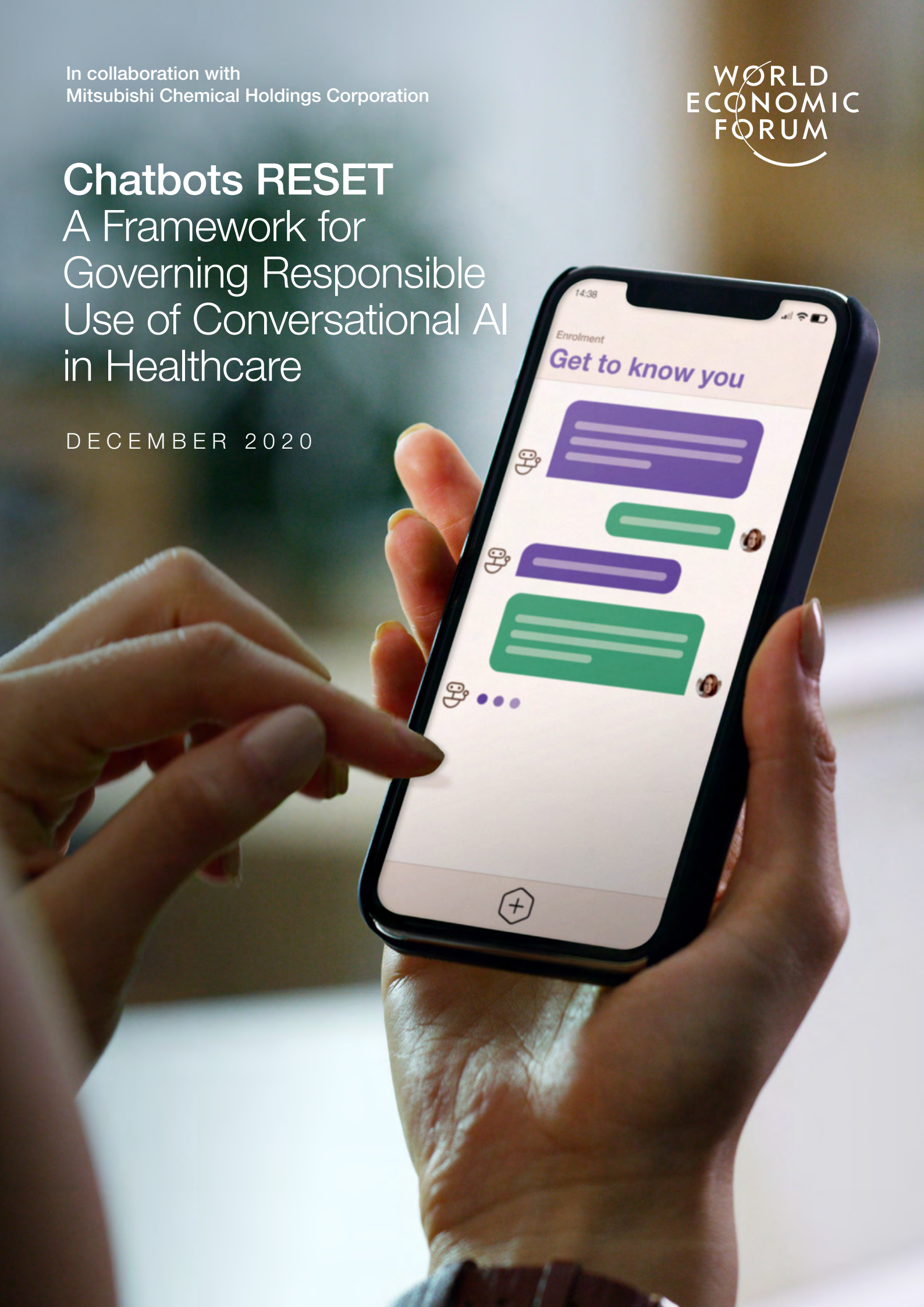
In collaboration with
Mitsubishi Chemical Holdings Corporation

WORLD ECONOMIC FORUM

Chatbots RESET

A Framework for Governing Responsible Use of Conversational AI in Healthcare

DECEMBER 2020



Cover: Sean Anthony Eddy/Getty Images
Inside: Cyano66/Getty Images, Marco VDM/Getty Images, Filadendron/Getty Images, Portra/Getty Images

Contents

Forewords

Larry Meixner
Chief Innovation Officer and CTO, Mitsubishi Chemical Holdings Corporation; Member of the Board, Mitsubishi Tanabe Pharma Corporation

The global COVID-19 pandemic has brought into sharp relief the importance of healthcare for the sustainable well-being of people, society and our planet. This coincides exactly with the corporate philosophy, known as "KAITEKI", of Mitsubishi Chemical Holdings Corporation (MCHC). We are deeply committed to these concepts for the long term, not only in times of emergency.

Artificial intelligence (AI) is enabling transformative advances in many domains, including healthcare. Yet even as such Fourth Industrial Revolution technologies show their power to solve social problems, we must remain conscious of gaps that appear when the technologies advance faster than our ability to govern them. This recognition motivated MCHC to partner with the World Economic Forum in launching and supporting the Chatbots RESET project and to second a Fellow from our company to co-lead the project.

The Chatbots RESET framework reflects months of effort by the project community, with contributions from numerous stakeholders bringing their diverse perspectives to bear on key governance challenges. The principles and recommended actions in this framework are designed to be ethical "guardrails" for the implementation of conversational AI systems to address healthcare problems. We believe this framework can help promote and generate examples of the responsible use of technology in healthcare.

Shobana Kamineni
Executive Vice-Chairperson, Apollo Hospitals

Use of AI and chatbots in healthcare is a true reflection of the current times, when the world had to move in a flash to digital systems. Though the technology advancements have been relentless in this domain in recent past, the pandemic pushed us to adopt and use it, rather abruptly. At Apollo 247 – Apollo Hospitals' Digital Health platform – we have been able to successfully use these technologies and reach out to millions in a very short period of time.

The Centre for the Fourth Industrial Revolution at the World Economic Forum has taken up this timely and impressive work to develop the frameworks and guiding principles for the development and use of chatbots in healthcare. Healthcare chatbot systems can improve and augment accessibility (reaching to the last mile), enhance effective interactions, deliver care faster and with higher accuracy. However, it has to be safe, maintain users' privacy and integrity and be delivered in a fair and inclusive manner. Hence, guiding the ecosystem of developers, users and regulators of this novel field remains as a paramount objective. I strongly believe this framework is developed at the right moment to provide the ecosystem with a guiding light and will help the ecosystem profoundly.

Kiran Thomas
Chief Executive Officer,
Jio Platforms Ltd

The exponential march of digital technologies like broadband connectivity, mobile devices, cloud computing and AI have transformed every aspect of human life. Digital technologies have been used to deliver price-performance-experience packages that are affordable and accessible to not just the privileged few but to billions of people across the globe.

But the solutions to many entrenched problems of humanity have been constrained by the relative scarcity of skilled manpower. A prominent example is healthcare, which has always depended on highly trained physicians. But modern AI services combined with smart sensors are making rapid strides in supplementing and even standing in for human doctors. This overcomes a critical constraint that has prevented widespread access to high-quality healthcare in developing nations, rural populations and other difficult to serve communities.

Even so, healthcare is a sensitive domain requiring due consideration for safety, security and privacy. Fortunately, there are robust standards, codes of conduct and ethics in healthcare that have evolved over centuries of human experience. This white paper is a timely and welcome effort to create a comprehensive framework that makes it easy for technology regulators and developers to conceptualize and incorporate these into AI solutions of the future.

C.P. Gurnani
Managing Director and
Chief Executive Officer,
Tech Mahindra

As the world battles an unseen enemy, the change for humanity is incumbent on AI and its augmentation of human knowledge. From drug discovery to patient care – the world has witnessed AI and its many principles in action fighting this pandemic.

While AI is the new electricity, and data the new fuel, a realization has dawned upon us, the realization of an augmented pollution (AP). This realization reflects upon the maleficent usage of this technology. Where AI could be useful for myriads of cases, the world has also witnessed numerous scenarios where usage of AI without transparency, without empathy and without proper controls causes havoc like deep fakes, maleficent chatbots, etc.

As we reflect on this change, I am pleased to introduce a governance framework for responsible use of conversational AI systems for healthcare, an initiative led by the World Economic Forum along with its partner companies. The framework offers insights from leading AI experts on stages of chatbots creation, adherence to principles like safety, efficacy, explicability, data protection and, above all, its operationalization.

I truly believe that we are at a precipice of a change and the change requires a sustained effort from the technology community to adhere to principles of responsible and ethical usage of AI. I hope these principles will provide you with the necessary insights for a practical implementation of chatbots in real-life healthcare scenarios.

Executive Summary

Chatbots, or conversational artificial intelligence (AI) systems, are used increasingly by organizations to communicate with customers in a natural and easy-to-use way by embedding chatbots in websites, social network apps, smart home devices, etc. The COVID-19 pandemic has accelerated the adoption of chatbots in healthcare applications. As examples, both the World Health Organization and the Centers for Disease Control deployed chatbots for coronavirus information dissemination and symptom checking. So, too, did many governments and healthcare providers. Beyond the pandemic, the rate of adoption of chatbots in healthcare applications is likely to be sustained due to the access and cost benefits they enable.

When a new technology is introduced in healthcare, especially one based on AI, it invites meticulous scrutiny and chatbots are no exception. The exchange of sensitive health information with chatbots is one of many governance challenges that require careful consideration in order to promote responsible use of chatbots in healthcare. Other challenges include performance assurance, patient considerations, legality, privacy and security, in addition to classic AI challenges such as fairness and explainability. To address these governance challenges, earlier this year the World Economic Forum assembled a multistakeholder community, which has co-created Chatbots RESET, a framework for governing the responsible use of chatbots in healthcare.

The Chatbots RESET framework consists of two parts: (1) A set of 10 principles carefully selected from AI ethics and healthcare ethics principles and interpreted within the context of the use of chatbots in healthcare; and (2) Operationalization actions for each principle in the form of recommendations to implement in various stages of deployment of chatbots in healthcare. The framework is an actionable guide for three groups of stakeholders to promote the responsible use of chatbots in healthcare applications: technology developers, healthcare providers and government regulators.

In the Chatbots RESET framework, chatbot developers will find actions they can incorporate within their project development process to create more responsible implementations of chatbots; healthcare providers can integrate actions from the framework within their workflows to promote responsible deployment; and regulators can choose actions that resonate with their national strategy to ensure responsible scale-up of chatbot technology.

Looking ahead, several partners of the Forum will be piloting the Chatbots RESET framework. Results of the piloting, including feedback on and enhancements to the framework, will be shared in a future publication.

1 Chatbots in Healthcare

A chatbot is an AI programme designed to converse in a natural manner with people via voice interfaces or text messages. Chatbots are typically found in websites, applications, or instant messaging. Chatbots are mainly used for tech support and lead generation.

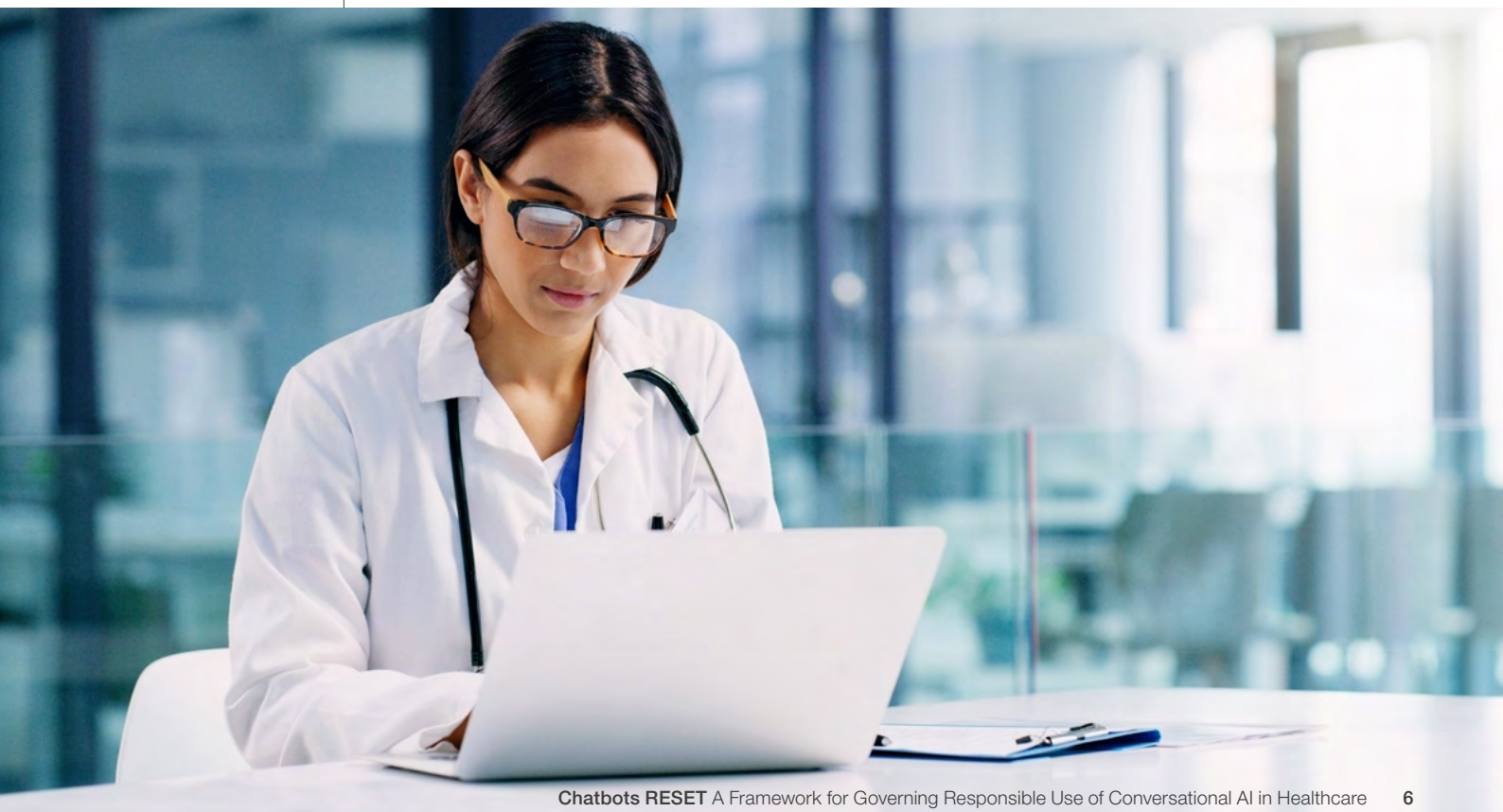
In healthcare, chatbots can be used in many ways to engage with patients, including to navigate frequently asked questions, find a doctor or service, schedule appointments, facilitate symptom checking, conduct triage in emergency care, help prepare for procedures and ensure adherence to post-discharge instructions. Chatbots can also act as virtual assistants to physicians, lowering administrative burden on physicians and giving them easy access to patient health records (see Figure 1 for a more comprehensive list of applications). These uses of chatbots in healthcare can result in better care management and customer engagement.

However, several issues could arise, such as miscommunication between chatbots and customers, customer perception of reduced choices when interacting with chatbots, and neglect of customer preferences in interacting with chatbots vs humans. More serious issues include incorrect/poor guidance, wrong diagnosis, or failure to achieve timely interventions. It is important to thoughtfully govern the deployment of chatbots in healthcare to avoid these issues and to ensure trust, transparency, reliability and security.

What are chatbots and why should you care?

The AIML foundation defines a chatbot as a "computer program designed to respond to text or voice inputs in natural language". Chatbots are also referred to as conversational AI or conversational agents. Typically, chatbots are preloaded with a set of rules or pre-trained using data in order to be able to have a meaningful conversation with the user in real time and to provide useful services in the process.

You might have come across chatbots when looking for help or tech support on websites – they are usually linked to a "chat" icon at the bottom-right corner of the site. You may also have smart home devices at home that you talk to. When you talk to a chatbot, you are talking to an AI system. Though you may find today's chatbots somewhat limited in what they can do, chatbot technology is advancing fast and soon you may not be able to distinguish between an automated system and a human. This can have critical implications if the topic of your chat is healthcare.



2 The Chatbots RESET Project

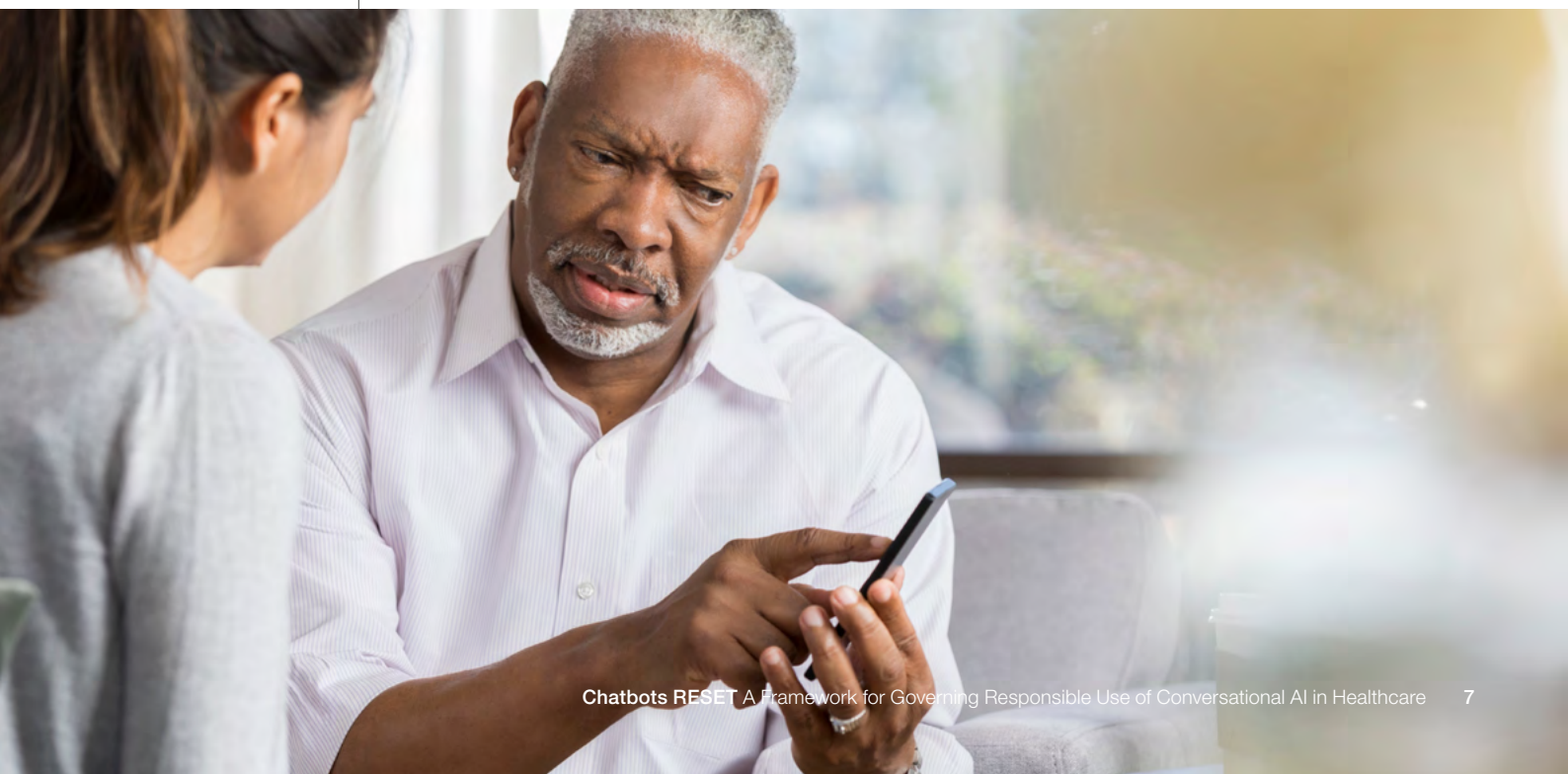
The goal of the World Economic Forum's Chatbots RESET project is to design a governance framework for the responsible use of chatbots in healthcare by bringing together chatbot developers, chatbot platforms, the medical community, civil society, academia and healthcare regulators. Through designing, piloting and scaling the framework, we strive to achieve broad adoption of chatbots in healthcare, maximizing their beneficial uses while minimizing negative consequences.

The Chatbots RESET project was launched in January 2020, with initial thoughts about governance actions alluded to by the word RESET (Reveal, Escalate, Substitute, Explain and Track). Soon thereafter, the importance of chatbots became apparent from their increased use for disseminating curated information about COVID-19 and the coronavirus. In March 2020, the Forum hosted a panel discussion titled "Chatbots for Coronavirus and Beyond", which included an in-depth discussion by a panel of experts from start-ups, platform providers and governments[1]. The discussion underscored the beneficial uses of the technology and its potential to make a positive impact on healthcare. In May 2020, the Forum brought together global experts in a design workshop to brainstorm the potential uses, the issues and stakeholder actions to maximize positive impacts and minimize negative consequences of the use of chatbots in healthcare.

Following the May workshop, the multistakeholder project community began co-creating the governance framework through a series of virtual meetings. The result of this work, the Chatbots RESET Governance Framework, containing 10 principles and 75 recommended actions, is the primary focus of this paper. Piloting the framework in actual use cases, working with partners drawn from the diverse stakeholder community, is ongoing. The results from the pilots will be shared in a subsequent publication.

"Conversational AI holds great promise in healthcare but there are also potential risks and harms which may have direct impact on patient care. Accordingly, the creation of a governance framework for chatbots is a matter of urgency. This work provides a global guide to enable stakeholder organizations to ensure efficacy, safety, privacy and other ethical considerations in their use of chatbots in healthcare and to build public trust in the technology.

– Kay Firth-Butterfield, Head of Artificial Intelligence and Machine Learning, World Economic Forum

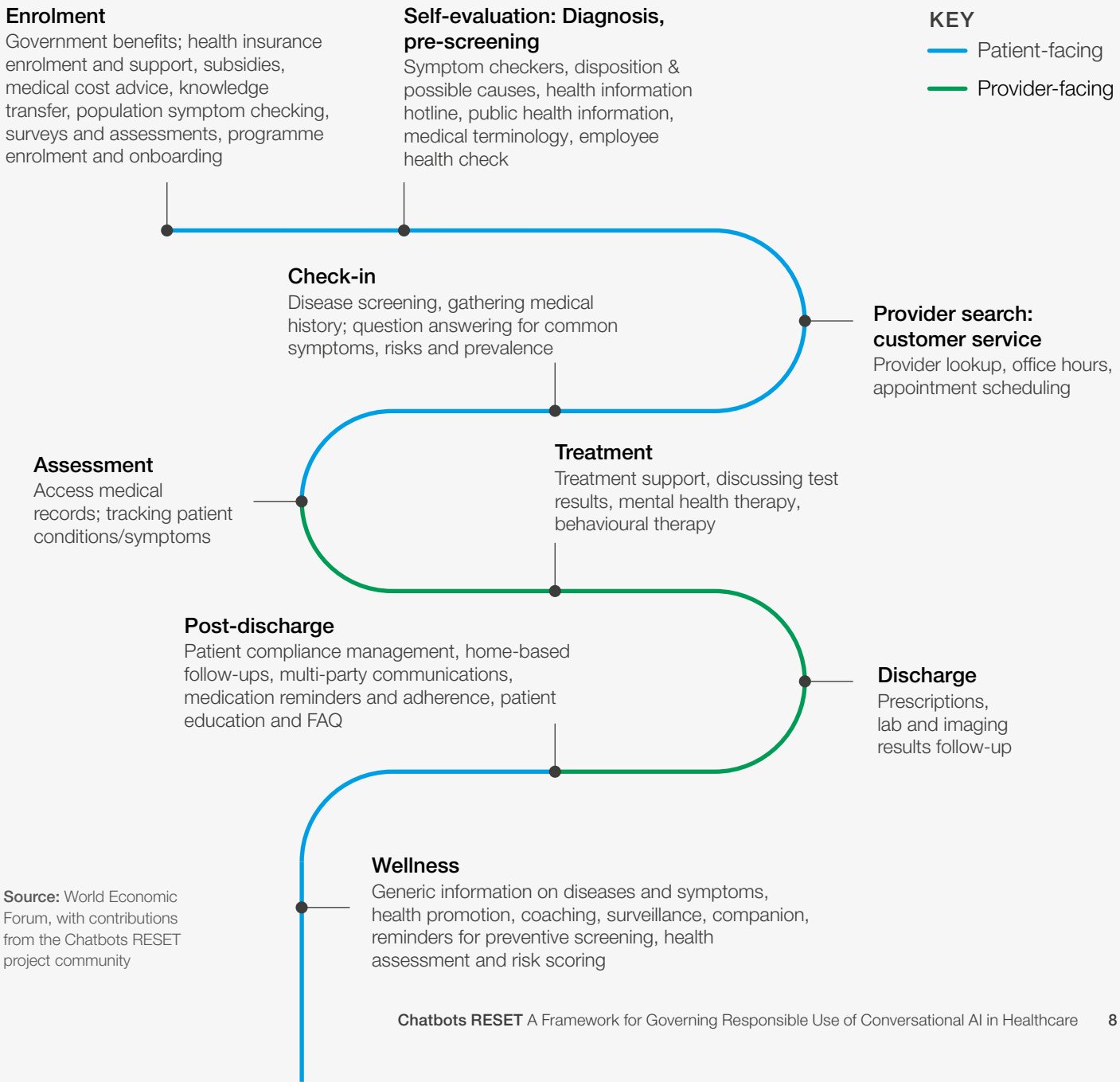


3 Applications of Chatbots in Healthcare

Chatbots are finding increasing uses in healthcare, recently propelled by the COVID-19 pandemic-related information dissemination needs. Many global and national healthcare organizations have adopted chatbots as a way of communicating with their constituents about the virus and the disease. As examples, the Centers for Disease Control uses a chatbot on its website for coronavirus self-checking[2], and the World Health Organization has a WhatsApp messenger chatbot for COVID-19 information. The Microsoft Healthcare Bot platform has been used in over 1,000 chatbot implementations related to the pandemic. Beyond COVID-19, there are numerous applications of chatbots in the healthcare ecosystem. See "Adoption of Chatbots during COVID-19 and Beyond".

There are many current and potential uses of chatbots in healthcare scenarios. Figure 1 provides a glimpse into the plethora of their possible uses,

FIGURE 1: **Chatbots can be deployed in numerous instances in the healthcare journey of an individual, some of which are shown here**



Source: World Economic Forum, with contributions from the Chatbots RESET project community

mapped on to a typical healthcare journey. It is worth noting that not all chatbots are patient-facing. Provider-facing chatbots, while in early stages today, are likely to expand to address many back-office functions in a virtual assistant format.

Chatbots are not always fully autonomous. A range of automation is possible, as we suggest in Table 1, which portrays increasing levels of automation with escalating role and decision-making authority for chatbots relative to human operators or providers. This type of categorization is similar to that used for autonomous vehicles[3].

TABLE 1: **Levels of automation in the use of AI in chatbots for healthcare, akin to levels of automation used to classify autonomous vehicles**

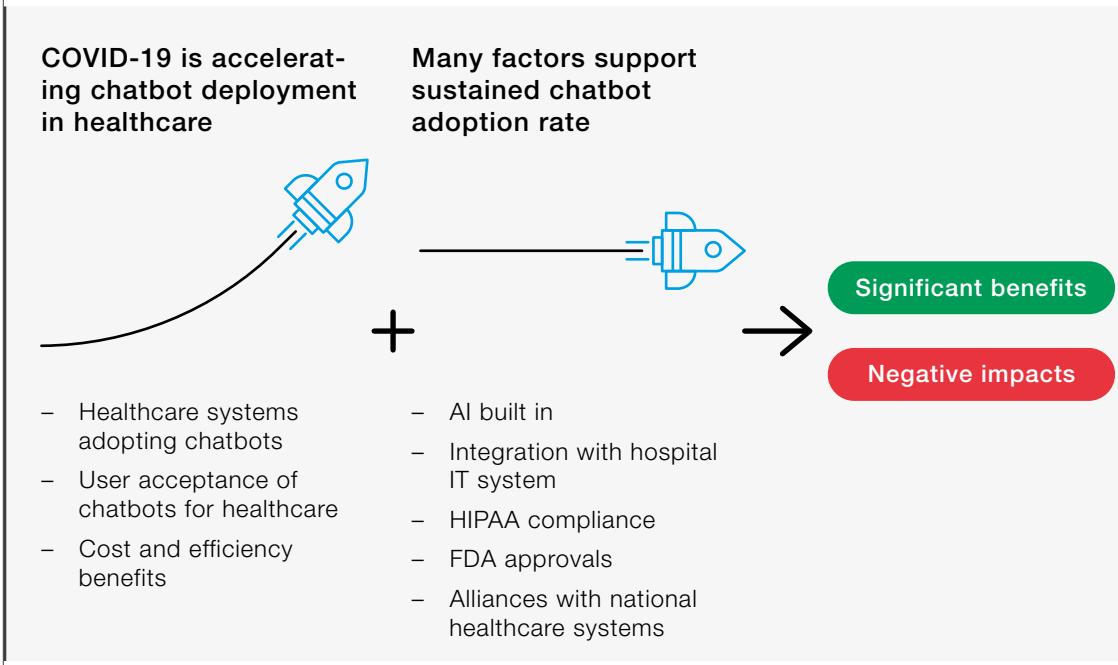
		0 No Automation	1 Basic Assistance	2 Personalized Assistance	3 Partial Automation	4 Conditional Automation	5 Full Automation
Description of roles	**Chatbots**	No role	Pre-diagnostic information collection	Collect and evaluate patient-specific information	Suggest diagnosis and treatment	Diagnosis and treatment of minor illnesses with no human supervision	Entire process of diagnosis and treatment planning
	Human	End-to-end	Analysis and decision-makig	Verification and decision-making	Verification and decision approval	Intervention in special or difficult cases	Approval only, if needed
Data sources		Open	Open	some EHR*	EHR	EHR	EHR
Decision-maker		Human	Human	Human	Hybrid (human + chatbot)	Hybrid (human + chatbot)	Chatbot
Example		Telemedicine	Wellness	Symptom checkers	Patient guidance	Self-service	Cognitive behavioural therapy

(* EHR refers to electronic health records)

Source: World Economic Forum, based on a suggestion by Murali Doraiswamy, a member of the project community

Adoption of chatbots during COVID-19 and beyond

FIGURE 2: **COVID-19 is an accelerator for chatbot adoption in healthcare**



Source:
World Economic Forum

In March 2020, we convened a panel of global experts to explore how Chatbots are being used during the COVID-19 pandemic, and on how they might be used beyond the pandemic[1]. Here is what we learned.

During the pandemic, healthcare systems ramped up adoption of chatbots to provide curated information about the virus and the disease. Most chatbots included symptom checking and guidance on next steps. User acceptance of chatbots for healthcare information is increasing. Significant cost and efficiency benefits are seen as a result of reduced workload on call centres and increased capacity to handle healthcare inquiry volumes.

Telehealth consultations using audio/video links have increased significantly during the COVID-19 pandemic[4]. These are harbingers for AI-based chatbot consultations in the future, as patients are willing to forgo office visits in favour of consulting with an expert – human or AI – on their smartphones.

Beyond the pandemic, there are many factors that will support sustained adoption rate for chatbots in healthcare:

– Most COVID chatbots already have AI built-in. Though heavy use of AI might be limited today to ensure consistent, curated information, the use of AI can be easily ramped up later

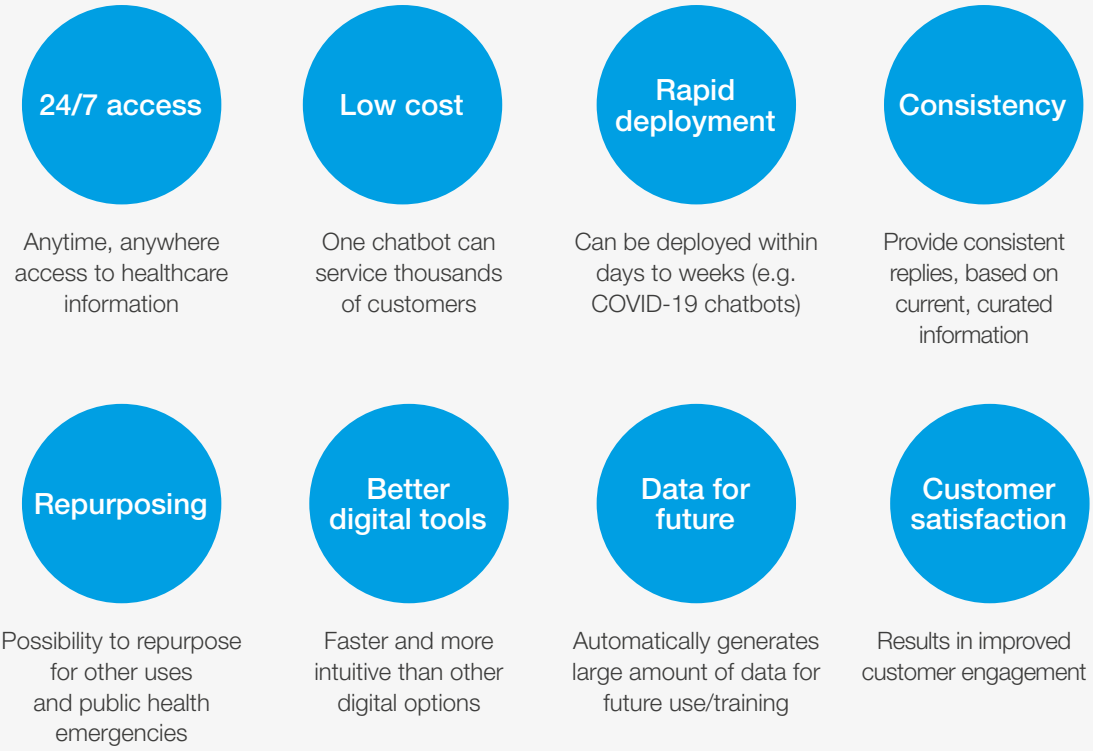
– Most chatbots are HIPAA-compliant, and some have FDA approvals as medical devices

– Many have been integrated with hospital IT systems, and some even have alliances with national healthcare systems (e.g., Babylon Health, with the National Health Service in the UK[5])

With high adoption of chatbots for healthcare, society can reap significant benefits (see box 'Benefits'). At the same time, there are many potentially negative impacts that create challenging governance gaps. Addressing the governance gaps is the focus of the Chatbots RESET framework.

Benefits

Healthcare can reap significant benefits by using chatbots. Some are listed below:



Source:
World Economic Forum

Governance gaps

The following is a list of key governance gap areas we have identified, along with examples of typical issues raised in each area. The questions in each area are meant to be representative of the kind of governance gaps issues, rather than being a comprehensive list.

Validation/accreditation

- What are the boundaries of chatbot operations? In other words, what are "approved" uses of chatbots in healthcare?
- Are current standards for regulating chatbots (e.g., as medical devices) adequate?
- Do chatbots need to be qualified, in a manner similar to qualifying medical doctors, assistants, etc.? If so, who will do it and how?

Performance assurance

- How do we ensure that the limits of performance of chatbots are well understood by everyone? For example, how will they deal with poor spelling, or speech in a noisy environment?
- How will chatbots catch errors (e.g., misunderstanding user inputs)? What will be the action/remedy if they don't?
- Is a second opinion needed every time a diagnosis is made?
- How will a chatbot escalate issues that are critical or those that it is unable to understand/handle?
- What are the expectations of the healthcare system (providers, payers, etc.) on what chatbots should/should not do? How will they verify that the performance of a chatbot meets these expectations?
- What is the oversight body for addressing deficiencies in chatbot performance?

Patient considerations

- Patient expectation management
  - Will patients be misled into believing they are talking to a human?
  - What if they want to opt out of talking to a chatbot?
  - What if they want to switch to talking to a human at any time of their choice (similar to "dial 0 for operator" in automated phone calls)?

- Patient access
    - Will the system understand their (native) language? How will chatbots operate in countries with thousands of dialects?
    - What are the hardware/software requirements? Is a smartphone needed? How about connectivity/bandwidth needs?

Legality
- Who is responsible for wrong diagnosis or misdirection or lack of timely response?
- What are preventive and punitive measures?
- How would consent work, for using the system, and for allowing access to and storage of personal data and chats?

Privacy
- Who will have access to electronic health records?
- What will be the rules governing the recording of chats?
- What if user accidentally reveals other private information in a chat?

Security
- Where will the chatbot and the AI reasoning system reside?
- Will conversations be recorded? If so, where will they be stored? Who will have access to the recordings, and when?
- Will there be options to "host" and/or store data on-premise?

"Classic" AI governance gaps
- How do we avoid "digital divide", from access, language and knowledge-level perspectives?
- What about transparency and explainability of AI-powered systems?
- How do we deal with bias and fairness and also make the solutions are relevant to the target population?
- How can we address data privacy/data rights issues?

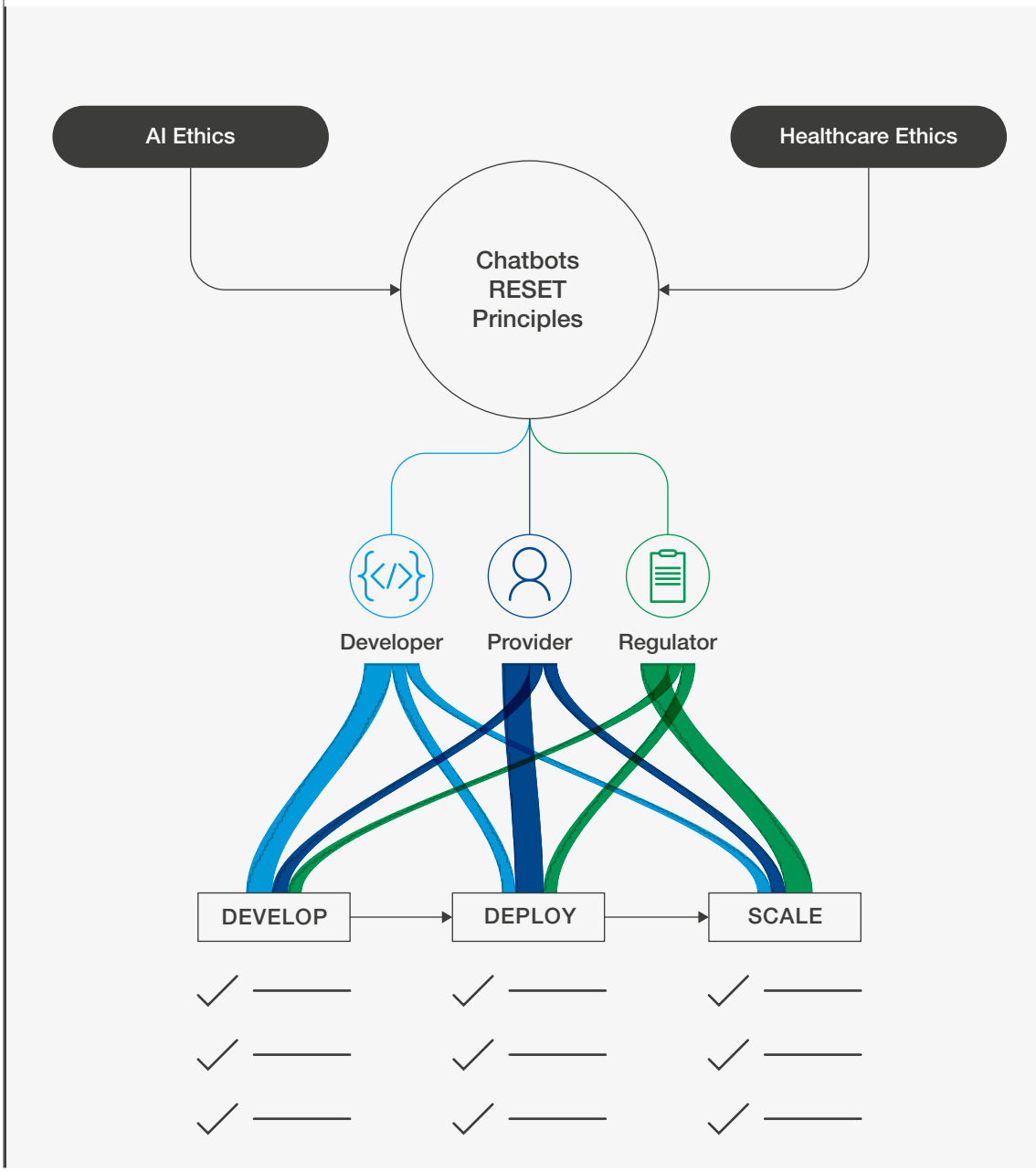
We strive to address these governance gaps in the Chatbots RESET framework, which is presented in the remainder of this document.

4 The Chatbots RESET Framework

The framework consists of two parts:

1. *A set of principles selected by the multistakeholder community to govern the use of chatbots in healthcare.* The principles have been drawn from AI ethics principles and healthcare ethics principles (see Appendix) and interpreted specifically for the use of chatbots in healthcare applications.
2. *Actions that stakeholders can take to operationalize the principles in various stages of the use of chatbots in healthcare.* For each principle, the framework recommends a set of actions that stakeholders can take to implement the principles.

FIGURE 3: **The Chatbots RESET framework consists of principles for responsible use of chatbots and actions to operationalize the principles**



Source:
World Economic Forum

Rationale of the framework

- By the process of starting from key AI and healthcare ethics principles and *interpreting* each principle for the specific use of chatbots in healthcare, the framework provides two benefits: (1) The users of the framework are relieved from the guesswork of interpreting the principles; and (2) The curated list of interpreted principles serves as a uniform standard.
- The operationalization actions provide specific recommendations to the users of the framework on how to implement the principles during the development, deployment, and scale-up stages of the use of chatbots in healthcare. This allows the users of the framework to focus on execution of the actions rather than development of the actions. (Caveat: The suggested actions are meant to be starting points to implement the principles, but they do not exhaustively cover all possible scenarios.)
- The framework has been co-created by a multistakeholder project community, with representation from start-ups, large companies, academia, governments and civil society. The diversity of the community and healthy debates during the creation of the framework have sharpened the focus on principles and actions of broad appeal, which have been carefully curated by the community.

Types of stakeholders

The framework has been developed with three types of stakeholders in mind, shown in Figure 4. Developers drive the creation of chatbots, providers pilot and deploy chatbots in healthcare applications at local/small scale, and regulators monitor and govern the widespread use of chatbots in society for healthcare purposes. Some examples of each type of stakeholder is shown in the figure below.

FIGURE 4: **We consider three types of stakeholders – developers, providers and regulators – and give examples under each**

Developer

- Chatbot developers
- Conversational AI platforms
- Chatbot hosting platforms

Provider

- Hospitals
- Insurance companies
- Technology providers (direct to patient)
- Government (as a provider)

Regulator

- Government (as a regulator)
- Consumer bodies
- Medical bodies

Source:
World Economic Forum

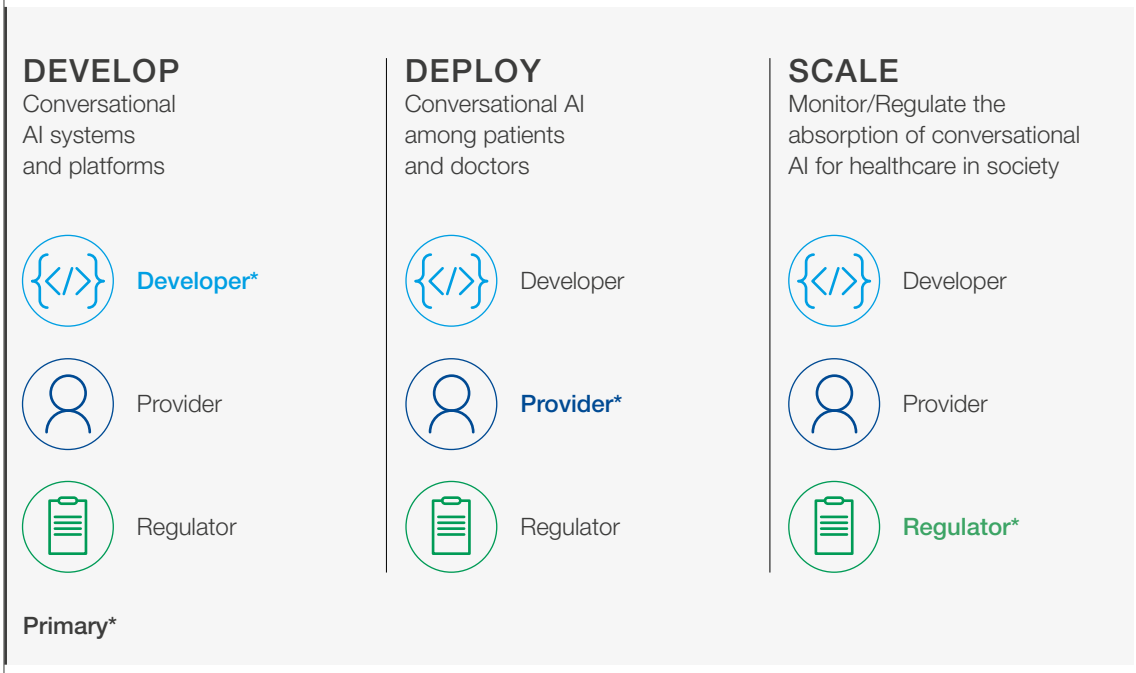
Stages of the use of chatbots in healthcare

The framework provides recommendations for actions to be performed during three operationalization stages (Figure 5):

1. **Develop:** Chatbots are designed and developed by technology developers using modern AI/ML techniques to meet potential needs in healthcare, in consultation with providers and in compliance with laws developed by regulators
2. **Deploy:** Chatbots are piloted and employed in small-scale by providers, with assistance from technology developers and guidance from regulators
3. **Scale:** As chatbots are broadly adopted by society, they begin to impact large and diverse populations, overseen by regulators, with support and compliance from technology developers and providers.

All the stakeholders have roles to play in all the stages, but one stakeholder takes on a primary role in each stage.

FIGURE 5: **The Chatbots RESET framework addresses three stages of operationalization. Develop, deploy and scale**



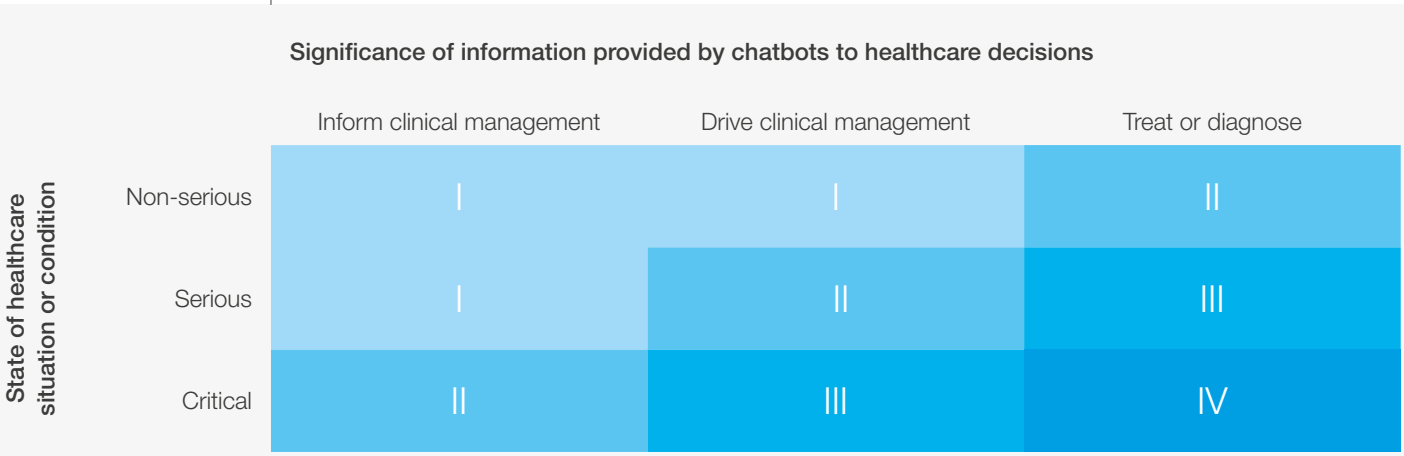
Source: World Economic Forum

Types of chatbots

Not all chatbots are created equal; they constitute a spectrum as they address a vast array of applications within healthcare. At one extreme, they address processes such as scheduling appointments and follow-ups or providing information on diseases and drugs. At the other extreme, they diagnose and suggest treatment plans for severe illnesses or provide therapeutic guidance for mental health. Because of the different types of risk levels involved in the use of different types of chatbots, the operationalization actions of the framework are not equally applicable across the spectrum of chatbots.

To address this diversity of risk levels, the framework includes a preliminary classification of Chatbots into four types (Types I, II, III, or IV) based on the severity of the healthcare condition and the significance of the information provided by the chatbots to healthcare decisions, as outlined in Figure 6, which is directly inspired by the approach

FIGURE 6: **The four "types" of chatbots**



Source: Based on International Medical Device Regulators Forum Final Document "Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations"[6]

of the International Medical Device Regulators Forum to software as a medical device[6]. In the framework, we recommend operationalization actions as they pertain to the four "types" of chatbots on a three-level scale: optional, suggested and required.

To facilitate the interpretation of the chatbot type classification in Figure 6, we provide examples in Table 2 that may be helpful to the users of the framework to more easily identify chatbot types based on their application scope.

TABLE 2: **Examples and risk-level labelling for the four types of chatbots**

Type	Risk level	Example
I	**Low**	Information only: addresses, office hours, find doctor, community health, pandemic information, medicine dosage, drug interactions; scheduling: appointments; post-visit follow-up
II	**Moderate**	Symptom checking without diagnosis; generic next step recommendations
III	**High**	Diagnosis; specific next step recommendations
IV	**Very high**	Treatment plan

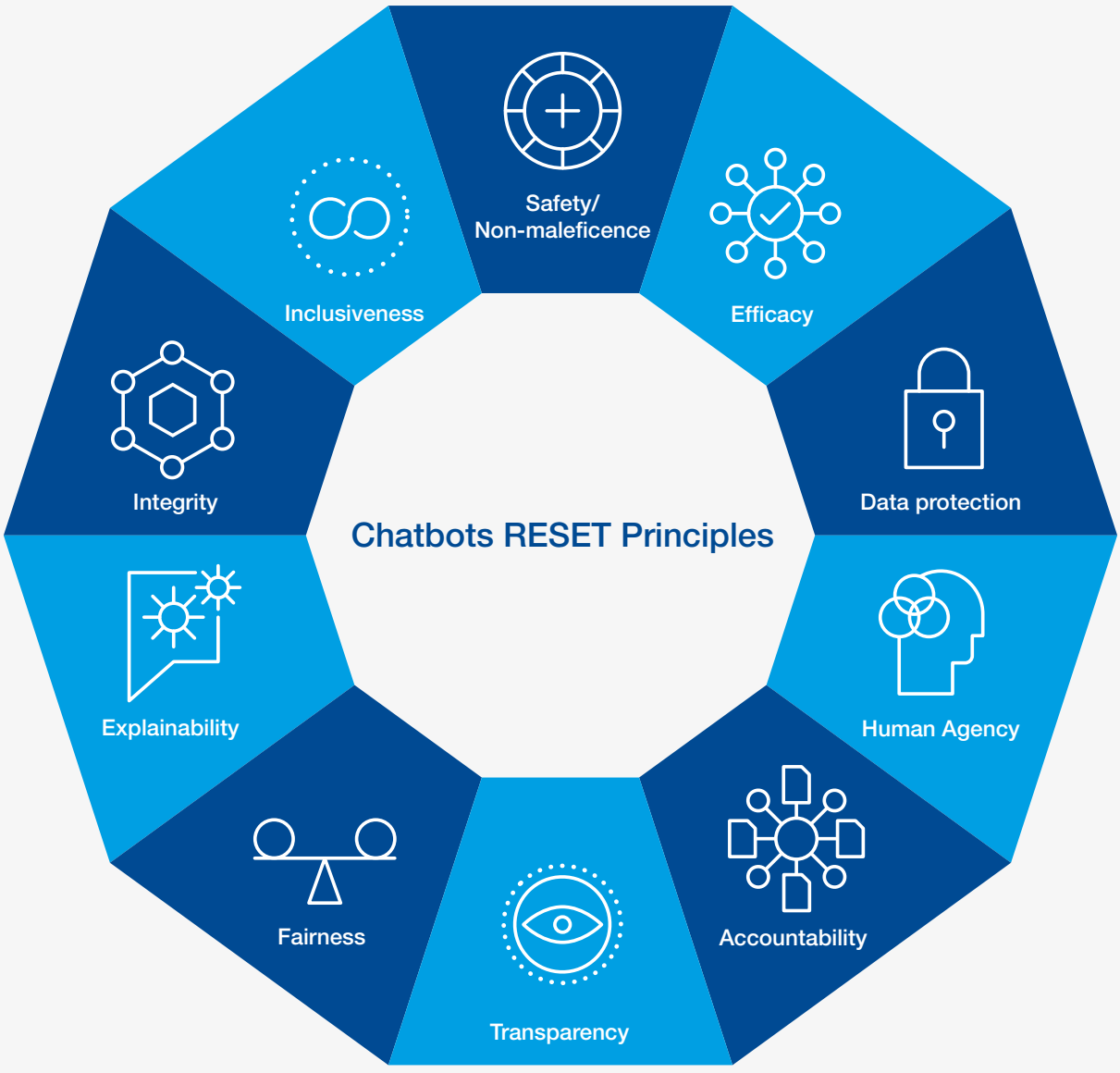
Source:
World Economic Forum

The following pages of the document focus on the two parts of the Chatbots RESET framework shown in Figure 3.

5 Chatbots RESET: Principles

Chatbots offer an unprecedented opportunity to enhance healthcare in a responsible and evidence-based manner.

– Murali Doraiswamy, Digital Health Innovator, Duke University School of Medicine



Source:
World Economic Forum

The principles for the framework have been derived from both AI and healthcare ethics principles, interpreted for the use of chatbots in healthcare. The 10 Chatbots RESET principles derived, interpreted and curated by the Chatbots RESET project's multistakeholder community are presented here. In this section, the word "users" refers to those who directly interact with chatbots.

Safety/Non-maleficence

- The actions of chatbots shall not result in avoidable harm to humans or other unintended consequences, including deception, addiction and lack of respect for diversity

Efficacy

- Chatbots shall be fully verified for the efficacy of their purported service, in compliance with accepted international standards
- Chatbot outputs shall be tailored to their intended users, while keeping in mind the medical nature of the information that is being communicated

Data protection

- All data and history of interactions, including intended and unintended revelations of private data and those collected with consent, shall be safeguarded and disposed of properly, respecting applicable privacy and data protection regulations/laws
- If any data is recorded during a session and/or used across sessions, the chatbot user consent and/or any applicable ethics body approvals for research and data collection purposes shall be required
- Chatbot users shall have the right and access to take ownership of personally identifiable information
- Data collected by chatbots shall not be used for surveillance or punitive purposes, or to unfairly and opaquely deny healthcare coverage to users

Human agency

- Chatbots shall support the user's agency, foster fundamental rights and allow for human oversight
- Chatbots shall respect the ability of patients to make their own decisions about healthcare interventions
- Chatbots whose operating model includes real-time human oversight shall yield to the desire of the user to interact with a human agent at any time the user wishes to do so

Accountability

- An entity (person or group) in the organization shall be accountable for the governance of chatbots
- Conclusions and recommendations of chatbots shall be auditable

Transparency

- Chatbot users shall at all times be made aware of whether they are interacting with an AI or a human or a combination of the two
- Chatbots shall clearly inform users about the limits of performance of the system, except in situations where not informing is required for the intended purpose of the chatbot
- Chatbot users shall be immediately informed if the chatbot is unable to understand the user or is unable to respond with certainty, except in situations where such communication interferes with the intended purpose of the chatbot

Fairness

- Chatbots shall not act in a systematically prejudiced manner with respect to ethnicity, geography, language, age, gender, religion, etc.
- If a chatbot "learns" from data, the training dataset should be representative of the target population

Explainability

- Decisions and recommendations made by chatbots shall be explainable in a way that can be understood by their intended users

Integrity

- Chatbots shall limit their reasoning and responses to those that are based on reliable, high-quality evidence/data, ethically sourced data and data collected for clearly defined purpose

Inclusiveness

- Every effort shall be undertaken to make chatbots accessible to all intended users, with special consideration given to identifying and enabling access for potentially excluded or vulnerable groups

6 Chatbots RESET: Operationalization

Principles require actions to implement them. In this section, we share the collection of actions developed by the Chatbots RESET project community. For each principle in the previous section, we present the following, one principle per page:

Principle: The name of the principle.

Interpretation: The principle is interpreted for the use of chatbots in healthcare (reproduced from the previous section).

Table of operationalization actions: The actions that can be taken by developers, providers and regulators to implement the principle. The table lists, by stage of implementation (as described in Figure 5:), the actions and the responsible stakeholder for each action. The last column of the table shows a code that corresponds to the applicability of each action to the four types of chatbots (outlined in Figure 6 and further elaborated in Table 2), using the following mnemonic: green, optional; yellow, suggested; red, required.

Here is an example of the code for an action that is optional for Type I, suggested for Type II, and required for Types III and IV:

Type I	Type II	Type III	Type IV
Optional	Suggested	Required	Required

For easy reference, we reproduce the icons for the stakeholders:

Developer

Provider

Regulator

Operationalization actions that cut across principles

While developing the operationalization actions for specific principles, we identified the following actions that have broader applicability beyond a single principle and often cut across all principles:

1. (Developers, providers, regulators) Create a redressal mechanism for users, including, but not limited to, the ability for users to provide feedback and seek recourse

2. (Regulators) Create a mechanism for providers to openly share causes, analyses and conclusions reached in situations that resulted in unexpected harm

3. (All) In all stages, follow well-established standards of inclusiveness. Some examples are provided here:

   – United Nations Convention on the Rights of Persons with Disabilities[7]

   – Know Your Rights: Three Important Federal Laws that Protect People with Disabilities[8]

   – ISO/IEC 30071-1:2019 Information technology – Development of user interface accessibility[9]

   – Digital Inclusion, Identity, Trust and Agency[10]

Safety/Non-maleficence

The actions of chatbots shall not result in avoidable harm to humans or other unintended consequences, including deception, addiction and lack of respect for diversity

“The focus on patient / user safety remains paramount in clinical chatbots with the principled aim – do no harm. However, as technology develops, we would need to address how AI and clinicians together aim for better accuracy and foster a culture of safe and effective communication between humans and machines.

– Sujoy Kar, Chief Medical Information Officer, Apollo Hospitals Group

Operationalization actions

Stage	Action	Responsible	Code			
			I	II	III	IV
DEVELOP	Build on existing guidelines to allow for determination of critical/serious/non-serious cases		🔴	🔴	🔴	🔴
DEVELOP	Design a robust hand-off system for situations when AI fails		🟢	🟡	🔴	🔴
DEVELOP	Install safeguards to identify abnormal behaviour and prevent manipulation		🟢	🟡	🔴	🔴
DEPLOY	Develop mechanism to govern consent for use/ treat/diagnose options of chatbots		🟢	🔴	🔴	🔴
DEPLOY	Track and document mistakes attributable to chatbot; share with developers and regulators for continual improvement		🟢	🟢	🟡	🔴
DEPLOY	While suggesting diagnostic/treatment options, consider issues related to patient safety		🟢	🟡	🔴	🔴
DEPLOY	Train all personnel on when and how to intervene		🔴	🔴	🔴	🔴
SCALE	Provide online user education (“How chatbots work”)		🔴	🔴	🔴	🔴
SCALE	Perform planned and unplanned audits of developer documentation		🟢	🟢	🟡	🔴

Efficacy

Chatbots shall be fully verified for the efficacy of their purported service, in compliance with accepted international standards.

Chatbot outputs shall be tailored to their intended users, while keeping in mind the medical nature of the information that is being communicated.

When users see a technology as effective, they begin to trust and use it. Health domain is all about trust, and efficacy principles will be key to getting chatbots adopted.

– Biplav Srivatsava, Professor of Computer Science, AI Institute, University of South Carolina

Operationalization actions

Stage	Action	Responsible	Code I	Code II	Code III	Code IV
DEVELOP	Identify intended users and understand their needs at the beginning of the development process		🔴	🔴	🔴	🔴
DEVELOP	Use verified and tested clinical protocols that have a regular cadence of updates		🟢	🟡	🟡	🔴
DEVELOP	Define bot-only actions, human-only actions and hybrid actions		🟢	🟢	🟡	🔴
DEPLOY	Define a functional test along with chatbot dialog flow to validate behaviour and test for regressions		🟢	🟡	🔴	🔴
DEPLOY	Include efficacy metrics as a central aspect in procurement, and request evidence of efficacy from developers		🟢	🟡	🔴	🔴
SCALE	Create a regionally relevant common testing framework (e.g., ITU/WHO FG-AI4H project[11], TRIPOD[12] for clinical AI) and a standard validation dataset to validate chatbot behaviour (minimum bar) and correctness of diagnosis		🟢	🟡	🔴	🔴
SCALE	Create patient education guidelines to ensure that adequate educational resources are available to lay users to interpret complex medical information		🟢	🟡	🔴	🔴

Data protection

All data and history of interactions, including intended and unintended revelations of private data and those collected with consent, shall be safeguarded and disposed of properly, respecting applicable privacy and data protection regulations/laws

If any data is recorded during a session and/or used across sessions, the chatbot user consent and/or any applicable ethics body approvals for research and data collection purposes shall be required

Chatbot users shall have the right and access to take ownership of personally identifiable information

Data collected by chatbots shall not be used for surveillance or punitive purposes, or to unfairly and opaquely deny healthcare coverage to users

Health and healthcare systems are undergoing a "Great Reset", using technology to enable greater and more cost-effective access to patients around the world. The protection of patient data remains paramount in the development of AI-supported applications. These governance principles are a strong step in the right direction to ensuring ethical and responsible collection of increasing amounts of such data.

– Genya Dana, Head of Shaping the Future of Health and Healthcare, World Economic Forum

Operationalization actions

Stage	Action	Responsible	Code			
			I	II	III	IV
DEVELOP	Require security review before launch		●	●	●	●
DEVELOP	Implement role-based access controls for conversation data		●	●	●	●
DEPLOY	Develop retention policy on conversation data		●	●	●	●
DEPLOY	Create accountability for stored data and penalties for leaks/loss		●	●	●	●
DEPLOY	Require provision of opt-in for users to manage stored data and their uses		●	●	●	●
DEPLOY	Train staff on consent mechanisms and data hygiene practices		●	●	●	●
SCALE	Facilitate private-public data sharing, especially when public funds are used		●	●	●	●
SCALE	Enhance existing health data protection regulation to include chatbot conversations		●	●	●	●

Human agency

Chatbots shall support the user's agency, foster fundamental rights and allow for human oversight

Chatbots shall respect the ability of patients to make their own decisions about healthcare interventions

Chatbots whose operating model includes real-time human oversight shall yield to the desire of the user to interact with a human agent at any time the user wishes to do so

Chatbots in healthcare increase human efficiency and productivity. These chatbots function at their best with the human agent in the loop. Humans and chatbots strengthen each other to deliver optimal care. That is why this governance framework is so important, because the patient deserves the best healthcare possible.

– Nupur Ruchika Kohli, Medical Doctor and Medical Adviser, Specialized Hospital Care and Expensive Medication; Curator, Amsterdam Hub, Netherlands, Global Shapers, World Economic Forum

Operationalization actions

Stage	Action	Responsible	Code			
			I	II	III	IV
DEVELOP	Include an option for seamless human real-time communication		🟢	🟡	🔴	🔴
DEVELOP	Have a diverse test group (and ideally, a diverse development team) representative of the target population group		🟢	🟢	🟢	🟡
DEVELOP	Provide tooling/adaptations for patients with visual impairments		🟡	🔴	🔴	🔴
DEVELOP	Provide for the human agent to view the chatbot conversation history		🟢	🟡	🟡	🔴
DEVELOP	Include a "signpost" to where user can get help in person		🟢	🟡	🔴	🔴
DEVELOP	Include a feature to generate a full transcript for user review		🟢	🔴	🔴	🔴
DEVELOP	Be explicit about how "human in the loop" is defined for the chatbot		🔴	🔴	🔴	🔴
DEPLOY	Verify safety of guidance/information provided when a human is not available for real-time oversight		🟢	🟡	🟡	🔴
SCALE	Conduct detailed qualitative assessment with a random user set and publish results		🟢	🟢	🟡	🔴

Accountability

An entity (person or group) in the organization shall be accountable for the governance of chatbots

Conclusions and recommendations of chatbots shall be auditable

With the growing demand of conversational AI, especially for healthcare, accountability of a deployed conversational agent needs to be owned by a person, entity or an organization. This accountability would form the quintessential core of the chatbot design principle. 'AI is the new electricity and data is the new fuel', it is said. We have to ensure we are accountable and responsible that when these two forces meet, we create more power for humankind and less pollution.

– Nikhil Malhotra, Chief Innovation Officer, TechMahindra

Operationalization actions

Stage	Action	Responsible	Code			
			I	II	III	IV
DEVELOP	Ensure that chatbot workflows are human audited at least every year internally to maintain current status and accuracy		🟢	🔴	🔴	🔴
DEVELOP	Within the audit, be transparent about whether the chatbot is dynamically learning or is static		🟢	🔴	🔴	🔴
DEVELOP	Seek clinical inputs in the decision to implement a chatbot		🟢	🟡	🔴	🔴
DEPLOY	Liability insurance should cover chatbot malpractice the way they cover healthcare providers		🟢	🟡	🔴	🔴
DEPLOY	In case of "accidents", provide full explanation of why the chatbot did what it did		🟢	🔴	🔴	🔴
DEPLOY	Proactively provide a framework and certification for the safety of chatbots		🟢	🟢	🟡	🔴
DEPLOY	Create a mechanism for accountability, especially if there is an adverse outcome		🟢	🟡	🟡	🔴
DEPLOY	Keep a comprehensive record of data governance		🟢	🟡	🔴	🔴
DEPLOY	Require evidence for engineering best practices (e.g., IEC62304, ISO14971, ISO27001)		🔴	🔴	🔴	🔴

Transparency

Chatbot users shall at all times be made aware of whether they are interacting with an AI or a human, or a combination of the two

Chatbots shall clearly inform users about the limits of performance of the system, except in situations where not informing is required for the intended purpose of the chatbot

Chatbot users shall be immediately informed if the chatbot is unable to understand the user or is unable to respond with certainty, except in situations where such communication interferes with the intended purpose of the chatbot

“The clear communication of limitations is critical because users and regulators need to develop an accurate theory of function for the chatbot that enables them to consider its recommendations and interactions rationally and critically. Because chatbots are engineered systems, users do not have priors to help them intuitively build a model of functionality and limitation as they do for human experts.

– Illah Nourbaksh, K&L Gates Professor of Ethics and Computational Technologies, Carnegie Mellon University

Operationalization actions

Stage	Action	Responsible	Code			
			I	II	III	IV
DEVELOP	Use a chatbot persona clearly distinct from that of a human		🟢	🟢	🟡	🔴
DEVELOP	Provide an explanation of decision or inference, in plain language, infographic or video		🟢	🟡	🔴	🔴
DEPLOY	Publish limitations of the chatbot (possibility of errors and consequences) and reliability of the chatbot		🟢	🔴	🔴	🔴
DEPLOY	Require developers to inform users when chatbot fails to understand the user		🟢	🟡	🔴	🔴
DEPLOY	Do testing in realistic conditions (pilot) to verify that the chatbot can inform the user if there are issues with understanding		🟡	🔴	🔴	🔴
DEPLOY	Set out guidelines on exceptions to transparency		🟢	🔴	🔴	🔴
DEPLOY	Inform users whether the matter is non-serious, serious or critical		🔴	🔴	🔴	🔴
DEPLOY	Develop policy to inform users when AI is involved		🟢	🔴	🔴	🔴
DEPLOY	Share limitations of the use of algorithm and data		🟢	🟡	🔴	🔴
DEPLOY	Be explicit in distinguishing between recommendation and information		🟢	🔴	🔴	🔴

Fairness

Chatbots shall not act in a systematically prejudiced manner with respect to ethnicity, geography, language, age, gender, religion, etc.

If a chatbot "learns" from data, the training dataset should be representative of the target population

In order to reach the maximum potential of these new technologies, it is critical that we take meaningful steps to ensure fairness and representation at every stage of their design, development and deployment. This framework emphasizes the importance of representative datasets, diverse development teams and collaboration at every level.

– Matthew Fenech, Medical Safety Lead, Ada Health

Operationalization actions

Stage	Action	Responsible	Code			
			I	II	III	IV
DEVELOP	Ensure that data used in development includes underrepresented groups		🟢	🟡	🔴	🔴
DEVELOP	Create and publish representative population statistics (including demographic and other) in a format appropriate to leverage for ML training		🟢	🟡	🔴	🔴
DEVELOP	Provide details on models, data source and collection methodology		🟢	🟡	🔴	🔴
DEPLOY	Define the distinction between unfair advice and personalized advice		🟢	🟢	🟡	🔴
DEPLOY	Allow for review by medical ethics commission		🟡	🔴	🔴	🔴
DEPLOY	Conduct evidence-based studies to prevent bias		🟢	🟡	🟡	🔴
SCALE	Set specific accountability when building solutions for sensitive groups		🟡	🔴	🔴	🔴
SCALE	Require open APIs to allow third-party testing		🟢	🟡	🔴	🔴

Explainability

Decisions and recommendations made by chatbots shall be explainable in a way that can be understood by their intended users

“Trust between patients and healthcare providers is fundamental in this sector. When using AI in healthcare, understanding how AI systems make recommendations and decisions is the first step to building that trust. This should be achieved through communication that can be easily understood by the intended audience.

– Zee Kin Yeong, Assistant Chief Executive (Data Innovation and Protection Group), Info-communications Media Development Authority of Singapore; Deputy Commissioner, Personal Data Protection Commission (Singapore)

Operationalization actions

Stage	Action	Responsible	Code			
			I	II	III	IV
DEVELOP	Perform stakeholder analysis to determine the level of explainability expected by users and providers		🟢	🟡	🔴	🔴
DEVELOP	Provide and integrate explanations into the conversation user interface		🟢	🟡	🔴	🔴
DEVELOP	State assumptions made, including user-knowledge level		🟢	🟡	🔴	🔴
DEVELOP	Use chatbot unit-testing tools for constant verification and transparency of chatbot behaviour		🟡	🔴	🔴	🔴
DEPLOY	Promote rating of systems based on testing performed		🟢	🟢	🔴	🔴
DEPLOY	Use lay-person terms in patient-facing chatbot interactions (and provide links to clinical concepts)		🟢	🟡	🔴	🔴
SCALE	Create and maintain a list of use cases where an “unexplainable” black box is unacceptable		🟢	🟡	🔴	🔴
SCALE	Determine acceptable range of performance of chatbot (with deviations triggering further investigation)		🟢	🟡	🔴	🔴
SCALE	Develop classification of levels of explainability (e.g., none; technical/clinical; partial/full, etc.)		🟢	🟡	🔴	🔴
SCALE	Create standardized queries to interrogate chatbots, and standardized metrics of confidence		🟢	🟡	🟡	🔴

Integrity

Chatbots shall limit their reasoning and responses to those that are based on reliable, high-quality evidence/ data, ethically sourced data and data collected for clearly defined purpose

“ In the near future we will have virtual assistants acting as digital twins – a single point of contact to our digital life. Constantly learning and improving to serve us in our business and private spheres. Like in most relationships, these intelligent, life-improving assistants won’t be trusted and adopted if they don’t come with a high level of integrity – especially in healthcare.

– Jascha Stein, Co-Founder and Chief Executive Officer, OmniBot.ai

Operationalization actions

Stage	Action	Responsible	Code			
			I	II	III	IV
DEVELOP	Implement functional tests to frequently validate the integrity of the chatbot and that the conversation behaves as expected		●	●	●	●
DEVELOP	Be open and transparent about article/data sources		●	●	●	●
DEPLOY	Build and communicate processes that protect and handle data, with stakeholders that may interact with the chatbot		●	●	●	●
DEPLOY	State intended use of chat logs for training/research		●	●	●	●
SCALE	Create a curated list of valid sources of data and reliable medical knowledge repository		●	●	●	●

Inclusiveness

Every effort shall be undertaken to make chatbots accessible to all intended system users, with special consideration given to identifying and enabling access for potentially excluded or vulnerable groups

“ Service to every life matters, with all its diversity; [this] is an unreachable aim today as constraints exist and the effort versus efficacy-funnel effect objectively limits service to some lives. Comprehensive primary healthcare provider chatbots of tomorrow should aim to reach the unreachable, speak to the unspeakable of today and not add another layer of constraint.

– Suresh Munuswamy, Head of Technology Innovations & Health Informatics, Public Health Foundation of India

Inclusiveness in a national context: India

The value of data should be measured in its diversity and dimensionality and not by its database size. India has 22 official scheduled languages and about 19,500 dialects, some of them spoken by fewer than 10,000 people with no written script. Medicine is primarily taught in English and practised in the regional language. Primary healthcare is delivered through trained healthcare workers who extend the service to a few more dialects. This graded hierarchical model favours simplicity and structural uniformity over diversity of the determinants of disease, leading to a direct treatment-centric approach versus a diverse prevention-centric approach.

A simple healthcare symptom data point like fever can possibly be expressed or sourced through hundreds of words. If one were to attempt to source and document the cause or actual determinant of fever as data points – across languages or dialects – it would be an impossible challenge in the current ecosystem in India.

Conversational AI systems, or chatbots, have the potential to address these limitations by bridging the gap, directly connecting healthcare professionals to patients and hopefully preserving or improving the diversity of conversation content if, and *only* if, they are programmed to value diversity and dimensionality.

Data and resultant service should be continuously evaluated for its population scale representativeness and diversity in inputs, outputs and outcomes. Extra effort should be directed to expand the reach in terms of diverse socio demographics and economics. Reaching comprehensively as in every person on the planet should be the aim rather than reaching to the constraint-less creamy layer.

– Suresh Munuswamy, Head of Technology Innovations & Health Informatics, Public Health Foundation of India

Operationalization: See the section on general actions

Conclusion

In this paper, we presented Chatbots RESET, a framework to govern the responsible use of chatbots in healthcare applications. The framework was co-created by the World Economic Forum's multistakeholder project community, with representation from start-ups, large businesses, academia, governments and civil society. The diversity of the community and healthy debates during the creation of the framework have sharpened the focus of the framework on principles and actions of broad appeal that have been carefully curated.

This has led to strong interest from Forum partners to pilot the framework. The pilots are designed to test the usability of the framework, gather feedback for further improvement and demonstrate its usefulness with a broad range of organizations and geographies. As we learn from the pilots, we will update the framework to capture the lessons and keep it relevant. We anticipate sharing updates to this ongoing project in future publications of the Forum.

We encourage government officials, industry players, civil society representatives and academics to join us on this journey to strengthen our frameworks and ensure their greater impact.

Appendix: AI and Healthcare Ethics Principles

AI ethics principles

Principle	Definition	Criterion
Accountability	It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.	– Auditability – Minimizing and reporting negative impact – Documenting trade-offs – Ability to redress – Liability
Human agency	AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights, and allow for human oversight.	– Fundamental rights – Human agency – Human oversight
Transparency	This requirement is closely linked with the principle of explicability and encompasses transparency of elements relevant to an AI system: the data, the system and the business models.	– Traceability – Communication
Explainability	Ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system).	
Reliability, robustness and security	Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimizing unintentional and unexpected harm, and preventing unacceptable harm.	– Resilience to attack and security – Accuracy – Reliability and reproducibility
Safety	It must be ensured that the system will do what it is supposed to do without harming living beings or the environment. This includes the minimization of unintended consequences and errors.	– Fall-back plan and general safety
Privacy and data governance	Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.	– Respect for privacy and data protection: – Quality and integrity of data – Access to data
Fairness	Bias affects the fairness of an AI solution and refers to a breach in the performance of AI solutions such that the results are systematically prejudiced; this is typically introduced into the system through three forms: Bias in data, Bias in algorithms, Bias in people.	– Unfair bias avoidance

Principle	Definition	Criterion
Diversity	We must enable inclusion and diversity throughout the entire AI system's life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment.	– Accessibility and universal design – Stakeholder participation
Beneficial AI: Societal and environmental well-being	The broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, such as for instance the Sustainable Development Goals. Ideally, AI systems should be used to benefit all human beings, including future generations.	– Sustainable and environmentally friendly AI – Social impact – Society and democracy

Source: Derived from the ETHICS GUIDELINES FOR TRUSTWORTHY AI, High-Level Expert Group on AI , European Commission https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Healthcare ethics principles

Principle	Definition	Criterion
Non-maleficence	Non-harming or inflicting the least harm possible to reach a beneficial outcome.	"Do no harm", safety
Beneficence	An act of charity, mercy and kindness with a strong connotation of doing good to others including moral obligation.	Fidelity, cultural understanding, empathy
Health maximization	Obligation to maximize health in populations.	Cost-effectiveness or cost-utility analyses
Efficiency	The use of evidence base and the performance of cost-benefit analyses to decide what should be done and how to do it.	
Respect for autonomy	Allowing or enabling patients to make their own decisions about which healthcare interventions they will or will not receive.	Informed consent, confidentiality, privacy
Justice	There should be an element of fairness in all medical decisions: fairness in decisions that burden and benefit, as well as equal distribution of scarce resources and new treatments, and for medical practitioners to uphold applicable laws and legislation when making choices.	Security, equity
Proportionality	It demands that in weighing and balancing individual freedom against wider social goods, considerations will be made in a proportionate way.	

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4196023/