficient and sustainable food ecosystem.

(d) Support the acceleration of access to and mass adoption of sustainable energy.

(e) Enable and promote the mainstreaming of sustainable infrastructure, sustainable business models and sustainable finance for sustainable development.

(f) Detect pollutants or predict levels of pollution and thus help relevant stakeholders identify, plan and put in place targeted interventions to prevent and reduce pollution and exposure.

86. When choosing AI methods, given the potential data-intensive or resource-intensive character of some of them and the respective impact on the environment, Member States should ensure that AI actors, in line with the principle of proportionality, favour data, energy and resource - efficient AI methods. Requirements should be developed to ensure that appropriate evidence is available to show that an AI application will have the intended effect, or that safeguards accompanying an AI application can support the justification for its use. If this cannot be done, the precautionary principle must be favoured, and in instances where there are disproportionate negative impacts on the environment, AI should not be used.

POLICY AREA 6: GENDER

87. Member States should ensure that the potential for digital technologies and artificial intelligence to contribute to achieving gender equality is fully maximized, and must ensure that the human rights and fundamental freedoms of girls and women, and their safety and integrity are not violated at any stage of the AI system life cycle. Moreover, Ethical Impact Assessment should include a transversal gender perspective.

88. Member States should have dedicated funds from their public budgets linked to financing gender-responsive schemes,

ensure that national digital policies include a gender action plan, and develop relevant policies, for example, on labour education, targeted at supporting girls and women to make sure they are not left out of the digital economy powered by AI. Special investment in providing targeted programmes and gender-specific language, to increase the opportunities of girls' and women's participation in science, technology, engineering, and mathematics (STEM), including information and communication technologies (ICT) disciplines, preparedness, employability, equal career development and professional growth of girls and women, should be considered and implemented.

89. Member States should ensure that the potential of AI systems to advance the achievement of gender equality is realized. They should ensure that these technologies do not exacerbate the already wide gender gaps existing in several fields in the analogue world, and instead eliminate those gaps. These gaps include: the gender wage gap; the unequal representation in certain professions and activities; the lack of representation at top management positions, boards of directors, or research teams in the AI field; the education gap; the digital and AI access, adoption, usage and affordability gap; and the unequal distribution of unpaid work and of the caring responsibilities in our societies.

90. Member States should ensure that gender stereotyping and discriminatory biases are not translated into AI systems, and instead identify and proactively redress these. Efforts are necessary to avoid the compounding negative effect of technological divides in achieving gender equality and avoiding violence such as harassment, bullying or trafficking of girls and women and under-represented groups, including in the online domain.

91. Member States should encourage female entrepreneurship, participation and engagement in all stages of an AI system life cycle

by offering and promoting economic, regulatory incentives, among other incentives and support schemes, as well as policies that aim at a balanced gender participation in AI research in academia, gender representation on digital and AI companies' top management positions, boards of directors and research teams. Member States should ensure that public funds (for innovation, research and technologies) are channelled to inclusive programmes and companies, with clear gender representation, and that private funds are similarly encouraged through affirmative action principles. Policies on harassment-free environments should be developed and enforced, together with the encouragement of the transfer of best practices on how to promote diversity throughout the AI system life cycle.

92. Member States should promote gender diversity in AI research in academia and industry by offering incentives to girls and women to enter the field, putting in place mechanisms to fight gender stereotyping and harassment within the AI research community, and encouraging academic and private entities to share best practices on how to enhance gender diversity.

93. UNESCO can help form a repository of best practices for incentivizing the participation of girls, women and under-represented groups in all stages of the AI system life cycle.

POLICY AREA 7: CULTURE

94. Member States are encouraged to incorporate AI systems, where appropriate, in the preservation, enrichment, understanding, promotion, management and accessibility of tangible, documentary and intangible cultural heritage, including endangered languages as well as indigenous languages and knowledges, for example by introducing or updating educational programmes related to the application of AI systems in these areas, where appropriate, and by ensuring a participatory approach, targeted at institutions and the public.

95. Member States are encouraged to examine and address the cultural impact of AI systems, especially natural language processing (NLP) applications such as automated translation and voice assistants, on the nuances of human language and expression. Such assessments should provide input for the design and implementation of strategies that maximize the benefits from these systems by bridging cultural gaps and increasing human understanding, as well as addressing the negative implications such as the reduction of use, which could lead to the disappearance of endangered languages, local dialects, and tonal and cultural variations associated with human language and expression.

96. Member States should promote AI education and digital training for artists and creative professionals to assess the suitability of AI technologies for use in their profession, and contribute to the design and implementation of suitable AI technologies, as AI technologies are being used to create, produce, distribute, broadcast and consume a variety of cultural goods and services, bearing in mind the importance of preserving cultural heritage, diversity and artistic freedom.

97. Member States should promote awareness and evaluation of AI tools among local cultural industries and small and medium enterprises working in the field of culture, to avoid the risk of concentration in the cultural market.

98. Member States should engage technology companies and other stakeholders to promote a diverse supply of and plural access to cultural expressions, and in particular to ensure that algorithmic recommendation enhances the visibility and discoverability of local content.

99. Member States should foster new research at the intersection between AI and intellectual property (IP), for example to

determine whether or how to protect with IP rights the works created by means of Al technologies. Member States should also assess how AI technologies are affecting the rights or interests of IP owners, whose works are used to research, develop, train or implement AI applications.

100. Member States should encourage museums, galleries, libraries and archives at the national level to use AI systems to highlight their collections and enhance their libraries, databases and knowledge base, while also providing access to their users.

POLICY AREA 8: EDUCATION AND RESEARCH

101. Member States should work with international organizations, educational institutions and private and non-governmental entities to provide adequate AI literacy education to the public on all levels in all countries in order to empower people and reduce the digital divides and digital access inequalities resulting from the wide adoption of AI systems.

102. Member States should promote the acquisition of “prerequisite skills” for AI education, such as basic literacy, numeracy, coding and digital skills, and media and information literacy, as well as critical and creative thinking, teamwork, communication, socio-emotional and AI ethics skills, especially in countries and in regions or areas within countries where there are notable gaps in the education of these skills.

103. Member States should promote general awareness programmes about AI developments, including on data and the opportunities and challenges brought about by AI technologies, the impact of AI systems on human rights and their implications, including children’s rights. These programmes should be accessible to non-technical as well as technical groups.

104. Member States should encourage research initiatives on the responsible and ethical use of AI technologies in teaching, teacher training and e-learning, among other issues, to enhance opportunities and mitigate the challenges and risks involved in this area. The initiatives should be accompanied by an adequate assessment of the quality of education and impact on students and teachers of the use of AI technologies. Member States should also ensure that AI technologies empower students and teachers and enhance their experience, bearing in mind that relational and social aspects and the value of traditional forms of education are vital in teacher-student and student-student relationships and should be considered when discussing the adoption of AI technologies in education. AI systems used in learning should be subject to strict requirements when it comes to the monitoring, assessment of abilities, or prediction of the learners' behaviours. AI should support the learning process without reducing cognitive abilities and without extracting sensitive information, in compliance with relevant personal data protection standards. The data handed over to acquire knowledge collected during the learner's interactions with the AI system must not be subject to misuse, misappropriation or criminal exploitation, including for commercial purposes.

105. Member States should promote the participation and leadership of girls and women, diverse ethnicities and cultures, persons with disabilities, marginalized and vulnerable people or people in vulnerable situations, minorities and all persons not enjoying the full benefits of digital inclusion, in AI education programmes at all levels, as well as the monitoring and sharing of best practices in this regard with other Member States.

106. Member States should develop, in accordance with their national education programmes and traditions, AI ethics curricula for all levels, and promote cross-collaboration between AI technical skills

education and humanistic, ethical and social aspects of AI education. Online courses and digital resources of AI ethics education should be developed in local languages, including indigenous languages, and take into account the diversity of environments, especially ensuring accessibility of formats for persons with disabilities.

107. Member States should promote and support AI research, notably AI ethics research, including for example through investing in such research or by creating incentives for the public and private sectors to invest in this area, recognizing that research contributes significantly to the further development and improvement of AI technologies with a view to promoting international law and the values and principles set forth in this Recommendation. Member States should also publicly promote the best practices of, and cooperation with, researchers and companies who develop AI in an ethical manner.

108. Member States should ensure that AI researchers are trained in research ethics and require them to include ethical considerations in their designs, products and publications, especially in the analyses of the datasets they use, how they are annotated, and the quality and scope of the results with possible applications.

109. Member States should encourage private sector companies to facilitate the access of the scientific community to their data for research, especially in LMICs, in particular LDCs, LLDCs and SIDS. This access should conform to relevant privacy and data protection standards.

110. To ensure a critical evaluation of AI research and proper monitoring of potential misuses or adverse effects, Member States should ensure that any future developments with regards to AI technologies should be based on rigorous and independent scientific research, and promote interdisciplinary AI research by including

disciplines other than science, technology, engineering and mathematics (STEM), such as cultural studies, education, ethics, international relations, law, linguistics, philosophy, political science, sociology and psychology.

111. Recognizing that AI technologies present great opportunities to help advance scientific knowledge and practice, especially in traditionally model-driven disciplines, Member States should encourage scientific communities to be aware of the benefits, limits and risks of their use; this includes attempting to ensure that conclusions drawn from data-driven approaches, models and treatments are robust and sound. Furthermore, Member States should welcome and support the role of the scientific community in contributing to policy and in cultivating awareness of the strengths and weaknesses of AI technologies.

POLICY AREA 9: COMMUNICATION AND INFORMATION

112. Member States should use AI systems to improve access to information and knowledge. This can include support to researchers, academia, journalists, the general public and developers, to enhance freedom of expression, academic and scientific freedoms, access to information, and increased proactive disclosure of official data and information.

113. Member States should ensure that AI actors respect and promote freedom of expression as well as access to information with regard to automated content generation, moderation and curation. Appropriate frameworks, including regulation, should enable transparency of online communication and information operators and ensure users have access to a diversity of viewpoints, as well as processes for prompt notification to the users on the reasons for removal or other treatment of content, and appeal mechanisms that allow users to seek redress.

114. Member States should invest in and promote digital and media and information literacy skills to strengthen critical thinking and competencies needed to understand the use and implication of AI systems, in order to mitigate and counter disinformation, misinformation and hate speech. A better understanding and evaluation of both the positive and potentially harmful effects of recommender systems should be part of those efforts.

115. Member States should create enabling environments for media to have the rights and resources to effectively report on the benefits and harms of AI systems, and also encourage media to make ethical use of AI systems in their operations.

POLICY AREA 10: ECONOMY AND LABOUR

116. Member States should assess and address the impact of AI systems on labour markets and its implications for education requirements, in all countries and with special emphasis on countries where the economy is labour-intensive. This can include the introduction of a wider range of “core” and interdisciplinary skills at all education levels to provide current workers and new generations a fair chance of finding jobs in a rapidly changing market, and to ensure their awareness of the ethical aspects of AI systems. Skills such as “learning how to learn”, communication, critical thinking, teamwork, empathy, and the ability to transfer one’s knowledge across domains, should be taught alongside specialist, technical skills, as well as low-skilled tasks. Being transparent about what skills are in demand and updating curricula around these are key.

117. Member States should support collaboration agreements among governments, academic institutions, vocational education and training institutions, industry, workers’ organizations and civil society to bridge the gap of skillset requirements to align training programmes and strategies with the implications of the future of work

and the needs of industry, including small and medium enterprises. Project-based teaching and learning approaches for AI should be promoted, allowing for partnerships between public institutions, private sector companies, universities and research centres.

118. Member States should work with private sector companies, civil society organizations and other stakeholders, including workers and unions to ensure a fair transition for at-risk employees. This includes putting in place upskilling and reskilling programmes, finding effective mechanisms of retaining employees during those transition periods, and exploring "safety net" programmes for those who cannot be retrained. Member States should develop and implement programmes to research and address the challenges identified that could include upskilling and reskilling, enhanced social protection, proactive industry policies and interventions, tax benefits, new taxation forms, among others. Member States should ensure that there is sufficient public funding to support these programmes. Relevant regulations, such as tax regimes, should be carefully examined and changed if needed to counteract the consequences of unemployment caused by AI-based automation.

119. Member States should encourage and support researchers to analyse the impact of AI systems on the local labour environment in order to anticipate future trends and challenges. These studies should have an interdisciplinary approach and investigate the impact of AI systems on economic, social and geographic sectors, as well as on human-robot interactions and human- human relationships, in order to advise on reskilling and redeployment best practices.

120. Member States should take appropriate steps to ensure competitive markets and consumer protection, considering possible measures and mechanisms at national, regional and international levels, to prevent abuse of dominant market positions, including by

monopolies, in relation to AI systems throughout their life cycle, whether these are data, research, technology, or market. Member States should prevent the resulting inequalities, assess relevant markets and promote competitive markets. Due consideration should be given to LMICs, in particular LDCs, LLDCs and SIDS, which are more exposed and vulnerable to the possibility of abuses of market dominance as a result of a lack of infrastructure, human capacity and regulations, among other factors. AI actors developing AI systems in countries which have established or adopted ethical standards on AI should respect these standards when exporting these products, developing or applying their AI systems in countries where such standards may not exist, while respecting applicable international law and domestic legislation, standards and practices of these countries.

POLICY AREA 11: HEALTH AND SOCIAL WELL-BEING

121. Member States should endeavour to employ effective AI systems for improving human health and protecting the right to life, including mitigating disease outbreaks, while building and maintaining international solidarity to tackle global health risks and uncertainties, and ensure that their deployment of AI systems in health care be consistent with international law and their human rights law obligations. Member States should ensure that actors involved in health care AI systems take into consideration the importance of a patient's relationships with their family and with health care staff.

122. Member States should ensure that the development and deployment of AI systems related to health in general and mental health in particular, paying due attention to children and youth, is regulated to the effect that they are safe, effective, efficient, scientifically and medically proven and enable evidence-based innovation and medical progress. Moreover, in the related area of digital health interventions, Member States are strongly encouraged to

actively involve patients and their representatives in all relevant steps of the development of the system.

123. Member States should pay particular attention in regulating prediction, detection and treatment solutions for health care in AI applications by:

(a) ensuring oversight to minimize and mitigate bias;

(b) ensuring that the professional, the patient, caregiver or service user is included as a “domain expert” in the team in all relevant steps when developing the algorithms;

(c) paying due attention to privacy because of the potential need for being medically monitored and ensuring that all relevant national and international data protection requirements are met;

(d) ensuring effective mechanisms so that those whose personal data is being analysed are aware of and provide informed consent for the use and analysis of their data, without preventing access to health care;

(e) ensuring the human care and final decision of diagnosis and treatment are taken always by humans while acknowledging that AI systems can also assist in their work;

(f) ensuring, where necessary, the review of AI systems by an ethical research committee prior to clinical use.

124. Member States should establish research on the effects and regulation of potential harms to mental health related to AI systems, such as higher degrees of depression, anxiety, social isolation, developing addiction, trafficking, radicalization and misinformation, among others.

125. Member States should develop guidelines for human-robot interactions and their impact on human-human relationships,

based on research and directed at the future development of robots, and with special attention to the mental and physical health of human beings. Particular attention should be given to the use of robots in health care and the care for older persons and persons with disabilities, in education, and robots for use by children, toy robots, chatbots and companion robots for children and adults. Furthermore, assistance of AI technologies should be applied to increase the safety and ergonomic use of robots, including in a human-robot working environment. Special attention should be paid to the possibility of using AI to manipulate and abuse human cognitive biases.

126. Member States should ensure that human-robot interactions comply with the same values and principles that apply to any other AI systems, including human rights and fundamental freedoms, the promotion of diversity, and the protection of vulnerable people or people in vulnerable situations. Ethical questions related to AI-powered systems for neurotechnologies and brain-computer interfaces should be considered in order to preserve human dignity and autonomy.

127. Member States should ensure that users can easily identify whether they are interacting with a living being, or with an AI system imitating human or animal characteristics, and can effectively refuse such interaction and request human intervention.

128. Member States should implement policies to raise awareness about the anthropomorphization of AI technologies and technologies that recognize and mimic human emotions, including in the language used to mention them, and assess the manifestations, ethical implications and possible limitations of such anthropomorphization, in particular in the context of robot-human interaction and especially when children are involved.

129. Member States should encourage and promote collaborative research into the effects of long- term interaction of people with AI systems, paying particular attention to the psychological and cognitive impact that these systems can have on children and young people. This should be done using multiple norms, principles, protocols, disciplinary approaches, and assessment of the modification of behaviours and habits, as well as careful evaluation of the downstream cultural and societal impacts. Furthermore, Member States should encourage research on the effect of AI technologies on health system performance and health outcomes.

130. Member States, as well as all stakeholders, should put in place mechanisms to meaningfully engage children and young people in conversations, debates and decision-making with regard to the impact of AI systems on their lives and futures.

V. MONITORING AND EVALUATION

131. Member States should, according to their specific conditions, governing structures and constitutional provisions, credibly and transparently monitor and evaluate policies, programmes and mechanisms related to ethics of AI, using a combination of quantitative and qualitative approaches. To support Member States, UNESCO can contribute by:

(a) developing a UNESCO methodology for Ethical Impact Assessment (EIA) of AI technologies based on rigorous scientific research and grounded in international human rights law, guidance for its implementation in all stages of the AI system life cycle, and capacity-building materials to support Member States' efforts to train government officials, policy-makers and other relevant AI actors on EIA methodology;

(b) developing a UNESCO readiness assessment methodology to assist Member States in identifying their status at specific

moments of their readiness trajectory along a continuum of dimensions;

(c) developing a UNESCO methodology to evaluate ex ante and ex post the effectiveness and efficiency of the policies for AI ethics and incentives against defined objectives;

(d) strengthening the research- and evidence-based analysis of and reporting on policies regarding AI ethics;

(e) collecting and disseminating progress, innovations, research reports, scientific publications, data and statistics regarding policies for AI ethics, including through existing initiatives, to support sharing best practices and mutual learning, and to advance the implementation of this Recommendation.

132. Processes for monitoring and evaluation should ensure broad participation of all stakeholders, including, but not limited to, vulnerable people or people in vulnerable situations. Social, cultural and gender diversity should be ensured, with a view to improving learning processes and strengthening the connections between findings, decision-making, transparency and accountability for results.

133. In the interests of promoting best policies and practices related to ethics of AI, appropriate tools and indicators should be developed for assessing the effectiveness and efficiency thereof against agreed standards, priorities and targets, including specific targets for persons belonging to disadvantaged, marginalized populations, and vulnerable people or people in vulnerable situations, as well as the impact of AI systems at individual and societal levels. The monitoring and assessment of the impact of AI systems and related AI ethics policies and practices should be carried out continuously in a systematic way proportionate to the relevant risks. This should be based on internationally agreed frameworks and involve evaluations of private and public institutions, providers and

programmes, including self-evaluations, as well as tracer studies and the development of sets of indicators. Data collection and processing should be conducted in accordance with international law, national legislation on data protection and data privacy, and the values and principles outlined in this Recommendation.

134. In particular, Member States may wish to consider possible mechanisms for monitoring and evaluation, such as an ethics commission, AI ethics observatory, repository covering human rights-compliant and ethical development of AI systems, or contributions to existing initiatives by addressing adherence to ethical principles across UNESCO's areas of competence, an experience-sharing mechanism, AI regulatory sandboxes, and an assessment guide for all AI actors to evaluate their adherence to policy recommendations mentioned in this document.

VI. UTILIZATION AND EXPLOITATION OF THE PRESENT RECOMMENDATION

135. Member States and all other stakeholders as identified in this Recommendation should respect, promote and protect the ethical values, principles and standards regarding AI that are identified in this Recommendation, and should take all feasible steps to give effect to its policy recommendations.

136. Member States should strive to extend and complement their own action in respect of this Recommendation, by cooperating with all relevant national and international governmental and non-governmental organizations, as well as transnational corporations and scientific organizations, whose activities fall within the scope and objectives of this Recommendation. The development of a UNESCO Ethical Impact Assessment methodology and the establishment of national commissions for the ethics of AI can be important instruments for this.

VII. PROMOTION OF THE PRESENT RECOMMENDATION

137. UNESCO has the vocation to be the principal United Nations agency to promote and disseminate this Recommendation, and accordingly will work in collaboration with other relevant United Nations entities, while respecting their mandate and avoiding duplication of work.

138. UNESCO, including its bodies, such as the World Commission on the Ethics of Scientific Knowledge and Technology (COMEST), the International Bioethics Committee (IBC) and the Intergovernmental Bioethics Committee (IGBC), will also work in collaboration with other international, regional and sub-regional governmental and non-governmental organizations.

139. Even though, within UNESCO, the mandate to promote and protect falls within the authority of governments and intergovernmental bodies, civil society will be an important actor to advocate for the public sector's interests and therefore UNESCO needs to ensure and promote its legitimacy.

VIII. FINAL PROVISIONS

140. This Recommendation needs to be understood as a whole, and the foundational values and principles are to be understood as complementary and interrelated.

141. Nothing in this Recommendation may be interpreted as replacing, altering or otherwise prejudicing States' obligations or rights under international law, or as approval for any State, other political, economic or social actor, group or person to engage in any activity or perform any act contrary to human rights, fundamental freedoms, human dignity and concern for the environment and ecosystems, both living and non-living.

NHÀ XUẤT BẢN HỒNG ĐỨC
Địa chỉ: 65 Tràng Thi - Hoàn Kiếm - Hà Nội

KỶ YẾU HỘI THẢO
CONFERENCE PROCEEDINGS

Đạo đức trong thời đại Trí tuệ Nhân tạo
Bioethics in the era of Artificial Intelligence (AI)

Chịu trách nhiệm xuất bản
Giám đốc
BÙI VIỆT BẮC
Chịu trách nhiệm nội dung
Tổng biên tập
LÝ BÁ TOÀN
Biên tập
ThS. Nguyễn Khắc Oánh

Trình bày, minh họa: Mai Phương
Sửa bản in: Đức Hiếu

In 250 cuốn, khổ 16x24 cm tại Công ty TNHH DVTM và Sản xuất Thiên Thành
Địa chỉ: Số 15 ngõ 43 Tôn Thất Thiệp, Ba Đình, Hà Nội
Số xuất bản: 215 - 2022CXBIPH/13 - 05/HĐ
Quyết định xuất bản số: 194/QĐ-NXBHĐ ngày 25/01/2022
ISBN: 978-604-351-825-2
In xong và nộp lưu chiểu năm 2022