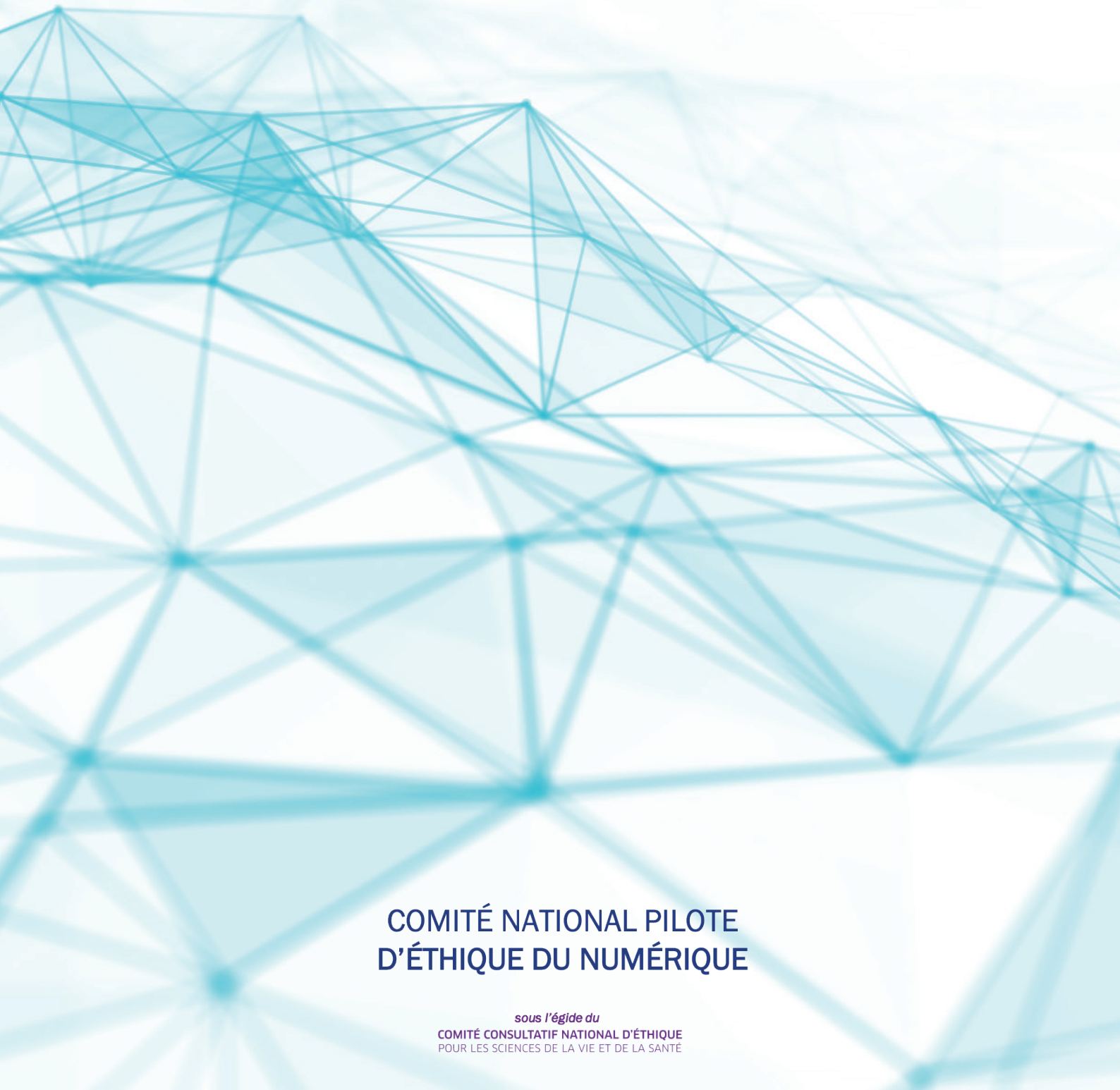
OPINION N°3
ETHICAL ISSUES OF CONVERSATIONAL AGENTS



COMITÉ NATIONAL PILOTE
D'ÉTHIQUE DU NUMÉRIQUE

sous l'égide du
COMITÉ CONSULTATIF NATIONAL D'ÉTHIQUE
POUR LES SCIENCES DE LA VIE ET DE LA SANTÉ

OPINION N°3

ETHICAL ISSUES OF CONVERSATIONAL AGENTS

OPINION ADOPTED ON SEPTEMBER 15, 2021,
BY UNANIMOUS VOTE OF THE MEMBERS PRESENT
AT THE PLENARY ASSEMBLY OF THE CNPEN

REFERRAL OF THE PRIME MINISTER

In his letter of July 15, 2019, the Prime Minister of France has given the President of the National Consultative Ethics Committee (CCNE) a mission to launch a probing investigation into the ethical questions of digital sciences, technologies, applications, innovations, and artificial intelligence. The Prime Minister has emphasized that the work carried out in this pilot phase should focus on medical diagnostics and artificial intelligence, conversational agents, as well as autonomous vehicles. The present opinion of the French National Pilot Committee for Digital Ethics (CNPEN) concerns conversational agents.

TABLE OF CONTENTS

I. INTRODUCTION

A conversational agent (also called a chatbot[1]) is a machine that interacts with users in natural language orally or in writing. Usually, a conversational agent does not function independently but is integrated into a multitask digital system or platform, e.g. a smartphone or a robot.

Ethical reflection on conversational agents stands out among other work in the ethics of artificial intelligence by its focus on language. It includes the analysis of the impact of machine learning systems on human language as well as the impact of language produced by these systems on their users and society in general. Since scientific studies dedicated to these topics are scarce, this opinion aims at shedding light on the issues and challenges involved in large-scale deployment of chatbots.

Similar to the treatment of other AI domains, ethical reflection on conversational agents relies on values in order to put forward design principles and recommendations that are listed below. UNESCO emphasizes[2] that considering the risks and ethical concerns should offer new technological avenues and stimulate research, innovation, and moral thinking, instead of hindering development and progress. Even if all the ethical values and design principles are desirable, a concrete situation may give rise to conflicts between them, for example in public safety and individual freedom, or the efficiency and transparency of artificial intelligence systems. Thus, it is necessary to make decisions on a case-by-case basis and to consider the contexts of design and use, while also respecting the principle of proportionality and fundamental human rights. In each specific application, ethical reflection must reply on the desired goals, technical constraints of the implementation, as well as the short and long-term interests of users.

Chatbots that are capable of written and oral dialogue already provide a wide array of services in the domains of healthcare, social support, human resources, customer support, education, banking, insurance and many others. For example, in healthcare, conversational agents provide services for diagnostics, monitoring, or assistance to patients. Private enterprises readily deploy chatbots in order to automate repetitive tasks, improve the communication with their clients, and reduce costs. Other applications of conversational agents can also have educational or entertainment goals.

Most current-generation chatbots respond to users by following strategies predetermined by their developers. From a user's point of view, such predetermined strategies are limited, because they leave the impression of a "conversational agent without imagination." The success of this approach in more complicated exchanges, as well as the conversational agent's ability to explain its reasons for action, are severely limited. However, there are factors that explain the widespread use of this technology. The state of the art is undergoing changes with the development of chatbots using language models that can hold realistic dialogue. Currently, the developers of conversational agents are striving to create personalized systems that engage with a user in the most efficient way. Scientific and technological research is being motivated by the ambitious visions of a "virtual friend" that imitates emotions and is capable of learning while interacting with a user, or that of a "guardian angel" that will oversee the security of one's personal data. These visions rely on advanced technologies in the domain of machine learning, developed by international research institutes and applied primarily through digital giants, e.g. 'transformer' neural networks that are informed by gigantic datasets. These and similar tools have recently expanded the range of possibilities for speech recognition and automatic text generation. The most recent chatbots raise ethical questions that also relate to the use of affective computing[3] – a range of techniques to influence user behaviour.

HOW ARE CHATBOTS MADE?

Conversational agents have long been developed with a modular architecture using Natural Language Processing (NLP) technologies. These modules rely on machine learning algorithms or, more often, on predetermined rules conceived and written as code by human developers. Such a conversational agent, e.g. a voice assistant, includes modules for input analysis, speech recognition, semantic processing, strategies and history of dialogue, access to internal or external resources (public or specialized databases, ontologies, data available on the web), response generation, and speech synthesis. In recent years, developing a rudimentary chatbot for written or oral speech has become quite approachable even for an individual developer due to many available developing tools.[4]

The new generations of chatbots are becoming more powerful because of the evolution in machine learning techniques, the increase of processing speeds and the size of datasets. Language models, that is, the models that predict a word or a sequence of words within the context of a conversation have become the "grail" of NLP applications, and chatbots in particular. The most recent models are called "transformers." They are neural networks that learn the most likely regularities from vast linguistic corpora without regard for the word order.

[1] These terms have had different meanings in the past - a chatbot was a system that only interacted in writing and had no memory. The present text uses the terms "conversational agent" and "chatbot" interchangeably, which corresponds to the current usage in the state of the art.

[2] UNESCO, Draft text of the Recommendation on the Ethics of Artificial Intelligence, 25 June 2021. https://unesdoc.unesco.org/ark:/48223/pf0000377897

[3] L'Affective computing is the development of systems with the ability to recognize, express, synthesize, and model human emotions.

[4] For example, LiveEngage, Chatbot builder, Passage.ai, Plato Research Dialogue System.

HOW ARE CHATBOTS MADE?

Transformers have been in development since 2017. Many breakthroughs have been made in the field since the launch of a language model called BERT (Bidirectional Encoder Representations from Transformers) by Google. We can highlight July 2020 when OpenAI has launched GPT-3 – a language model with 175 billion parameters and 570 gigabytes of training data; or the launch of LaMDA (Language Model for Dialogue Applications) by Google that was trained specifically on data from conversations that allow the model to engage in free dialogue on a potentially infinite number of topics. More recently, significant advances have been recorded in the beginning of 2021 with the Switch-C model, also developed by Google, that boasts 1600 billion parameters, in the Summer of 2021 with the launch of Jurassic-1 Jumbo by AI21 Labs (Israel) consisting in 178 billion parameters or YaML by Yandex (Russia) with 13 billion parameters in the Russian language, or even WuDao 2.0 by BAAI (China) with 1750 billion parameters – a model oriented towards the English and Mandarin languages. To our knowledge, WuDao is currently the biggest neural network ever created. These models are often not accessible to the public and contain many implicit biases, particularly those related to the opaque nature of the training data. In the interest of transparency, a consortium called "Big Science" led by Huggingface that includes many public research labs is trying to create a language model that is equivalent in size but is open to the public to explore and improve the understanding of the functioning and limits of these huge transformers.

The technological state of the art and the scientific knowledge on conversational agents is evolving at a high pace. The shift towards deep learning algorithms, reinforcement learning, and transformers create new ethical tensions that the developers must tackle with respect to the norms and values of society. However, the norms and values are also evolving under the influence of digital technologies. These technologies not only pose the difficulty of distinguishing artificial dialogue from human speech but also make an impact on the cognitive and emotional state of their human users. Since conversational agents use human language, it is natural for the users to anthropomorphize the chatbots. The evolution of these technologies tends to blur the perceived line between humans and machines.

To nourish its thinking, CNPEN has solicited input from citizens and stakeholders via a call for contributions on ethical tensions related to conversational agents (Annex 2). The contributions received are often split and strongly polarized. Some responses show significant anthropomorphization of the conversational agent, pushing the respondent to equate it with a human person; others, on the contrary, tend to reduce it to an innocuous automatic tool. However, the responses remain univocal on the fact that ethical issues and judgments depend on the aims, concrete applications, and persons concerned in specific use cases. Legal questions also figure prominently in the survey. Those issues that are specific to conversational agents are addressed in the present document, as well as in Annex 1, which relates to the issues of consent and the protection of personal data.

The ethical tensions raised by conversational agents call for a thoughtful and responsible development of these systems. The question of responsibility arises in all its forms: legal and moral, individual and collective, that of the developer, manufacturer, user, and political agent, that arising from possible malfunctions, and that linked to the long-term impact of these technologies. Some of these issues are already subject to regulation. For example, in its white paper on vocal assistants, published in September 2020, CNIL has identified questions that fall within the framework of the General Data Protection Regulation (GDPR)[6].

Going beyond the existing regulation, questions arise about the meaning of human-machine relationships and their respective responsibilities. What behaviours and beliefs do people hold towards conversational agents? What behavioural models should be allocated to the chatbot by the developer, should it systematically imitate human behaviour? Should a chatbot be allowed to lie to its user? Will the errors of conversational agents be judged more or less severely than those of a human being? What are the limits of this comparison?

Human language is an essential element in shaping, or even determining, cultural characteristics, human perceptions, or even entire worldviews. But the linguistic representations presented by the conversational agents are devoid of lived experience. Dialogue generated by a language model that cannot physically perceive, feel or reason like a human, creates a linguistic universe without bodily experience and understanding of meaning. The way that conversational agents use language removes the unambiguous link between language and humans – we converse with machines that can neither take the responsibility for what they say, nor be held responsible for it.

However, a conversational agent is likely to influence the thinking of its user by imprinting notions, perceptions, ideas or even beliefs into their thinking. The user creates a world in which the language of the machines is integrated into reality and social environment. This newly reshaped world feels increasingly real to the individual and tacitly transforms the meaning of their values, such as their own autonomy and dignity in the face of a conversational agent, capable of lying to them and manipulating them. At the societal level, the issue of fairness and non-discrimination must be carefully considered. In the long term, the effects of chatbots, including "deadbots," may produce a significant change in the human condition. The co-adaptation of language between human users and conversational agents is the driving force behind this potential change.

[5] https://blog.google/technology/ai/lamda

[6] https://www.cnil.fr/fr/votre-ecoute-la-cnil-publie-son-livre-blanc-sur-les-assistants-vocaux

[7] « In its widest sense, culture may now be said to be the whole complex of distinctive spiritual, material, intellectual and emotional features that characterize a society or social group. It includes not only the arts and letters, but also modes of life, the fundamental rights of the human being, value systems, traditions and beliefs," Mexico City Declaration on Cultural Policies. Mundiacult World Conference on Cultural Policies, Mexico City, 26 July - 6 August 1982.

[8] A conversational agent that purposely mimics the way a dead person speaks or writes is called a "deadbot".

II. USING CONVERSATIONAL AGENTS: ETHICAL QUESTIONS

Conversational agents are increasingly integrated into various aspects of human life. Their use raises ethical tensions, which in turn lend to the question of responsibility as conceived in moral philosophy. It is rightly understood as the responsibility of human beings because the machine is not a moral agent and should not be considered a person. Thus, responsibility must be shared among the user of a conversational agent, its developer (a computer scientist or a group of developers with specialised knowledge), the 'trainer' (an individual or a group of individuals carrying out the selection and sorting of data and the optimisation of the machine learning algorithms), and the manufacturer (a natural or a legal person who releases the chatbot to the market). The sharing of responsibility is evaluated on a case-by-case basis, depending on the technical aspects and the role played by the user, the developer, and the manufacturer in each of the situations that cause ethical tensions.

The design and use of current and future conversational agents must be explored in the light of ethical issues[9]. This opinion is therefore dedicated primarily to researchers in computer science who are bound to question their design methods and ethically evaluate conversational agents. It is also aimed at industry leaders who must be made aware of the tensions between ethics and trust, as well as the consequences that conversational agents place on the market. In supporting the development of these applications, they must also support work that will address the arising ethical concerns. Finally, this opinion is also addressed to the public authorities, who must increase their support for training and education on ethical issues, but also evaluate short-term effects of conversational agents and facilitate society-wide experiments that will allow to understand their long-term effects.

The approach of all these stakeholders must observe transparency and explicability to ensure that values such as human autonomy, dignity and equity are respected. All their actions should also be in line with the "ethics by design" framework put forward by the European Union.[10]

CNPEN has identified eight directions of ethical reflection relating to the uses of chatbots (chapter II) and five directions relating to particular technologies (chapter III), which inform the thirteen recommendations on uses, the ten principles of design for chatbots, and the eleven research questions formulated in this opinion. Among these different issues, the three main tensions concern the status of conversational agents, the imitation of language and emotions by chatbots, and public awareness of the capabilities and limitations of conversational agents, including their ability to manipulate.

The technological reality and scientific knowledge of conversational agents evolve very rapidly. Accordingly, the reflection on the ethical issues of conversational agents will necessarily have to evolve in parallel to be able to cover the emerging cultural and technological changes. It is necessary that this ethical reflection be continued in the next three to five years.

Since users are not machine learning specialists, they have little or no knowledge of the capabilities of chatbots. Moreover, technology marketing often oversells the features of conversational agents. User beliefs of users are thus fuelled by fiction. At the same time, the users' responsibility can come into the picture in case of malicious or inappropriate use.

As for the developers of conversational agents, they are often unaware of the ethical tensions that may emerge during the use of chatbots. This is either due to the unforeseeable consequences of the technology or to a lack of vigilance or experimentation during the design stage. However, both developers and manufacturers carry responsibility in all cases.

HISTORY

The first conversational agent in the history of computing is ELIZA (1966) developed by Joseph Weizenbaum at MIT, which is also one of the first conversational tricks. ELIZA, which plays the role of a Rogerian psychotherapist, simulates a written dialogue by reusing the user's or "patient's" words and phrases. If a user's sentence contains the word "computer," Eliza may ask, for example, "Are you talking about me in particular?" Today, the term "ELIZA effect" refers to the tendency to unconsciously equate a dialogue with a computer with a human conversation.

[9] E Ruane, A Birhane, A Ventresque, Conversational AI: Social and Ethical Considerations. AICS, 104-115, 2019.

[10] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

[11] Rogerian Psychotherapy is based on the idea that the patient has the solutions and resources within themselves to solve their problems. The therapist helps the patient to develop their own choices through dialogue without directing or influencing them.

HISTORY

The history of conversational agents has its origins in Alan Turing's Imitation Game (Turing, 1950). Turing proposed a test of intelligence for a machine through its ability to converse in natural language. Since 1991, competitions have been organised to support the development of chatbots capable of passing the Turing test. This test is an imitation game in which a hidden entity, which can either be a human subject, or a machine, interacts in writing with another human subject. The aim is for the hidden machine to make the latter person think that they are conversing with a human. Alan Turing predicted that by the year 2000, machines would be able to fool about 30% of human judges during a five-minute test. There is an annual competition created in 1990 called the "Loebner Prize," which rewards the program considered closest to passing the Turing test. Despite the advances in technology, this test does not seem to capture the full complexity of human intelligence. Other tests have been proposed to complement it, such as the "Lovelace test," in which an artificial agent passes a creative task only if its programmer cannot explain how the text produced by the machine was generated. This criterion of intelligence has received criticism, partly because, like all these tests, it is formulated with a negative criterion – if a machine does not pass it, then it is not intelligent[13]; but if it does pass it, this does not necessarily make it intelligent. There is a natural tendency to separate the notion of intelligence from mere problem solving.

With increasing frequency, the manufacturers that produce chatbots look to make an impression for their human users that they are interacting with a "virtual character" endowed with intelligence. Regardless of the personalisation infused during development, the user projects human characteristics on the chatbot in a way that is often spontaneous and unconscious. This unavoidable projection, as well as its anthropological, psychological, legal, and political ramifications, raise many ethical questions. The responses submitted to the call for contributions (Annex 2) are often contradictory and demonstrate the complexity and richness of these questions.

1. STATUS OF CONVERSATIONAL AGENTS

Since the times of ancient oracles, non-human speech has been regarded as a source of revelation and fascination. Whatever the nature of the interlocutor, real or imaginary, embodied as a sculpture, a stone, a god, an animal or a machine, humans naturally project human traits on whoever is speaking to them in natural language. This includes gender, thought, will, desire, consciousness, and representation of the world. For the duration of a conversation, any such interlocutor appears to be an individual with traits that are seemingly familiar, even if, ontologically speaking, it is a non-human or virtual being. The projection of human traits concerns many ethical values and principles: human autonomy and freedom, dignity, responsibility, loyalty, non-discrimination, justice, security, and respect for privacy.

The projection of human traits on a conversational agent is spontaneous. It is usually experienced as a momentary illusion, but in some cases it may persist. Moreover, it can be reinforced through the technical means of personifying a conversational agent, for example by configuring a tone of voice or a manner of speaking. The effects of the projections of traits depend on the technical knowledge of the human conversant, his or her state of mind and affective disposition, but also on the degree of personification of the conversational agent. Clear and understandable communication about the status of the chatbot helps to control the effects of this projection without eliminating it: the user is more quickly and easily aware that they are interacting with a machine but some users, seriously or for entertainment, engage emotionally with a conversational agent despite knowing that they are conversing with a machine. This shows that merely informing the user cannot be sufficient to dissolve all the effects. What is at stake is a true blurring of status distinctions. This is a source of ethical tension, for instance, in matters of human dignity or manipulation.

The moral and ontological difference between a conversational agent and a human being is particularly important with regard to the purpose and the role attributed to a chatbot. All computer systems are designed to achieve a goal determined by their developers, meanwhile human beings are free to set their own goals or to hold conversations without an explicit purpose. From an ethical point of view, the conservation or, on the contrary, the blurring of status distinctions must be evaluated in context. In some situations, the anthropomorphism may cause malicious confusion; in others, despite the danger, it can prove useful for the user.

Depending on the goals, it may be necessary to enforce a distinction between a conversational agent and a human interlocutor during the dialogue, in order to avoid the nefarious effects that it may cause.

The difference in status can also be assessed by the type of application. For example, chatbots in the domain of healthcare are used for medical advice or, e.g. in psychiatry or psychology, for treatment and diagnostics. A "virtual doctor" would be capable of making a diagnosis and assign treatment for common diseases; a "virtual nurse" can monitor the patients. Such chatbots rely on the spontaneous projections of human traits to enforce medical protocols. Others produce positive effects through the explicitly non-human nature of the dialogue.

[12] Bringsjord, S., Bello, P. & Ferrucci, D. Creativity, the Turing Test, and the (Better) Lovelace Test. *Minds and Machines* 11, 3–27 (2001).

[13] Floridi, L., Chiriatti, M. GPT-3: Its Nature, Scope, Limits, and Consequences. Minds & Machines 30, 681–694 (2020).

[14] Ruane E., Farrell S., Ventresque A. (2021) User Perception of Text-Based Chatbot Personality. In: Følstad A. et al. (eds) Chatbot Research and Design. CONVERSATIONS 2020. Lecture Notes in Computer Science, vol 12604. Springer, Cham. https://doi.org/10.1007/978-3-030-68288-0_3

In the justice system, chatbots are used by professionals and the public with increasing frequency as decision aides, allowing them to access databases containing verdicts and juridical precedents. Conversational agents suggest solutions based on the statistical analysis of legal documents (predictive justice) or based on the data provided to them about a case. In the latter case they serve as a type of virtual judge (simulative justice). It is much less likely that the risks of assimilation of the machine with a person could arise in this field. On the other hand, when public institutions or private operators make conversational agents available to the public to deliver legal information or to facilitate the resolution of disputes, they will have to make sure that the language and, if applicable, the vocal expression are sufficiently distant from a human conversation, so that the conversational agent is not mistaken for a person.

VIRTUAL JUDGES

Some countries or jurisdictions are considering the implementation of virtual judges in courtrooms, which have themselves moved to a virtual space. This prospect seems distant because of the technical challenges. It also appears to be inconsistent with the European Union's Article 22 of the GDPR, but at least one state has considered it.[15] In France, such implementation would be discarded on the basis of the same European regulation and in view of the provisions of articles 47 and 120 of the law of January 6, 1978, known as the Data Protection Act that claims a human guarantee being necessary for the formation of a legal decision. There are, however, Chinese[16] and Canadian examples. Yet the creation of a virtual assistant to the judge could be reasonably considered if a state wanted to facilitate preparatory hearings using a conversational agent to gather answers to certain questions. Would it be designed in such a way that it could not be confused with a human judge? Or, if it were to resemble a human, how would its appearance and its oral expression be defined?

The CNPEN emphasizes that a chatbot should never be perceived by the user as a responsible person, even by projection. In general, the appropriate attitude is neither to give free rein to anthropomorphization nor to wish to eliminate it at all costs, but to define limits in concrete applications. Anthropomorphising as far as to project responsibility on a chatbot poses a major risk for society, i.e. novel uncontrollable agents may emerge who will not obey the existing norms and conventions. Therefore, it is necessary to continuously monitor the development and diffusion of "virtual characters" with clarity and vigilance. This may eventually lead to a regulatory measure.

RECOMMENDATION 1

REDUCE THE PROJECTION OF MORAL TRAITS ON A CONVERSATIONAL AGENT

To reduce the spontaneous projection of moral traits on the conversational agent and to limit the attribution of responsibility to such systems, the manufacturer must limit its personification and inform the user about biases that may result from the anthropomorphization of the conversational agent.

THE CALIFORNIAN LAW ON THE DEVELOPMENT OF CHATBOTS

California has passed the Bolstering Online Transparency Act (California Senate bill 1001), a law that requires the developers of all conversational agents to reveal their artificial nature. This was preceded by a debate that caused great concern to the major Internet platforms. As of July 1, 2019, the law requires the developers of conversational agents used to sell a product or convince voters to reveal to their users that they are conversing with a machine. There is currently no comparable obligation in the French law. Article 52 of the proposed European regulation on artificial intelligence[17] claims that providers of conversational agents must, risking the penalty of a fine, ensure that their users are informed that they are communicating with an artificial intelligence system. It is expected that this obligation will be exempted in a few specific cases, for example in chatbots that help detect, prevent, investigate, and prosecute crimes.

RECOMMENDATION 2

AFFIRM THE STATUS OF A CONVERSATIONAL AGENT

Any person communicating with a conversational agent must be informed in an appropriate, clear and intelligible way that they are conversing with a machine. The format and timing of this communication must be adapted on a case-by-case basis.

[15] https://e-estonia.com/artificial-intelligence-as-the-new-reality-of-e-justice/
https://www.alexsei.com/

[16] Chris Young, China has Unveiled an AI Judge that Will 'Help' With Court Proceedings, Interesting Engineering (Aug 19, 2019) available at https://interestingengineering.com/china-has-unveiled-an-ai-judge-that-will-help-with-court-proceeding.

[17] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

2. IDENTITY OF CONVERSATIONAL AGENTS

A name may be given to the chatbot by the user or the developer.

From the developer's perspective, naming a machine can help to better perform its function, e.g. in customer support, personal assistance, or entertainment. If the name chosen by the developer represents a brand or a company, this will help to enhance the connection between the user and the manufacturer. On the other hand, a public institution that uses chatbots to expedite its services may voluntarily choose not to give it a name, in order to emphasize the impartial nature of its authority.

Naming can be also performed by the user of a conversational agent, somewhat similarly to the way children name toys. Naming allows the user to individualize the chatbot and to give it a "virtual character," often a "friend" or a "partner." The act of naming is meant to be understood as an expression of freedom enjoyed by the human user, who can parametrize their "own" chatbot as they wish.

The possibility of choosing, deleting, or changing a chatbot's name gives the user an illusion of control and mastery over the "virtual character." Explicitly or implicitly, the users conceive of themselves as co-creators of a unique "individual." In some applications, e.g. "virtual companions," users may perceive the chatbot as an autonomous being somewhat analogous to a pet, although this impression is merely an illusion.

Naming has multiple significance: it creates a convenient shortcut for the user or an illusion of the chatbot's autonomy, but it also has a long cultural history. Certain interpretations of mythological stories present Adam as co-creator of Nature through his action of naming all living beings, or Prometheus as the creator of language. Such interpretations make ethical and anthropological conclusions based on narratives. For example, according to some interpretations, the story of the Tower of Babel highlights the excessiveness of human action without condemning humanity for its own ambition to create. It therefore opens a debate on the meaning of *giving a name*. For Gilbert Simondon, individuation of technical objects (naming being an example of individuation) endows them with dignity.[18] Then destroying such an object or getting rid of it becomes morally problematic. For conversational agents, this reasoning would amount to questioning the decision to reset the chatbot's memory or to erase its history. Despite its apparent convenience, the act of naming a chatbot can thus lead to ethical tensions.

The choice of the name is as important as the act of naming itself. The tendency to name things is quite natural: it is unfeasible to forbid users from naming the machines they engage with. On the other hand, it is important to recognize the confusion of status that can arise from the name. Depending on the application, such confusion may prove to be harmful and must be examined on a case-by-case basis. Whether the given name is human (e.g. "Sophia" or "Albert") or non-human (e.g. "R2D2"), the act of naming takes part in the dynamic of anthropomorphization and individuation with regard to the chatbot.

If the chatbot's given name has grammatical gender, it can have a significant effect on users. Gendered naming, like other linguistic elements, notably personal pronouns, may lead to anthropomorphism and gender bias. Limits need to be defined for the creative liberty of the users, especially for the assignment of gendered proper names to a conversational agent.

If a conversational agent uses its assigned name in a dialogue, the question of self-reference arises: to whom or to what does this name refer to exactly? The chatbot lacks corporeality but it assumes its virtual "identity" and shapes the user's perception of reality. The use of the pronoun "I" by a conversational agent is conceptually troubling, but avoiding it in all contexts through constraining the chatbot's vocabulary would be equally problematic.

Indeed, when a conversational agent refers to its role in the first person, which creates an analogy to the human role, it can help its function: "I am your doctor," "I am here to help you," "I will give you advice," etc. These roles, albeit merely stated by the chatbot, employ the existing notions of human professions and responsibilities. Despite the utility of this spontaneous projection, the use of the pronoun "I" must not be used to assign competences and responsibilities to a conversational agent.

RECOMMENDATION 3

CONFIGURE THE IDENTITY OF CONVERSATIONAL AGENTS

To avoid bias, especially gender bias, the settings by default of a conversational agent for public use (name, personal pronouns, voice) should be made in an equitable way whenever possible. In the case of personalized conversational agents for private or domestic use, the user must be able to modify the default settings.

[18] Gilbert Simondon, *Du mode d'existence des objets techniques*, Paris, Aubier, 1958.

3. ABUSING A CONVERSATIONAL AGENT

General purpose voice assistants occasionally get insulted by their users. However, it is a complex problem for the conversational agent to define or recognize the insults in the flow of the dialogue. The internet is teaming with examples of derogatory speech by users seeking entertainment or people who were dissatisfied with the services they received. The client facing services of many businesses confirm that it is a widespread practice. Like in everyday life, an insult can only be understood and evaluated within its context and circumstances. In addition to that, the case of conversational agents brings in their dialogue strategies that are defined by the developer, and they might provoke an insult (malicious incitement), or they can try to prevent it.

The problem of insulting behaviour while engaging with a conversational agent has been already observed in 2000s.[19] For example, the authors of XiaoIce, a virtual companion initially developed by Microsoft China in 2014 and available in several languages, acknowledge that solving this problem posed a real design challenge.[20]

According to some responses to the CNPEN call for contributions (Appendix 2), there is nothing immoral about insulting or abusing a computer. A chatbot being only a computer program, devoid of understanding, conscience and sensitivity, does not differ from a car or a refrigerator. Insulting or abusing it would therefore not be morally reprehensible.

Other contributors expressed the opinion that insulting a chatbot is a morally degrading act for the person who does it. This is exactly because the user is not talking to another person when talking to a chatbot. The user is the only one who is aware of the content of the conversation. In a way, the user "receives" their own input by a mirroring effect. The ethical argument from the "negative transfer"[21] states that users may experience defective moral development when they get accustomed to the liberty of using demeaning phrases with human interlocutors.

The latter position among the contributions is based on a fundamental observation that the use of language cannot be completely dehumanized or completely desocialized. The very use of language, which proceeds with conscious thought and judgment, causes a projection of human traits on the machine. This projection shows that we cannot morally neutralize the conversation simply by separating the chatbot's language from the meaning, associations, and judgments conveyed by human language.

The insults that users address to chatbots bring forth the limits of anthropomorphization conversational agents. They push against the boundary of individual morality, prevalent in the private sphere, and the collective morals, evident in the public sphere. A user may be surprised by the conduct of a chatbot, for example a chatbot may not respond to any of the insults, or an insult tolerated in the private sphere may appear embarrassing when uttered by a chatbot in public.

RECOMMENDATION 4

ADDRESS THE INSULTS

If situations in which the user engages in insulting a conversational agent cannot be avoided, the manufacturer should anticipate them and define specific response strategies. In particular, the conversational agent should not respond to insults with insults and should not report them to an authority. Manufacturers of chatbots that use machine learning techniques should exclude such phrases from the training data.

RESEARCH QUESTION 1

AUTOMATICALLY RECOGNIZING INSULTS

It is necessary to develop methods for the chatbots to automatically detect inappropriate language, especially insults.

4. MANIPULATION BY A CONVERSATIONAL AGENT

Some applications use conversational agents to influence their users through the architecture or language of the dialogue. Manipulation by a conversational agent can be direct (including inaccurate or skewed information) or indirect, using the "nudging" strategies.

NUDGING

Nudging is a term that means a suggestion, an incitement or a boost. Nudging deals with inconspicuously pushing the person in a "good" direction. For example, a chatbot could encourage a user to do more sports by referring to the example of their athletic friends. The concept of moderate and non-invasive incentives that do not prohibit or restrict a person's options was first described by the economist Richard Thaler.[22]

From an ethical standpoint, it is necessary to determine who the goal of nudging serves – the developer, the user, or the community. The intentional decision to manipulate or deceive a user must be assessed in view of its purpose. For example, a conversational agent could refuse to order a fast-food meal because a user has not done enough physical exercise. But in that case, there is a dilemma – should we make the chatbot tell a lie ("they ran out of supplies") or provide a detailed explanation, including medical recommendations that contraindicate the user's choice ?

[19] Brahnam, Sheryl. 2005. Strategies for handling customer abuse of ECAS. Abuse: The Darker Side of Human Computer Interaction, pages 62–67.

[20] Li Zhou, Jianfeng Gao, Di Li, Heung-Yeung Shum. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. arXiv:1812.08989 v2 (2019)

[21] Ph. Brey, The ethics of representation and action in virtual reality", Ethics and Information Technology 1: 5–14, 1999.

[22] R.H. Thaler and C.R. Sunstein. Nudge: Improving Decisions About Health, Wealth, and Happiness. Penguin Books, 2009.

If a recommendation system employs manipulative means, it must ethically consider a balance between the well-being of a generic user, constructed through a statistical approach that addresses the largest number of people (e.g. following a balanced diet or doing physical exercises), and the swell-being of a particular person, i.e. the particular user. It is the responsibility of the developer to define this balance. If most users agree that the intended purpose is consistent with their well-being, it will mitigate the negative judgment associated with manipulation and deception.

But manipulation remains morally problematic regardless of its utility. While the use of a nudging is not necessarily morally wrong, deception infringes on users' autonomy and freedom if it is not clearly presented to them. At a societal level, the use of nudging and deception can lend itself to political manipulation. This calls for enforcing strict limits to manipulation independently of the utility and context of application.

LEGAL MEASURES CONCERNING MANIPULATION

The European Union law is moving to include measures to regulate manipulation by digital systems. Article 5 of the proposed regulation on artificial intelligence forbids the placing on the market, putting into service or use of an AI system that implements subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm. The same section prohibits artificial intelligence systems that exploits any of the vulnerabilities of a specific group of persons in order to influence their behaviour and cause harm to them. Article 71 of the text defines the penalties for disregarding these prohibitions. In addition, a person who has suffered harm may seek financial compensation. Moreover, the Council of Europe has called for "open-ended, informed and inclusive public debates with a view to providing guidance on where to draw the line between forms of permissible persuasion and unacceptable manipulation" (Declaration of the Committee of Ministers on the manipulative capabilities of algorithmic processes, February 13, 2019).

RECOMMENDATION 5
INFORM ABOUT DELIBERATE MANIPULATION

If the design of a conversational agent includes the capacity to influence user behaviour as part of its intended use, the manufacturer must inform the user about the existence of this functionality and obtain consent. The user must be able to withdraw this consent at any time. The manufacturer of a conversational agent that may influence user behaviour must inform the users about the nature and the origin of messages formulated by the chatbot as well as its communication methods. The manufacturer must ask users to exercise vigilance before sharing such messages.

RECOMMENDATION 6
AVOID MALICIOUS MANIPULATION

The manufacturer must seek to avoid the technical possibility of malicious manipulation or threats issued by the conversational agent. The user must have the ability to flag unwanted expressions, leading to a modification of the conversational agent by the developer.

VIRTUAL INFLUENCERS

Conversational agents known as "virtual influencers" are increasing their presence on social networks, such as Twitter or Instagram. These virtual influencers imitate humans and manipulate other users, most worryingly by spreading misinformation or disinformation.

One of the virtual influencers, Lil Miquela, created in 2016, dwells on Instagram's social networks and currently has over three million followers. Lil Miquela is a chatbot and an animated character that takes on the narrative of a young woman that appears as a muse of famous music brands or finds herself in the company of real world celebrities. This virtual influencer often pleads against racism, sexism, and police violence, and even talks about "sexual abuse" of which she was supposedly a victim. She plays on the empathy and ambiguity of her character to attract the interest of Instagram users.

A chatbot that is telling lies is a particularly complicated case. Not all lies are morally wrong. Other moral principles, such as modesty, generosity, usefulness, justice, or peace can motivate human beings to lie. An example of socially acceptable lying is sometimes called "white lying," which does no harm to others. Another one might be lying by omission of details. When confronted with sensitive questions (for example, "Do I have cancer?"), should the chatbot refuse to answer and refer to a human interlocutor?

LEGAL MEASURES REGARDING LYING

The legal problem of lying and misinformation concerns the responsibility of the manufacturer of a conversational agent, and not the conversational agent itself since artificial intelligence does not constitute a legal person. The legal texts on this subject are quite limited because they essentially relate to the formation of the contract. Article 1104 of the Civil Code imposes a requirement of good faith in contractual relations. Article 1112-1 provides an obligation to disclose information to the party who knows information which is decisive for the consent of the other party, when the latter is unaware of this information or trusts his co-contractor. Moreover, for consent to be informed, it must not be obtained by fraud, on pain of rendering the contract void (article 1137).[24] Moreover, unfair commercial practices aiming at deceiving the consumer are prohibited by the Consumer Code (article L. 120 and following). The abuse of weakness is sanctioned by the criminal code (article 223-15-2).

[23] See the bulletin no. 2 of CNPEN
(https://www.ccne-ethique.fr/fr/actualites/comite-national-pilote-dethique-du-numerique-bulletin-de-veille-ndeg2)

[24] Article 1137 provides that fraud "is the fact that a contracting party obtains the consent of the other party through manoeuvres or lies. Fraud (also results from an intentional concealment of information by one of the contracting parties which it knows is decisive for the other party."

A conversational agent is rarely equipped with a function to evaluate the statements it utters in terms of their truth or falsity. Even when this function is present, it is a formal operation that does not allow to evaluate the content of the statements. If the data that the conversational agent relies on a specific database (phrases spoken by the user or extracted from the Internet) contains false sentences, the chatbot will not easily identify it. For a chatbot, truth is only a result of an algorithmic evaluation.

The conversational agent has no "human" understanding of the meaning of the statements it produces. Therefore, when a chatbot utters a lie, this process is not the result of an evil intention or moral choice. It is performed without awareness, simply carrying out the programmed functions with the data available to the program. These arguments point towards the absence of moral judgment regarding a chatbot that tells a lie. However, if a chatbot engages in lying, the responsibility of its manufacturer should be assessed in terms of the measures that they took to limit the possibility of manipulation or, if no measures were taken, it should be noted that the manufacturer has not anticipated these issues during development or when selecting the data.

RESEARCH QUESTION 2

STUDYING LIES TOLD BY A CONVERSATIONAL AGENT

The empirical significance of lies told by a conversational agent requires further study. It is also necessary to avoid the projection of moral traits on a conversational agent via a narrative of its actions explicitly different from a narrative that characterizes lies told by humans.

5. CONVERSATIONAL AGENTS AND VULNERABLE GROUPS

Conversational agents can encounter vulnerable people or people in a position of vulnerability[25] in a variety of fields, including health and education. Dialogues with conversational agents are recorded in the form of "logs." When a chatbot converses with vulnerable people or people in a position of vulnerability, these logs can contain sensitive information. The collection of logs may be necessary to fulfil the purpose of the system. It is important to include the collection, storage and use of these traces in a legal framework.

We can distinguish a few special cases of the use of conversational agents by vulnerable people or people in a position of vulnerability.

For example, children are naturally inclined to talk to inanimate objects such as toys or stuffed animals.[26] An even stronger attachment is formed when they can respond and interact, like the Furby.[27] Unlike traditional toys, a gadget with a chatbot can have a verbal and emotional influence on the child.

RECOMMENDATION 7

SET UP A FRAMEWORK FOR THE USE OF CHATBOTS IN TOYS

In the toy industry, particularly with regard to toys for young children, public authorities must assess the effects of user interactions with chatbots having a potential to influence children's behaviour. Public authorities must regulate the use of such conversational agents with regard to the impact on children's linguistic, emotional and cultural development.

In the field of education, chatbots can help students understand difficult concepts. For example, some distributors of voice assistants provide their users with instructions of how to use their systems for educational purposes.[28] However, learning while interacting with a chatbot is not equivalent to learning with a human educator. For instance, a conversational agent could teach a student to pronounce better in a foreign language by accurately pointing out mistakes and training on repeating a number of sounds, but it can also happen that a language that is taught will have a limited or inadequate vocabulary compared to the one that is naturally learned. In particular, a chatbot could teach its student to use sentences that are too literary, lacking in stylistic sensibility, because it applies the same conversational strategies without awareness of context or the status of the conversation. Moreover, a conversational agent could end up teaching the student to pronounce sounds inhumanly, based on statistical averages on tone, energy and rhythm of voice calculated by a machine that would not resemble a human.

Conversational agents are often used in the education of autistic children or in the rehabilitation of disabled people, thanks to the machine's ability to repeat instructions a large number of times, which is not always the case with a human educator. Unlike a human educator, the machine does not "get impatient " and does not take on impatience from interacting with vulnerable people. Technically, this requires special attention because machine learning based on the imitation of human behaviour, i.e., the set of data based on human educators, also risks importing these undesirable traits into the conduct of the chatbot.

RESEARCH QUESTION 3

ASSESSING THE UNFORESEEN EDUCATIONAL EFFECTS OF CHATBOTS

In education, public authorities need to evaluate the consequences of interactions between pupils and chatbots, especially when vulnerable or young children are involved.

[25] Vulnerable individuals are understood as minors or adults, whose vulnerability is related to age or to physical or mental disabilities, disorders, or conditions (e.g., autism, Alzheimer's disease, phobias, anxiety, depression, etc.).

[26] https://www.hadopi.fr/sites/default/files/sites/default/files/ckeditor_files/2019_05_24_Assistants_vocaux_et_enceintes_connectees_FINAL.pdf

[27] https://www.whoson.com/chatbots-ai/hey-furby-did-the-popular-90s-toy-influence-the-chatbot-timeline/

[28] Voir https://aws.amazon.com/fr/education/alexa-edu/ or https://dialogs.yandex.ru/store/categories/education_reference

[29] An example of learning with the Siri chatbot was described as early as 2014, viz. J. Newman, "To Siri with Love," New York Times, 18 Octobre 2014.

In the field of health, conversational agents contribute to the set of digital tools that help answering the recurrent problems of this sector: access to care, shortage of doctors, medical deserts, repetitive tasks that get unloaded on the supporting staff. Personal assistance through chatbots for managing pathologies, monitoring, or medical advice can involve a patient's private life in a major way; however, it has become a necessary part of the medical field.

The use of chatbots for medical advice is most often conducted through smartphone applications. Implemented in such a way, the chatbot can give health advice directly to the user or collect health information to be transferred to a professional. A chatbot can also answer questions of patients who strive to take action and become responsible for their health. Repetitive care tasks, such as informing the patient before and after an operation or educating and monitoring diabetic patients, are increasingly assigned to chatbots.[30] Also, because these are intimate issues that people may be reluctant to share with other people, there are chatbots dedicated to sexuality, accessible exclusively on personal smartphones and aimed at young adults and teenagers.[31]

In psychiatry, chatbots are used to conduct prevention, diagnostics and follow-up interviews. In this field, chatbots are increasingly used as platforms for personal transformation to rediscover oneself, one's history, and one's relationship with others. Until recently, automated systems in healthcare performed only simple and repetitive tasks in the form of a questionnaire. The arrival of elaborate chatbots that mimic the behaviour of human psychiatrists can be a source of new ethical tensions. Usually, psychiatrists spend the first few interviews to gain the patient's trust. However, some people find it easier to trust a chatbot than a human.[32] This effect comes the patient's perception that the chatbot is neutral and does not express moral judgments. Patients feel it does not provoke feelings of guilt, as can happen with a human interlocutor. This becomes all the more true for vulnerable patients that feel like the chatbot's voice is embodying a "caregiver." This feeling is linked to the degree of chatbot's personalization. Some people provide information to conversational agents more easily than to humans. This information can eventually be used by a human doctor. Finally, medical chatbots are available without downtime, at all hours of the night, and can help reassure a patient.

Conversational agents can be used to promote healthy habits (like diet or sport) in their users and to increase their health and well-being. This is especially true for conversational agents that are connected to devices collecting physiological data, like smartwatches that measure heart rate, temperature, electrodermal response, oxygen levels, etc. A conversational agent that makes use of such data and explicitly acknowledges it to the user, can increase or decrease their engagement. However, the biological feedback (biofeedback) of body measurements (Quantified Self) together with an interpretation to the collected digital data by the chatbot is likely to cause anxiety or a stressful state of mind. These systems can influence users or make them dependent. The more vulnerable the person, the greater the effect is likely to be.

LEGAL PROTECTION OF VULNERABLE GROUPS

Article 5 of the proposed European regulation on artificial intelligence prohibits the use of any artificial intelligence system that exploits the vulnerability of a group of individuals to influence the behaviour of any of these individuals and cause harm to them.

RECOMMENDATION 8

RESPECT VULNERABLE INDIVIDUALS

In the case of a dialogue between a conversational agent and a vulnerable individual, the manufacturer of the conversational agent must seek to respect the dignity and autonomy of this person. In particular, medical chatbots must be designed to avoid excessive trust in these systems by the patient and to ensure that any possible ambiguity between the conversational agent and a qualified physician is eliminated.

RECOMMENDATION 9

ANALYSE THE EFFECTS OF CONVERSATIONAL AGENTS USING PHYSIOLOGICAL DATA

In the case of conversational agents with access to physiological data ("Quantified Self"), designers must study the risk of creating dependency. Public authorities must supervise the use of these systems with regard to their impact on personal autonomy.

[30] Klonoff, D. C., & Kerr, D. (2016). Digital Diabetes Communication: There's an App for That. Journal of Diabetes Science and Technology, 10(5), 1003–1005.

[31] https://roo.plannedparenthood.org/onboarding/intro ; J. Brixey, R. Hoegen, W. Lan, J. Rusow, K. Singla, X. Yin, R. Artstein, and A. Leuski, "SHIHbot: A Facebook chatbot for Sexual Health Information on HIV / AIDS,"

Proceedings of 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue, August 2017, pp. 370-373.

[32] Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. Can J Psychiatry. 2019;64(7):456-464. doi:10.1177/0706743719828977

[33] Quantified self means the practice of "self-measurement" and refers to a Californian movement that encourages knowing oneself by measuring data related to one's body and activities (https://www.cnil.fr/fr/glossaire).

6. LABOUR AND CONVERSATIONAL AGENTS

Conversational agents can serve different functions at a workplace and can be easily implemented in collaborative digital platforms. Chatbots can be used to assign tasks to collaborators, monitor the progress of a project, remind the team of norms, procedures, and goals, help understand different roles, contributions, and areas of expertise of collaborators, set appointments, monitor completed and ongoing tasks, make lists of assignments agreed during meetings, or they can even train employees.[34]

The use of chatbots can facilitate the sharing of information between human collaborators and optimize workload to achieve project deadlines. Such chatbots are developed and implemented as assistants that are available at all times. By projection, they are sometimes understood as virtual collaborators.

The use of conversational agents in teams of professionals can have organizational effects that vary across industries, but include increased informational and emotional load, potential decrease of direct interactions between human collaborators, the rise of impersonal mediation, the feeling of unity or, conversely, the isolation of workers, effects on employee morale and autonomy, as well as problems of equality and merit recognition within companies. There are currently no systematic studies to assess the validity of these concerns.

When a conversational agent performs a task in a company, it is problematic to determine who controls it and who is responsible for its utterances. These systems must be technically evaluated to avoid social discrimination. It is also necessary to analyse the biases and possible gender or age inequalities propagated through the implementation of chatbots. The company must clearly declare the purposes and internal procedures involving conversational agents to respect worker rights.

LEGAL ISSUES SURROUNDING E-RECRUITMENT SYSTEMS

Chatbots are used by human resources managers for recruitment as well as for career follow-up and employee training. Legal regulations are starting to be applied to implementations in human resources. Article 6 of the proposed European regulation on artificial intelligence and its annex III consider recruitment systems to be high-risk. Therefore, legal compliance is mandatory ex ante, including risk management processes, monitoring, bias detection and correction, technical documentation, event logs, user consent, human oversight, robustness, security, accuracy, and proportionality.

RECOMMENDATION 10

DEFINE RESPONSIBILITIES FOR THE USE OF CONVERSATIONAL AGENTS IN THE PROFESSIONAL ENVIRONMENT

The manufacturer should envisage control and audit mechanisms to facilitate the attribution of responsibilities for the functioning or malfunctioning of a conversational agent in the professional environment. In particular, the manufacturer must study the chatbot's secondary or unintended effects.

RESEARCH QUESTION 4

STUDYING THE EFFECTS OF CONVERSATIONAL AGENTS ON THE ORGANIZATION OF LABOUR

Public authorities and private enterprises should support empirical research on the effects of conversational agents on the organization of labour across different industrial sectors.

7. CONVERSATIONAL AGENTS AND THE MEMORY OF THE DEAD

With the recent developments in chatbot technology, it has become possible to create conversational "digital twins"[35] that replicate the speech and language patterns of deceased individuals. A chatbot is able to converse by imitating a deceased individual by a learning process based on conversational data collected from this person.[36] Even though the "deadbot" technology is not yet well known by the public, there are already several companies working in this field, e.g. HereAfter AI[37] and MBC Design Center.[38]

[34] See the report of the Global Partnership on Artificial Intelligence (GPAI) Working Group on the Future of Work, https://gpai.ai/fr/projets/avenir-du-travail/pmia-groupe-de-travail-sur-l_avenir-du-travail-novembre-2020.pdf

[35] In general, a "digital twin" is a virtual model that represents a real object or system as closely as possible to enable simulations and the evaluation of the effect of modifications or actions performed on it. Research is being done to develop digital twins in several fields, including medicine (for organs or the human body).

[36] French law includes articles that deal with the processing of personal data that remains on the internet after the death of the person (the notion of digital death). These provisions are not specific to conversational agents and do not consider the possibility of creating "digital twins" after death. On the other hand, Article 86 of loi informatique et libertés states that: "any person may give instructions regarding the retention, erasure, and communication of their personal data after their death. The instructions may be general or specific."

[37] https://www.hereafter.ai

[38] https://www.ubergizmo.com/2021/01/virtual-reality-husband-meet-deceased-wife/

Typically, conversational agents do not repeat the training data word by word. They have the ability to generate new phrases that the person being imitated has never uttered in their lifetime. Microsoft has engaged with such applications and has filed for a patent that combines a deadbot and a transformer conversational agent.[39] Deadbots can be purely conversational or include and additional visual imitation, but at least some of these systems hold a very realistic dialogue that can be further enhanced by the chatbot's ability to mimic emotions. A human interlocutor can genuinely experience being in the presence of the person being imitated, even if they are explicitly informed that they are conversing with a machine. In a striking example, a young Canadian man has harnessed the power of the GPT-3 transformer neural network to imitate a dialogue with his deceased girlfriend.[40] This topic has been the subject of passionate reactions from "deadbot" users and the public for several years.[41]

Some respondents find this application fascinating and even consider it as a way to "overcome" or "cheat" death (Appendix 2). Others, frightened by the illusion of extendable life, think that it undermines the respect for human dignity, even if this major moral concept[42] is difficult to define.[43] According to them, the generation of new speech by imitating a dead person should not be allowed after their death and any digital interference with this fundamental element of human nature should be prohibited. The subject of deadbots condenses the technological fantasies and concerns and raises critical questions about our conception of human dignity.

Conceptions of death and its different stages vary with cultures and times. Funeral rites take very different forms depending on the customs (mummification, cremation, burial, etc.) and can extend over several months. Similarly, the posthumous relationship to the bodies and spirits of the dead varies according to religions and cultures. It can even become a cult of the dead. Western literature since Homer and Virgil, including Dante and Molière, also contains numerous examples of dialogues with the dead. In Japanese culture, which is influenced by multiple religious traditions but especially by Shintoism, ghosts or doppelgangers of the dead appear abundantly in literature and in film.

AN EXAMPLE OF A "DEADBOT" IN JAPAN

In 2018, a Japanese company has offered the relatives of a deceased individual to use a humanoid robot equipped with a 3D printed mask imitating the face of this person. The users would interact with this robot through a system that is able to imitate certain traits of the deceased person's personality by using their pre-recorded speech and gestures, but without creating or producing anything that the person had not actually uttered during their lifetime. The lifespan of this machine was designed to be only 49 days, corresponding to the length of the traditional mourning period in Japanese culture.

Photographs and recordings with audio and video already provide a way to recall a person after their death. Chatbots could be considered as yet another step in this path enabled by technology since the invention of writing. But the ability of a conversational agent to generate original outputs that the person they are imitating never uttered during their lifetime requires special attention because it sets deadbots apart from any other technique of remembrance.

Original yet convincing outputs can only be attributed to the person being imitated in a conditional mode: this person could have said such words, even if they had not actually said them. However, their effect is very real. The reaction of the young Canadian man who trained a chatbot with the conversational data of his deceased girlfriend offers a disturbing example: "Most mysterious of all: the chatbot seemed to perceive emotions. It knew how to say the right sentence, at the right time, with the right accent." The responsibility for the verisimilar but invented words is a new ethical and legal problem. It must be analysed in the context of the current debate on artificial intelligence in France and in Europe. Moreover, OpenAI, the owner company of GPT-3, has decided to restrict access to the neural network of the Californian computer scientist who enabled the young Canadian man to create the digital twin of his deceased girlfriend.[46]

[39] https://www.forbes.com/sites/barrycollins/2021/01/04/microsoft-could-bring-you-back-from-the-dead-as-a-chat-bot/

[40] https://www.sfchronicle.com/projects/2021/jessica-simulation-artificial-intelligence/

[41] https://www.liberation.fr/futurs/2017/07/19/un-journaliste-discute-avec-son-pere-decede-grace-a-un-programme-qu-il-a-cree_1584849

[42] In moral philosophy, the concept of human dignity arises from Immanuel Kant's description of the categorical imperative. It is found in Article 1 of the 1948 Universal Declaration of Human Rights: "All human beings are born free and equal in dignity and rights."

[43] The concept of human dignity is legally difficult to define. It is not defined by any normative text and in domestic law takes its meaning from the jurisprudence of the Constitutional Council: it is a principle of constitutional value defined for the first time in the decision n° 94-343/344 of 27 July 1994, so-called "bioethics". It was then applied in criminal law for deciding cases of deprivation of liberty and hospitalization without consent. The Council of State made it a constituent of public order in its decision of October 27, 1995, Commune of Morsang-sur-Orge, n° 136727: the "dwarf-tossing" events in which a small person is thrown by spectators, and which involves using a person affected by a physical handicap as a projectile, undermines the dignity of the human person and, thus, the authority invested with the power of municipal policing could prohibit it despite the fact that protective measures had been taken to ensure the safety of the person concerned, who had given their consent and was remunerated for this service.

[44] https://starts-prize.aec.at/en/digital-shaman-project/

[45] https://www.sfchronicle.com/projects/2021/jessica-simulation-artificial-intelligence/

[46] https://gadgets.ndtv.com/internet/news/openai-chatbot-gpt-3-samantha-shut-down-dilute-jason-rohrer-possible-misuse-2537388

The debate regarding "digital twins" that replicate human beings is still in its infancy. CNPEN would like to draw the stakeholders' attention to the ethical complexity of the issue. The creation of "digital twins" of people – especially when deceased individuals are involved – must ethically analysed in terms of the process itself and of the purpose why such systems can be developed. The ethical analysis must follow along the development at all stages, during the initial conception, data collection with the consent of the person that is the originator of the data, as well as at the implementation stage when the deadbot is in use. It is also necessary to define legal limits to this technology, adapting them as necessary in light of future developments. Consent for the collection of data from the deceased must be framed according to its temporality (before death or after death) and its author (by the person during their lifetime or by their heirs, which by itself is problematic, given the non-transferability of personal data according to current law). As the deceased individual cannot withdraw their consent, the decision to use the "deadbot" lies entirely with the user, unless a specific legal framework is put into place. Another risk resulting from the "digital twins" is to enable impersonating a living or deceased individual through a conversational agent.[47]

A conversational agent that imitates a deceased individual would most often be used by someone who knew the person while they were alive. Even if the user knows that the exchange is only coming from a chatbot, and not from the person, they are not always aware of that. The user can enact their wish to remember the deceased by letting go of the knowledge and by embracing the illusion of their presence. This practice is not too distant from the spiritual seances performed in the past, except that the user is informed that they are interacting with a machine.

In some cases, the use of deadbots could result in impaired judgment, which is cognitively and morally problematic. Note that the chatbot does not understand the meaning or context of its language. Its words could thus cause the "uncanny valley" effect for the interlocutor by either the chatbot uttering offensive words, or, after a sequence of familiar lines, producing a sentence that is totally different from what the person being imitated might have said. In these moments, the user would naturally wonder whether it was a mistake, a misunderstanding, or a meaningful statement. They could undergo a quick and painful psychological flux. This calls for defining the limits of chatbots that simulate the speech of deceased people or present themselves as their "digital twins. "

Respect for the memory and dignity of the dead is a widely shared principle. The development of conversational agents that imitate a deceased individual is already controversial. The society can decide to forbid their development, but it can also regulate its development by legal measures. In the latter case, a specific legal framework must be worked out alongside a set of technical constraints limiting the side effects, most importantly the potential negative effects on the mourning process.

RECOMMENDATION 11

CONDUCT A REFLECTIVE PUBLIC DEBATE BEFORE REGULATING "DEADBOTS"

The legislator should adopt specific regulation concerning conversational agents that imitate the speech of deceased persons after an extensive ethical reflection at the societal level.

RECOMMENDATION 12

SET UP A TECHNICAL FRAMEWORK FOR "DEADBOTS "

The developers of "deadbots" must respect the dignity of the human person, which does not end with death, while seeking to protect mental health of the users of "deadbots". Rules must be defined concerning the consent of the deceased person, the collection and reuse of their data, the operating time of a "deadbot," the vocabulary used, the name given to the chatbot, and the specific conditions of its use.

8. LONG-TERM EFFECTS OF CONVERSATIONAL AGENTS

In addition to the known vulnerabilities discussed in section II.5, there are emerging long-term risks for the users of conversational agents, like the lack of interaction with other humans, cognitive biases, or gullibility. These risks can arise from users' interactions with chatbots that evoke excessive trust in conversational agents due to the different roles they assume (teacher, banker, doctor, or friend). Moreover, the change in behavioural norms that conversational agents bring about is likely to create new personal and collective vulnerabilities. This change is already underway, propagated via user interaction with widespread voice assistants found on smartphones (SIRI or Google Assistant) or on voice activated speakers (Amazon Alexa, Google Home).

[47] See in particular article 226-1-4 of the French Penal Code: "*impersonating a third party or making use of data of any kind to identify a person in order to disturb their tranquillity or that of others, or damage their honour or consideration, is punishable by one year's imprisonment and a fine of €15,000. This offence is punishable by the same penalties when committed on a public communication network online.*"

Considering the medium and long terms, the habitual use of chatbots can have a lasting impact on human language and on the change of behavioural patterns. For example, if chatbots respond with short, linguistically simplistic sentences without any politeness, people may mimic these language traits when they address other people. Such consequences are not certain, and it is necessary to study them prospectively by measuring the lasting impact on users. Interactions with chatbots also have the potential to influence the lifestyles, opinions, and decisions of humans. It is important to raise awareness about of the importance and extent of widespread effects of conversational agents on users' beliefs, opinions, and decisions at all levels of their development – from the engineer to the politician. The performance of language models used in the most recent machine learning systems (e.g., neural networks such as GPT-3 or LamDA) marks a real turning point in the development of chatbot technologies. Today, transformer type neural networks allow to use any dialogue strategy to generate responses. They far surpass the functionality of previous generation connected speakers that were the subject of studies in the humanities and social sciences. These models integrate very large volumes of data collected on the web, cross-reference them, and rewrite them, often in a non-reproducible way, which is tailored to the input, without abstracting the meaning or reasoning like a human being.

The machine only makes calculations. It returns a calculated answer after receiving a question. However, this does not mean that the transformer type neural networks cannot produce sentences which do not reproduce any of those used during learning. A human user may find them original. Users project meanings on these original utterances. This projection complicates the attribution of responsibility for the words spoken by the machine. It means that transformer type neural networks should not be considered as "neutral" because, despite its asemantic character, it takes part in the construction of the ethical and political meaning of its statements.

By memorizing human utterances and actions, conversational agents are able to deduce information about our opinions, decisions, or even worldviews. For example, a chatbot can recall memories that the user has forgotten. Potentially, this can influence people to be more revealing about themselves. In the long term, the very notion of personal intimacy can change under the influence of conversational agents.

GATEBOX

A Japanese company called Gatebox has put to market a chatbot – Azuma Hikari – that plays the role of a "virtual girlfriend" in hologram. It can turn the lights on and off, send SMS, recognize people, and converse with them. An attachment to such an application can lead to a vulnerability rooted in emotional dependence and the intimate relationship that develops over time.

A "guardian angel" chatbot can be, for example, a conversational agent that is designed to always protect its user's personal data. It looks over the privacy of its user.

Daily interactions with such a conversational agent or a "virtual friend" can change the notion of private life and the relationship with other humans. Constant interactions can create a dependency, especially in children whose development largely relies on the relationship with their circle of friends, which now includes chatbots. At the same time, chatbots can offset various social deficiencies and respond to traumas. This therapeutic function of a conversational agent serves the need of people to be reassured and receive answers to their questions. A therapeutic chatbot can act as a role model or an educational reflection of exemplary behaviours. On a societal scale, the long-term effects of these chatbots can produce a significant change in the human condition. The language co-adaptation between human users and conversational agents is the driving force behind this change.

RECOMMENDATION 13

SET UP A FRAMEWORK FOR THE USE OF "GUARDIAN ANGEL" CHATBOTS

To limit paternalism and to respect human autonomy, public authorities must set up a framework for the use of "guardian angel" conversational agents that are designed to protect personal data.

RESEARCH QUESTION 5

STUDYING LONG-TERM EFFECTS OF USING CHATBOTS

Public authorities and private enterprises must invest in research on long-term effects on humans and society of the use of conversational agents. All societal stakeholders must remain aware of the potential future effects of conversational agents on users' beliefs, opinions and decisions, and avoid considering this technology as neutral or devoid of ethical and political significance.

The market for conversational agents is growing rapidly, especially with the help of transformer neural networks (see Section III.). Conversational agents will become increasingly hungry for computing power and memory size due to the use of machine learning technologies to exploit very large databases and to support the continuous adaptation to their users. Thus, the speedy evolution of chatbots raises questions of power consumption, although it is not specific to conversational agents.

RESEARCH QUESTION 6

STUDYING THE ENVIRONMENTAL IMPACT

Public authorities and private enterprises should conduct studies on energy consumption and environmental impact of the technology that enables conversational agents.

[48] https://arxiv.org/abs/2005.14165

[49] https://blog.google/technology/ai/lamda/

[50] For example, Usage et valeur 62 (10/2019) ; Réseaux 2020/2-3 (N° 220-221) ; H. Kempt, Chatbots and the Domestication of AI: A Relational Approach, Springer, 2020.

[51] https://www.gatebox.ai

III. DESIGNING CONVERSATIONAL AGENTS: ETHICAL PRINCIPLES

Conversational agents raise ethical questions about their design. Some questions are common to all chatbots, even those that follow a deterministic algorithm with a limited number of predefined answers. Others are specific to chatbots that deal with emotions and, more broadly, with user behaviour. Still other questions concern chatbots that use adaptive learning or transformer neural networks to process the dialogue. Some major ethical issues of chatbot design were also raised in the call for contributions (see Annex 2) and are grouped under the following five sections.

1. ETHICS BY DESIGN

The notion of "ethics by design"[52] is based on the idea of respecting fundamental values when designing a technical system. Ethics by design can be understood within different theoretical and methodological frameworks[53], namely "value-sensitive design"[54] or "technology assessment."[55] These approaches have been in development for more than three decades. They aim to integrate human values into the design process of technical systems. This does not mean that human values are bluntly translated into computer code. Their integration requires a complex design process that involves coders, entrepreneurs, users, and policy makers. These theoretical approaches provide a method to analyse the redistribution of responsibilities brought about by the spread of artificial intelligence systems, including conversational agents. They also offer a framework for training and education.

The process of evaluation, which contains within its concept an etymological link to the notion of value, is an integral part of "ethics by design." If an ethical framework is formulated in terms of values, it aims at determining the degree of correspondence between them and the way that a system operates. The most obvious example of this is the evaluation of biases involved in the development and training of algorithmic systems that rely on statistical learning from large data sets. An artificial intelligence system should not merely claim to not discriminate against user-groups. The bias must be measured with specific quantitative indicators. A large body of scientific work on bias assessment already exists as a part of the "ethics by design" approach and it includes conversational agents. Several enterprises, including some digital giants,[56] already integrate tools for measuring explicit or implicit biases into the design process for their products.

DESIGN PRINCIPLE 1

"ETHICS BY DESIGN" OF CONVERSATIONAL AGENTS

The developers of a conversational agent must analyse during the design phase every technological choice that may cause ethical tension. If a potential ethical issue is identified, the developers must envisage a technical solution seeking to reduce or eliminate it. They should subsequently evaluate this solution in realistic usage contexts.

RESEARCH QUESTION 7

DEVELOPING THE "ETHICS BY DESIGN" METHODOLOGIES FOR CHATBOTS

Public authorities should support research to elaborate the "ethics by design" methodologies suitable for the development of conversational agents.

2. BIAS AND NON-DISCRIMINATION

The sentences produced by a conversational agent may contain biases. For instance, a corpus of voice profiles may consist entirely in adult voices when the system is developed to at least in part interact with children, or a corpus of text may statistically use female pronouns more frequently. Even though algorithms can be used positively to identify these biases, they also import social or historical biases. The system will keep reproducing biases unless it is equipped with specific modules to offset them, which already presupposes prior knowledge of these biases and the ability to correct them. However, we may not know all the biases in advance. The presence of biases in the behaviour of conversational agents is a major source of ethical conflicts or even blatant discrimination – one person could be treated less favourably than others with regard to age, sex, gender, handicap, or skin colour, when applying for a job, housing, or other goods.[57] They can also lead to indirect discrimination – for example, the early job candidates could be disadvantaged at an interview if the parameters of the chatbot evolve as a result of adaptive learning, influenced by the data of previous candidates.

[52] See the deliverables of Horizon-2020 projects SIENNA and Sherpa funded by the European Commission.

[53] J. van den Hoven et al. (eds.), Handbook of Ethics, Values, and Technological Design. Springer, 2015.

[54] B. Friedman and D.G. Henry. Value Sensitive Design. Shaping Technology with Moral Imagination. MIT Press, 2019.

[55] A. Grunwald et R. Hillerbrand. Handbuch Technikethik. 2e éd. J.B.Metzler, 2021.

[56] R. K. E. Bellamy et al., «AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias,» IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 4:1-4:15, 1 July-Sept. 2019, doi: 10.1147/JRD.2019.2942287.

[57] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

DESIGN PRINCIPLE 2

REDUCE LANGUAGE BIAS

To reduce language bias and seek to avoid discrimination, especially cultural discrimination effects, developers must implement a technical solution at three levels: in the implementation of the algorithm, in the selection of optimization parameters, and in the choice of training and validation data for the different conversational agent modules.

LEGAL AND TECHNICAL TREATMENT OF BIAS AND DISCRIMINATION

In 2019 and 2020, the Defender of Rights and CNIL have focused on the risks of discrimination that can result from algorithmic biases.[58] A debate is also underway at the European level with a goal to adapt a regulatory framework to mitigate such risks. In 2020, the Council of Europe recommended that developers, manufacturers, and service providers should avoid any potential bias, including unintentional or hidden bias, as well as the risks of discrimination in the new Convention 108 guidelines. In the resolution of February 2019, the European Parliament claimed that "outputs should be reviewed in order to avoid all forms of stereotypes, discrimination and biases, and where appropriate, make use of AI to identify and correct human biases where they might exist. "

The proposed European Artificial Intelligence Regulation, published by the European Commission on April 21, 2021, names measures to limit discriminatory biases and employs the notion of human oversight as the key to fighting them. The proposed regulation stresses that training, validation, and test datasets must be subject to appropriate data governance and management practices to mitigate possible biases. It is not specified how systems will be tested for such biases. Should they be benchmarked against the equality of opportunities, equality of outcomes, or other criteria?

3.TRANSPARENCY, REPRODUCIBILITY, INTERPRETABILITY, AND EXPLAINABILITY

The principles of transparency, reproducibility, interpretability, and explicability are essential to "ethics by design," even if tensions can appear among these principles in concrete applications. The implementation of these principles depends on the context and must be understood with respect to the principle of proportionality and respect for fundamental rights.

Among other meanings, the transparency of a system implies that its functioning should not be opaque or incomprehensible by its user. In the case of a conversational agent, the issue mostly revolves around the traceability of the chatbot's responses.

DESIGN PRINCIPLE 3

DECLARE THE CHATBOT'S PURPOSE

The developer must ensure that a conversational agent clearly declares its purpose to the user in an easily understandable way at an appropriate moment, for example at the beginning or at the end of each conversation.

DESIGN PRINCIPLE 4

TRANSPARENCY AND TRACEABILITY OF THE CHATBOT

In compliance with the GDPR, a conversational agent should be able to save parts of the conversation (the extent of which needs to be defined) for evidential purposes or to satisfy security requirements. This need creates a tension with the protection of personal data. Chatbot architecture, used data, and dialogue strategies should be made available for audit and legal proceedings if needed. This recommendation may result in a regulatory measure to define the precise application terms.

DESIGN PRINCIPLE 5

PROCESSING DATA COLLECTED BY CONVERSATIONAL AGENTS

Following the existing example of health data, it is necessary to develop ethical and legal rules in compliance with the GDPR for the collection, storage, and use of linguistic data resulting from the interactions with conversational agents.

In their current iterations, chatbots can be strongly personalized by creating their individual replicas. For example, the history of conversations between a conversational agent and a patient, including the knowledge about the patient's state of mind and beliefs, can be used to improve treatment. The method to predict the psychological, economic, or other traits of the user during the dialogue with a chatbot is called profiling or behavioural analysis.

LEGAL MEASURES REGARDING PROFILING

Article 4 of the GDPR defines profiling as any form of automated processing of personal data that consists of using that data to evaluate certain aspects of an individual, including analysing or predicting issues related to work performance, economic situation, behaviour, etc. Decisions resulting from profiling are governed by Article 47 of the French Data Protection Act and Article 22 of the GDPR, as long as they are likely to have an effect on the individual. According to Article 22 of the GDPR, "the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." There are three exceptions to this provision, namely the explicit consent of the subject, the existence of a contract making it necessary for automated decision making, and specific legal provisions authorizing an automated decision. Furthermore, under Articles 13– 2 f) and 14– 2 g) of the GDPR, individuals who are subject to a fully automated decision must be informed, at the time of collection of their data and at any other time about "the existence of automated decision-making [and] *meaningful information about the logic involved.*"

[58] https://www.defenseurdesdroits.fr/fr/communique-de-presse/2020/05/algorithmes-et-discriminations-le-defenseur-des-droits-avec-la-cnil

DESIGN PRINCIPLE 6

INFORM ABOUT THE FEATURES OF CONVERSATIONAL AGENTS

In the interest of transparency, the user should be informed in an appropriate, clear and intelligible manner, either orally or in writing, of the data collection, the adaptive features of the conversational agent, the data it collects during use, and profiling.

Conversational agents that use statistical learning present a reproducibility problem as well. A model obtained by statistical learning always returns results that are close in the computational space, but they may differ from the user's point of view, although the system parameters and input data remain the same.

RESEARCH QUESTION 8

STUDYING REPRODUCIBILITY OF THE CHATBOT'S LANGUAGE

Reproducibility requires storing data and involves a characterisation of the right amount of repetition in the chatbot's language. These issues need to be studied.

Although most of the digital stakeholders agree on the importance of the principles of explicability and interpretability, the meaning of these principles seems to vary according to the source. In general, they represent the extent to which an observer can understand the machine's decisions and their reasons. Even though all technological decisions are implemented by an algorithm's developer, this does not mean that the developer intentionally and consciously participates in all of the decisions made by the algorithm. In machine learning systems, the causal chain that leads to a decision is opaque by construction. Research in the field of explainable AI (xAI) aims at establishing explanatory heuristics for these systems.

Pragmatically speaking, the explicability of a conversational agent presupposes the implementation of technical solutions that allow the user to understand the chatbot's responses by recognizing a coherent "reasoning" behind them. This is despite the fact that the conversational agent does not actually think. A chatbot does not "understand" the meaning of the sentences it generates or gets as inputs. It has no common sense. It is thus prone to formulating sentences that do not correspond to any human reality (e.g. "lyrical milk"), answering without appreciating the context (e.g. "How are you?" – "It's a nice day"), or using inappropriate vocabulary. The immediate effects of such distorted dialogue on the user can be significant. It can evoke strong emotional reactions, a breakdown in understanding, a termination of the dialogue, or a disconnection from the system. All these effects call for the responsibility of the developers.

DESIGN PRINCIPLE 7

PROMOTE EXPLAINABLE CHATBOT BEHAVIOUR

Developers must devise solutions to facilitate understanding of the chatbot behaviour by the users.

4. AFFECTIVE INTERACTION WITH HUMANS AND SELF-ADAPTATION

Some conversational agents implement modules for predicting the emotional, attentional, or intentional behaviour of humans. They can also simulate affection in their written and oral replies. Moreover, the affective information can be used by the dialogue system to choose a response strategy. It can be very well integrated in artificial neural networks, especially in transformers. Many examples have been cited in sections II.7 and II.8, e.g. conversational agents that mimic a virtual girlfriend (Gatebo) or a dead person ("deadbot").

Attempts to develop technologies related to human emotions, called affective computing, date back to the work of Rosalind Picard, a researcher at the Massachusetts Institute of Technology, who published an article laying the foundations of this new discipline in 1995. Affective computing brings together three technologies: emotion detection, analysing emotional states through dialogue strategies, and emotional generation or synthesis.

Affective computing is based on the theory of mind, and especially on the notion that individuals possess the cognitive ability to attribute unobservable mental states (intention, desire, belief, emotion) to themselves and to other individuals. When humans perceive emotions in other humans, their perception belongs to them. On the other hand, the machine's perception is just a probabilistic estimate obtained by a calculation in a stochastic network built from emotional data that was originally produced by numerous human subjects.

Artificial empathy aims at improving cooperation with conversational agents. It uses emotion detection to simulate some aspects of human empathy, i.e. to allow the conversational agent to behave as if it were putting itself in the user's shoes. This results in affective anthropomorphism, which serves its purpose for the conversational agent but can potentially harm a user. Likely negative consequences for humans range from reactions of attachment, to guilt, or unjustified trust towards the conversational agent or other humans. Defining the limits for affective simulation of chatbots should involve reflecting on applications and contexts as well as on human vulnerabilities. For instance, a "sad" chatbot could worsen depression but in other cases it may relieve the pain.

DESIGN PRINCIPLE 8

RESPECT PROPORTIONALITY WHEN IMPLEMENTING AFFECTIVE COMPUTING TECHNOLOGIES IN CHATBOTS

To limit the spontaneous projection of emotions on conversational agents and to reduce assigning them with an inner self, the developer should respect the proportionality and adequacy between the intended purposes and the necessity of affective computing to achieve them. In particular, the detection of human emotions and artificial empathy of the chatbot should be carefully considered. The developers should also inform the user of the potential biases of anthropomorphism.

RESEARCH QUESTION 9

INVESTIGATING THE EFFECTS OF CHATBOTS ON HUMAN EMOTIONAL BEHAVIOUR

In the emerging field of empathetic conversational agents, designers need to perform research and undertake risk analysis with regard to the impact of these systems on the emotional behaviour of human users, especially in the long term.

Some emotions are culturally and socially dependent. When this is the case, tensions may emerge regarding their representation in language. For example, emotional small talk seems necessary in some cultures to establish a friendly relationship, but in other societies the same type of small talk is considered a sign of insincerity or even hypocrisy. A chatbot will be judged differently depending on the cultural context. Whether one takes a universalist, cosmopolitan, or communitarian perspective, affective conversational agents must respect the values of the cultures to which their users belong.

DESIGN PRINCIPLE 9

ADAPT CONVERSATIONAL AGENTS TO CULTURAL CODES

Chatbot developers should adapt conversational agents to cultural codes, including codes of emotional conduct, in different parts of the world.

DESIGN PRINCIPLE 10

INFORM ABOUT THE FEATURES OF AFFECTIVE CONVERSATIONAL AGENTS

When informing about emotional conversational agents, developers should seek to explain the actual limitations and features of these systems, so that the users do not overestimate the simulation of emotions.

5. EVALUATION OF CONVERSATIONAL AGENTS

A conversational agent provides a response by applying dialogue strategies that emerge based on the automated learning models. The most advanced models use large data sets. The evaluation of systems is essentially dynamic and belongs to the "ethics by design" approach (section III.1). It is problematic in at least two respects: a) the difficulty to predict user's language patterns; b) and the difficulty to predict dialogue strategies, which contributes to the difficulty of reproducing the system's behaviour. This theoretical and practical uncertainty is inseparable from the learning techniques that provide these systems their high efficiency. This makes the evaluation of how these machine learning systems behave a key issue.

The issues of evaluation are particularly pressing for adaptive machine learning systems, although they are still quite rare among commercialized chatbots. In most conversational agents, the success of a dialogue is often measured by the users' engagement, i.e. their willingness to continue the dialogue with the chatbot. Reinforcement learning techniques apply metrics that optimize user engagement, like the duration of exchanges as well as external metrics of satisfaction or interest (laughter, smiling, hesitation, nodding, etc.), to influence the algorithms in choosing responses.

REINFORCEMENT LEARNING IN CHATBOTS

It is usual to say in the field of reinforcement learning that the conversational agent interacts with the "environment" to find the optimal solution. Reinforcement learning is special because of its interactive and iterative aspects. During the same interaction the chatbot tries several solutions and depending on the reactions of the environment in which it operates, it adapts in order to arrive at the best strategy. The goal is to learn how to respond in different situations from experience and, more technically, to optimize a quantitative global reward over time. Some researchers have tried to prove that reinforcement learning techniques would be sufficient to account for all signs of human intelligence. This important debate deserves to be deepened in the context of conversational agents.

[59] D. Silver, S. Singh, D. Precup, and R. S. Sutton, "Reward is enough". Artificial Intelligence 299 (2021) 103535.

These dialogue strategies may be ethically unacceptable. For example, if a conversational agent observes that, statistically, users tend to respond to insults addressed to them, it could insult its user in order to maximize their engagement in the dialogue.

AN ADAPTIVE CHATBOT: TAY

In April 2016, Microsoft's chatbot Tay was equipped with the ability to learn adaptively from its interactions with users on the Internet. During its operation, it quickly learned to make racist comments. DeepCom, another chatbot developed by Microsoft China in 2019 was developed to comment on news on social networks. Researchers themselves recognized the chatbot "to be likely to generate biased content, even propaganda, following strong reactions in the research community."[60]

The development of "deepfakes" using reinforcement learning systems can pose a security problem for entities that rely on verbal or informal procedures. Chatbots using "deepfakes" can be used to make fraudulent financial operations.

BOX: LEGAL MEASURES RELATED TO DEEPFAKES

Article 52 of the proposed EU Regulation on Artificial Intelligence claims that the manufacturer of an artificial intelligence system "that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated." Failure to do so would be punishable by fine (Article 71).

Errors inevitably arise when a machine learning system classifies data that does not resemble data in the training corpus. In the case of conversational agents, this includes homophones, homographs, homonyms, and other examples of linguistic ambiguity. Hackers may abuse this inherent instability of machine learning system. They may manipulate chatbots to engage in inappropriate speech choices, harmful advice, or to malfunction.

Chatbots will make use of ever larger data sets. Since the use of transformer neural networks does not have a long history, there is no experimental data to evaluate their effects. For instance, GPT-3 is currently not able to systematically filter out racist, sexist, and hateful outputs. This is a complex technical problem. It is therefore necessary to develop evaluation methods adapted to the very large neural networks.

RESEARCH QUESTION 10

DEVELOPING SPECIFIC EVALUATION METHODS FOR CONVERSATIONAL AGENTS

Public authorities and private enterprises should support research on the evaluation of conversational agents during their use and propose new tests fitting various use contexts.

RESEARCH QUESTION 11

INVESTIGATING THE POTENTIAL OF TRANSFORMERS FOR SIMULATING DIALOGUE

In view of the potential to process and generate language using transformers, research should be supported on conversational agents that use these neural networks. Special attention should be given to evaluating their conformity with ethical values.

[60] arXiv:1909.11974.

IV. LIST OF RECOMMENDATIONS, DESIGN PRINCIPLES, AND RESEARCH QUESTIONS

RECOMMENDATIONS

RECOMMENDATION 1: REDUCE THE PROJECTION OF MORAL TRAITS ON A CONVERSATIONAL AGENT

To reduce the spontaneous projection of moral traits on the conversational agent and to limit the attribution of responsibility to such systems, the manufacturer must limit its personification and inform the user about biases that may result from the anthropomorphization of the conversational agent.

RECOMMENDATION 2: AFFIRM THE STATUS OF A CONVERSATIONAL AGENT

Any person communicating with a conversational agent must be informed in an appropriate, clear and intelligible way that they are conversing with a machine. The format and timing of this communication must be adapted on a case-by-case basis.

RECOMMENDATION 3: CONFIGURE THE IDENTITY OF CONVERSATIONAL AGENTS

To avoid bias, especially gender bias, the settings by default of a conversational agent for public use (name, personal pronouns, voice) should be made in an equitable way whenever possible. In the case of personalized conversational agents for private or domestic use, the user must be able to modify the default settings.

RECOMMENDATION 4: ADDRESS THE INSULTS

If situations in which the user engages in insulting a conversational agent cannot be avoided, the manufacturer should anticipate them and define specific response strategies. In particular, the conversational agent should not respond to insults with insults and should not report them to an authority. Manufacturers of chatbots that use machine learning techniques should exclude such phrases from the training data.

RECOMMENDATION 5: INFORM ABOUT DELIBERATE MANIPULATION

If the design of a conversational agent includes the capacity to influence user behaviour as part of its intended use, the manufacturer must inform the user about the existence of this functionality and obtain consent. The user must be able to withdraw this consent at any time. The manufacturer of a conversational agent that may influence user behaviour must inform the users about the nature and the origin of messages formulated by the chatbot as well as its communication methods. The manufacturer must ask users to exercise vigilance before sharing such messages.

RECOMMENDATION 6: AVOID MALICIOUS MANIPULATION

The manufacturer must seek to avoid the technical possibility of malicious manipulation or threats issued by the conversational agent. The user must have the ability to flag unwanted expressions, leading to a modification of the conversational agent by the developer.

RECOMMENDATION 7 : SET UP A FRAMEWORK FOR THE USE OF CHATBOTS IN TOYS

In the toy industry, particularly with regard to toys for young children, public authorities must assess the effects of user interactions with chatbots having a potential to influence children's behaviour. Public authorities must regulate the use of such conversational agents with regard to the impact on children's linguistic, emotional and cultural development.

RECOMMENDATION 8 : RESPECT VULNERABLE INDIVIDUALS

In the case of a dialogue between a conversational agent and a vulnerable individual, the manufacturer of the conversational agent must seek to respect the dignity and autonomy of this person. In particular, medical chatbots must be designed to avoid excessive trust in these systems by the patient and to ensure that any possible ambiguity between the conversational agent and a qualified physician is eliminated.

RECOMMENDATION 9 : ANALYSE THE EFFECTS OF CONVERSATIONAL AGENTS USING PHYSIOLOGICAL DATA

In the case of conversational agents with access to physiological data ("Quantified Self"), designers must study the risk of creating dependency. Public authorities must supervise the use of these systems with regard to their impact on personal autonomy.

RECOMMENDATION 10 : DEFINE RESPONSIBILITIES FOR THE USE OF CONVERSATIONAL AGENTS IN THE PROFESSIONAL ENVIRONMENT

The manufacturer should envisage control and audit mechanisms to facilitate the attribution of responsibilities for the functioning or malfunctioning of a conversational agent in the professional environment. In particular, the manufacture must study the chatbot's secondary or unintended effects.

RECOMMENDATION 11 : CONDUCT A REFLECTIVE PUBLIC DEBATE BEFORE REGULATING "DEADBOTS"

The legislator should adopt specific regulation concerning conversational agents that imitate the speech of deceased persons after an extensive ethical reflection at the societal level.

RECOMMENDATION 12 : SET UP A TECHNICAL FRAMEWORK FOR "DEADBOTS"

The developers of "deadbots" must respect the dignity of the human person, which does not end with death, while seeking to protect mental health of the users of "deadbots". Rules must be defined concerning the consent of the deceased person, the collection and reuse of their data, the operating time of a "deadbot," the vocabulary used, the name given to the chatbot, and the specific conditions of its use.

RECOMMENDATION 13 : SET UP A FRAMEWORK FOR THE USE OF "GUARDIAN ANGEL" CHATBOTS

To limit paternalism and to respect human autonomy, public authorities must set up a framework for the use of "guardian angel" conversational agents that are designed to protect personal data.

DESIGN PRINCIPLES

DESIGN PRINCIPLE 1: "ETHICS BY DESIGN" OF CONVERSATIONAL AGENTS

The developers of a conversational agent must analyse during the design phase every technological choice that may cause ethical tension. If a potential ethical issue is identified, the developers must envisage a technical solution seeking to reduce or eliminate it. They should subsequently evaluate this solution in realistic usage contexts.

DESIGN PRINCIPLE 2: REDUCE LANGUAGE BIAS

To reduce language bias and seek to avoid discrimination, especially cultural discrimination effects, developers must implement a technical solution at three levels: in the implementation of the algorithm, in the selection of optimization parameters, and in the choice of training and validation data for the different conversational agent modules.

DESIGN PRINCIPLE 3: DECLARE THE CHATBOT'S PURPOSE

The developer must ensure that a conversational agent clearly declares its purpose to the user in an easily understandable way at an appropriate moment, for example at the beginning or at the end of each conversation.

DESIGN PRINCIPLE 4: TRANSPARENCY AND TRACEABILITY OF THE CHATBOT

In compliance with the GDPR, a conversational agent should be able to save parts of the conversation (the extent of which needs to be defined) for evidential purposes or to satisfy security requirements. This need creates a tension with the protection of personal data. Chatbot architecture, used data, and dialogue strategies should be made available for audit and legal proceedings if needed. This recommendation may result in a regulatory measure to define the precise application terms.

DESIGN PRINCIPLE 5: PROCESSING DATA COLLECTED BY CONVERSATIONAL AGENTS

Following the existing example of health data, it is necessary to develop ethical and legal rules in compliance with the GDPR for the collection, storage, and use of linguistic data resulting from the interactions with conversational agents.

DESIGN PRINCIPLE 6: INFORM ABOUT THE FEATURES OF CONVERSATIONAL AGENTS

In the interest of transparency, the user should be informed in an appropriate, clear and intelligible manner, either orally or in writing, of the data collection, the adaptive features of the conversational agent, the data it collects during use, and profiling.

DESIGN PRINCIPLE 7: PROMOTE EXPLAINABLE CHATBOT BEHAVIOUR

Developers must devise solutions to facilitate understanding of the chatbot behaviour by the users.

DESIGN PRINCIPLE 8: RESPECT PROPORTIONALITY WHEN IMPLEMENTING AFFECTIVE COMPUTING TECHNOLOGIES IN CHATBOTS

To limit the spontaneous projection of emotions on conversational agents and to reduce assigning them with an inner self, the developer should respect the proportionality and adequacy between the intended purposes and the necessity of affective computing to achieve them. In particular, the detection of human emotions and artificial empathy of the chatbot should be carefully considered. The developers should also inform the user of the potential biases of anthropomorphism.

DESIGN PRINCIPLE 9: ADAPT CONVERSATIONAL AGENTS TO CULTURAL CODES

Chatbot developers should adapt conversational agents to cultural codes, including codes of emotional conduct, in different parts of the world.

DESIGN PRINCIPLE 10: INFORM ABOUT THE FEATURES OF AFFECTIVE CONVERSATIONAL AGENTS

When informing about emotional conversational agents, developers should seek to explain the actual limitations and features of these systems, so that the users do not overestimate the simulation of emotions.

RESEARCH QUESTIONS

RESEARCH QUESTION 1: AUTOMATICALLY RECOGNIZING INSULTS

It is necessary to develop methods for the chatbots to automatically detect inappropriate language, especially insults.

RESEARCH QUESTION 2: STUDYING LIES TOLD BY A CONVERSATIONAL AGENT

The empirical significance of lies told by a conversational agent requires further study. It is also necessary to avoid the projection of moral traits on a conversational agent via a narrative of its actions explicitly different from a narrative that characterizes lies told by humans.

RESEARCH QUESTION 3: ASSESSING THE UNFORESEEN EDUCATIONAL EFFECTS OF CHATBOTS

In education, public authorities need to evaluate the consequences of interactions between pupils and chatbots, especially when vulnerable or young children are involved.

RESEARCH QUESTION 4: STUDYING THE EFFECTS OF CONVERSATIONAL AGENTS ON THE ORGANIZATION OF LABOUR

Public authorities and private enterprises should support empirical research on the effects of conversational agents on the organization of labour across different industrial sectors.

RESEARCH QUESTION 5: STUDYING LONG-TERM EFFECTS OF USING CHATBOTS

Public authorities and private enterprises must invest in research on long-term effects on humans and society of the use of conversational agents. All societal stakeholders must remain aware of the potential future effects of conversational agents on users' beliefs, opinions and decisions, and avoid considering this technology as neutral or devoid of ethical and political significance.

RESEARCH QUESTION 6: STUDYING THE ENVIRONMENTAL IMPACT

Public authorities and private enterprises should conduct studies on energy consumption and environmental impact of the technology that enables conversational agents.

RESEARCH QUESTION 7: DEVELOPING THE "ETHICS BY DESIGN" METHODOLOGIES FOR CHATBOTS

Public authorities should support research to elaborate the "ethics by design" methodologies suitable for the development of conversational agents.

RESEARCH QUESTION 8: STUDYING REPRODUCIBILITY OF THE CHATBOT'S LANGUAGE

chatbot's language. These issues need to be studied.

RESEARCH QUESTION 9: INVESTIGATING THE EFFECTS OF CHATBOTS ON HUMAN EMOTIONAL BEHAVIOUR

In the emerging field of empathetic conversational agents, designers need to perform research and undertake risk analysis with regard to the impact of these systems on the emotional behaviour of human users, especially in the long term.

RESEARCH QUESTION 10: DEVELOPING SPECIFIC EVALUATION METHODS FOR CONVERSATIONAL AGENTS

Public authorities and private enterprises should support research on the evaluation of conversational agents during their use and propose new tests fitting various use contexts.

RESEARCH QUESTION 11: INVESTIGATING THE POTENTIAL OF TRANSFORMERS FOR SIMULATING DIALOGUE

In view of the potential to process and generate language using transformers, research should be supported on conversational agents that use these neural networks. Special attention should be given to evaluating their conformity with ethical values.

ANNEX 1: CONSENT

When the debate on conversational agents is evaluated from the perspective of data protection, the main issue is informing the user about the processing of their personal data to obtain their consent and, if necessary, to offer avenues of limiting the processing or withdrawing from it.

CONSENT IN THE CONTEXT OF GDPR AND THE FRENCH LAW OF JANUARY 6, 1978

The issue of personal data protection has become crucial with the development of digital technology, the explosion of data processing, and the availability of free services in return for the use of data. The protection from the collection of personal data was considered to be part of privacy as early as the law of January 6, 1978,[61] known as the Data Protection Act in France. The European Union has begun regulating it in 2002[62] with regard to communication technologies. It is now governed by the European regulation of April 27, 2016,[63] known as the GDPR, especially its chapters II and III,[64] that constitute a set of protective rights for the individual.

Consent is one of the six legal bases that permit personal data processing. The others consist in legal obligation, contract (contractual or pre-contractual relations), public task, vital interests, and legitimate interests (e.g. commercial prospecting operations with customers of a company without conclusion of a contract).

In accordance with Article 12 of the GDPR, the user must be informed about the privacy policy in a concise, transparent, intelligible, and easily accessible form, using clear and plain language. This information must include the purposes of the processing, whether the provision of personal data is a contractual requirement, and the possible consequences of failure to provide such data, as well as the categories of the data, the identity and the contact details of the controller, the contact details of the data protection officer, the recipients or categories of recipients of the personal data, the possible transfer of data outside the EU (to third countries, with appropriate safeguards are provided for the transfer), the period for which the personal data will be stored, the right to obtain from the controller confirmation as to whether or not personal data are being processed, and, where that is the case, access to the personal data and the following information, the right to rectification, or the right to erasure...

The user of a conversational agent, as a consumer, must give free, specific, informed, and unambiguous consent (unless otherwise provided by law). The user can withdraw their consent. Article 7 of the GDPR states that the controller must ensure that it is as easy for the data subject to withdraw consent as it is to provide it, and that he or she can withdraw it at any time. An easy withdrawal procedure known to the user is a guarantee for a valid consent. It should be noted that the GDPR prohibits the processing of biometric data (e.g. voice parameters or patterns), which are considered sensitive data, with certain limited exceptions.

The control of personal data protection falls under a national regulator, the CNIL in France, which monitors compliance with the GDPR and the French Data Protection Act,[65] mostly by issuing opinions and formal notices and by applying sanctions under the oversight of the Council of State.[66] Although the national judge and the Court of Justice of the European Union are progressively developing case law on data protection, there are questions on the quality of consent, its meaning, and the conditions under which it is collected (legibility, clarity, and precision of clauses). There are tensions between the law and the actual collection of data, which stimulates the current reflections in this area.

How can one consent to clauses that are not easily explained and may be numerous? How can parental consent be effectively obtained for minors?[67] Can the perpetual demands for consent lead to a fiction of law? Should we review the methods of informing and find new ways of presenting the committing clauses of the digital consent? Perhaps, by using games, graphic design, and visual aids? How can the effects of consent be presented at the time of the proposal that entices the user?

[61] Law no. 78-17 of January 6, 1978 relating to data processing, files and freedoms and various provisions concerning the protection of personal data.

[62] Directive 2002/58/EC of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (e-privacy Directive), which should be replaced by the e-privacy Regulation.

[63] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[64] Principles and rights of the data subject.

[65] The French Data Protection Act in its current version contains a number of provisions taken from the GDPR, includes clarifications in relation to it, incorporates articles from the Law of 7 October 2016 for a Digital Republic and transposes the e-privacy directive of 12 July 2002.

[66] See CNIL sanction of January 21, 2019 against Google. One of its arguments was that users are not always able to apprehend the effect of data processing on their privacy. The requirement of informed, specific, and unambiguous consent was not respected. The appeal against this decision was rejected by a decision of the Council of State of June 9, 2020 Google LLC n° 430810.

[67] Less than sixteen years old (article 8 of the GDPR) and less than fifteen years old (article 7-1 of the French Data Protection Act).

What is the meaning of consent for the use of a conversational agent? The notion of the transparency of information may become illusory – it depends on the collected data, the functions of the chatbot, and its mode of expression (written or oral). The consent information cannot be standardized and needs to be adapted. Providing adequate information for the consent about conversational agents that replace physical persons in customer service or in advertising is particularly complicated because of the tension between the requirements of transparency and efficiency. In point of fact, the user's time is limited and when they are in a hurry they are not particularly interested in the protection of their data as long as their primary intention is to solve a technical problem or find a commercial deal. The companies tend to aim for promptness rather than thoroughness in order to be attractive and efficient and they may risk their trustworthiness by not explaining the purpose of the conversational agent.

The issue of consent can be particularly sensitive when a conversational agent is installed in-home or embedded on a mobile device and intended as a confidant or an assistant. Users may engage in very personal conversations with them, thus risking a significant invasion of privacy. Although the consent information must be adequate, it must also be delivered in such a way that the consent is informed. However, it is not easy to assess whether the user has understood the issues in the information that was provided.

ANNEX 2: CALL FOR CONTRIBUTIONS

The CNPEN opened a call for contributions on the ethical issues of conversational agents in July 2020 to meet the Prime Minister's referral. The call had three objectives: to enrich the reflection thanks to the responses, to raise awareness among the various digital stakeholders, and among all citizens about the ethical issues raised by conversational agents. The call for contributions was formulated in French and English and distributed to academic and industry audience both in France, and abroad.

Ninety-six individual and collective contributions representing personal and professional opinions were collected between July 2020 and January 2021. Some contributions were delivered by academic students and staff, e.g. educational managers, teachers, and researchers, from France and abroad. A significant number of these respondents came from the background in law or computer science. Several contributions also came from professionals in the industry (telecommunication specialists, chatbot developers, etc.) or service sectors (bankers, data analysts, consultants). The remainder of the contributions mostly came from the health, justice, or digital regulation areas.

The questionnaire has also been used as a heuristic tool by teachers. It also served as a basis for discussion on the ethics of conversational agents within companies. These applications were not necessarily anticipated by the committee, but they helped to solidify the relevance and usefulness of the consultation process. In fact, these heuristic applications show the lack of concrete tools for training, cultural integration, and public awareness.

The questions were chosen to highlight ethical tensions arising from the uses of chatbots in concrete examples. The contributions reflect the difference in the level of the knowledge about these technologies as well as the familiarity with the ethical issues.

For example, some contributors thought that the anthropomorphization of conversational agents is necessary for their interactions. Others held that the anthropomorphization must be avoided as much as possible, even banning the use of the first person by the chatbot. In addition, questions about chatbot mistakes, lies, and emotions, have evoked polarized answers that reflect moral choices rooted in beliefs, images, and traditions of various origins.

Although the collected contributions cannot be considered representative, some answers shared a wide consensus. For example, the questions about free choice highlighted the acceptance of the principle of transparency and respect for the user's autonomy. In the eyes of the respondents, the developers must find the technical means and the design to allow the implementation of these principles.

It was interesting to observe the difference between the answers of chatbot developers and users. The trust in chatbots is particularly interesting in this regard. The developers thought that there are inevitable situations in which the chatbot cannot provide accurate information to the user, whereas users were not keen on accepting mistakes, which they saw as failures in the chatbot's design. It is difficult to grasp the causes of the chatbot's behaviour without the technical knowledge and prior experience.

NATIONAL PILOT COMMITTEE FOR DIGITAL ETHICS (CNPEN)

ETHICAL ISSUES OF CONVERSATIONAL AGENTS

ANSWERS SHOULD BE SENT TO CNPEN-CONSULTATION-CHATBOTS@CCNE.FR

The National Pilot Committee for Digital Ethics (CNPEN) was established in December 2019 at the request of the Prime Minister. Composed of 27 members, this committee brings together computer scientists, philosophers, doctors, lawyers, and members of civil society. One of the three referrals submitted by the Prime Minister to the CNPEN concerns the ethical issues of conversational agents, commonly known as chatbots, which communicate with the human user through spoken or written language. This work of the CNPEN is an extension of the work initiated by CERNA, the Allistene Alliance's Commission for Research Ethics in Digital Science and Technology.

This call is intended to allow stakeholders and the public to express their views on ethical issues related to chatbots. We ask readers to answer all twenty questions or any subset thereof. Contributors' names will not be attached to any answers quoted in the future opinion.

Under the conditions defined by the French Data Protection Act of 6 January 1978 and by the European Regulation on the Protection of Personal Data which came into force on 25 May 2018, each contributor has the right to access, rectify, query, limit, transfer, and delete data concerning him/her. Each contributor may also, on legitimate grounds, object to the processing of such data. The contributor may exercise all of the abovementioned rights by contacting the CNPEN at the following email address: cnpen-consultation-chatbots@ccne.fr. The following data will remain confidential and will be stored on the servers used by the CNPEN. They will be used exclusively by members of the CNPEN for the purpose of analyzing contributions to this call.

You are answering this questionnaire:

- In your personal capacity (specify first and last name if you wish)

- As part of your professional activity or on behalf of an organization:

- *Researcher or research institute (specify the name of your institution)*
- *Company or group of companies (specify which)*
- *Consumer association or similar (specify which)*
- *Public authority (specify which)*
- *Professional consultant*
- *Think tank (specify which)*
- *Other*

INTRODUCTION

WHAT IS A CONVERSATIONAL AGENT?

A conversational agent, commonly called a chatbot, is a computer program that interacts with its user in the user's natural language. This definition includes both voice agents and chatbots that communicate in writing.

The conversational agent is most often not an independent entity but is integrated in a system or digital platform, e.g. a smartphone or a voice speaker. In terms of visual appearance, chatbots can also be integrated into an animated conversational agent, represented in two or three dimensions on a screen, or even be part of a social, including humanoid, robot. In this case, the dialogue capacity is only one of the functions of the overall system.

The history of conversational agents has its origin in Alan Turing's imitation game. Turing's interest was in language comprehension to the extent to which it is manifest in answers that appear intelligible and sensible to a human examiner (the Turing Test). Since 1991, an annual competition has been held to support the development of chatbots capable of passing the Turing Test.

The first conversational agent in the history of computer science is Joseph Weizenbaum's ELIZA program, which is also one of the first conversational tricks. ELIZA simulates a written dialogue with a Rogerian psychotherapist in Rome by simply rephrasing most of the "patient's" responses in the form of questions. Today, the term "ELIZA effect" refers to the tendency to unconsciously equate dialogue with a computer with that with a human being.

FROM A TECHNICAL POINT OF VIEW, HOW DOES IT WORK?

The design and operation of a chatbot is divided into several modules for automatic natural language processing (NLP). Schematically, a chatbot can include modules for speech recognition (for voice conversational agents), semantic processing (out of and in context), dialogue history management, dialogue strategy management, access management ontology, management of access to external knowledge (database or internet), language generation, and speech synthesis (for voice conversational agents).

A conversational agent follows rules decided and transposed into code by human designers or obtained by learning. Learning chatbots, such as Microsoft China's XiaoIce, for example, are still quite rare among commercialized applications, but their proportion will continue to grow as mastery of this technology advances.

In recent years, developing a rudimentary or single-task chatbot yourself has become relatively easy thanks to the availability of many design tools, such as "LiveEngage", "Chatbot builder", "Passage.ai", "Plato Research Dialogue System", etc.

SOME RESEARCH CHALLENGES IN CONVERSATIONAL AGENT DESIGN

- Learn adaptively by evolving the knowledge base in use.
- Be able to converse freely on generic topics. commun », le caractère ironique ou le sens au « second degré » d'un énoncé.
- Grasp the common sense, ironic, or tongue-in-cheek meaning of a statement.
- Set up a dialogue strategy.
- Detect the user's emotions and intentions.

SOME RESEARCH CHALLENGES REGARDING USERS' UNDERSTANDING OF CONVERSATIONAL AGENT CAPABILITIES

- What data do chatbots record? Are they anonymized?
- How can chatbots' behavior be audited (automatic measurement and/or human evaluation)?
- Are the responses selected by the chatbots explicable? Can the chatbots make themselves more understandable?
- Which of the user's profile parameters do chatbots calculate? Are humans aware of this?
- Does the user's idea of the chatbot's strategy correspond to the actual strategy implemented in the chatbot?

[68] "Google Assistant", "Google Home", "Apple Siri", "Amazon Alexa" et "Amazon Echo", "Yandex Alisa", "Mail.ru Marusia", "Baidu DuerOS", "Xiaomi XiaoAI", "Tencent Xiaowei", "Samsung Bixbi", "Orange Djingo", etc.

[69] A. Turing, "Computing Machinery and Intelligence", Mind 59(236) 433–460, 1950.

[70] J. Weizenbaum, "ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine", Communications of the Association for Computing Machinery 9, 36-45, 1966.

[71] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum, "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot", Computational Linguistics 46(1), 53-93, 2020.

ETHICAL QUESTIONS

Language is a constituent element of human identity and the foundation of human life in society. Conversational agents are thus naturally compared to a human being, whether or not their user is aware of their artificial nature. This natural aspect of dialogue is likely to influence the human being: this is the fundamental problem of the ethics of chatbots. Since their deployment is a recent phenomenon, there is not enough experimental data to assess their long-term effects on human beings.

Recently, the performance of speech recognition has made it possible to use voice interfaces. In addition to language dialogue, the voice carries information of various kinds, such as the speaker's age, gender, body size, mother tongue, accent, living environment, sociocultural background, education, health status, understanding, and emotions. Many ethical issues are related to these aspects of human life.

Like technical systems in general and autonomous systems in particular (e.g. automatic image recognition or self-driving vehicles), conversational agents must meet a large number of requirements in terms of security, transparency, traceability, usefulness, privacy, etc. The systems of each type implement these properties according to the specific context of their use. In all cases, these are key constraints for both the designer and the user.

Some conversational agents create new ethical tensions, such as the impossibility of explaining in natural language the chain of decisions leading to a particular medical recommendation. Recommendations are made in this respect in the CERNA opinion on the ethical issues of research in machine learning.[72]

[72] http://cerna-ethics-allistene.org/Publications%2bCERNA/apprentissage/index.html

CONSULTATION

I. ETHICAL FACTORS IN THE USE OF CHATBOTS

1) STATUS CONFUSION

Several factors help to confuse a conversational agent with a human being. A blurring of status distinctions may occur as a brief illusion or, on the contrary, it may persist throughout a dialogue. It may also be voluntary or spontaneous, have psychological or legal consequences, or give rise to varying degrees of manipulation. This confusion of status is caused by a more general phenomenon.

A human being spontaneously projects human traits onto an interlocutor, of whatever nature: thought, will, desire, conscience, internal representation of the world. This behavior is called "anthropomorphism". The interlocutor then appears as an autonomous individual endowed with thought, expressed through words.

To date, only a law in the State of California explicitly requires an interaction with a chatbot to be mentioned when this interaction is intended to encourage the purchase or sale of products or services in the context of a commercial transaction or to influence voting in an electoral context. There is no equivalent to this provision in French or European law, even though this point is now being considered.[74]

1.1 Should the user be informed of the nature of the interlocutor (human being or machine)? And, if so, what information about the chatbot should be communicated to the user (purpose, training corpus, name of the designer, etc.)?

1.2 Do you think that in Europe we should adopt a legislative framework comparable to that of the State of California?

1.3 Free comments:

2) NAMING

People often give a conversational agent a name, as children do with their dolls.

Sometimes, the naming is intended by the designer: addressing the machine by a name can help it to function better, in the personal assistance or entertainment sectors, for example. In these cases, the use of the name heightens the user's emotional response.

Currently, this use of a name and of emotional response is still often used to mask the lack of semantic and contextual performance of conversational agents. Assigning a name to the machine is part of the dynamics of projection, i.e. the anthropomorphization of this machine. However, when the conversational agent itself uses its "name" in a dialogue, the question of self-reference arises: to whom or what does this name refer?

2.1 Should the user be able to choose the name and the gender of the name (masculine, feminine, neutral) assigned to a chatbot, or is this choice up to the designer?

2.2 Could or should a chatbot be given a human name (e.g. "Sophia"), a non-human name (e.g. "R2D2"), or no name at all?

2.3 Free comments:

3) BULLYING OF CHATBOTS.

The projection of human qualities onto chatbots is a common and important phenomenon. In particular, users may mistreat a conversational agent.

While your chatbot reminds you of protective measures during an epidemic, you might respond by insulting it or ordering it to be quiet. This could affect children who hear the exchange.

Voice assistants (Siri, etc.) are sometimes insulted by users. In this case, they respond according to strategies predetermined by their designers.

3.1 Is insulting a chatbot in a conversation a morally reprehensible act? Do you think it is permissible to use the chatbot as a punching bag?

3.2 Should a chatbot who is insulted be able to respond by insulting the user in turn?

3.3 If a chatbot with a feminine name or even a feminine voice is abused, do you see this as abuse towards women? The same question applies to male names.

3.4 Free comments:

4) TRUST IN CHATBOTS

A certain amount of user confidence in the chatbot's purpose is necessary for the chatbot to perform its functional tasks.

Trust is not only an emergent psychological phenomenon, but also the result of a technical effort: conversational agent designers seek to establish and maintain trust, but may also consider avoiding giving it unthinkingly to the chatbot.

Assessing the level of user trust in chatbot behavior and performance is an important research topic.

4.1 If a chatbot's "I don't know" weakens the user's trust, for example in the case of an after-sales service, should trust be promoted by modifying the answer?

4.2 In order to gain trust, can the chatbot introduce itself as the user's "assistant / advisor / friend"?

4.3 Free comments:

[73] https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=BPC&division=7.&title=&part=3.&chapter=6.&article

[74] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

5) CHATBOT CONFLICTS.

While most chat systems are designed for a specific task, many others are general-purpose conversational agents. Their interaction with humans can be part of a conflict. The question then arises as to the conversational agent's role in this conflict and its judgment. *For example, a chatbot could give the user unfortunate advice, lie, or behave like an informer by calling the police if it rightly or wrongly detects a threat.*

Current research focuses on the development and use of systems that can adapt to users, their wishes, intentions, and beliefs, by responding as would a relative. These adapted or even "intelligent" responses to human questions or behaviors can only lead users to believe in the "skills" or supposed "mindset" of the machine. Humans therefore adapt to the conversational agents with which they chat, either by distrusting them or, on the contrary, by giving them a certain credence. By relying on this credence, a chatbot could lie.

Tension arises when the chatbot, for example, answers a question about the user's health. A doctor may conceal the whole truth from the patient in the interest of the patient's well-being.

5.1 Is a lie told by a chatbot more or less acceptable than a human lie? Does the answer depend on the context (voice assistant, education, psychotherapy, recruitment, etc.)?

5.2 If chatbots can lie to users, who should decide on the permissible purposes and the limits of such behavior, and how?

5.3 Free comments:

6) MANIPULATION OF CHATBOTS (NUDGE THEORY).

The American Richard Thaler, winner of the Nobel Prize in Economics, has highlighted the concept of nudge, which consists in encouraging individuals to change their behavior without coercing them, simply by using their cognitive biases. In the case of chatbots, nudges are defined as suggestions or manipulations, overt or covert, designed to influence a user's behavior or emotions.

Conversational agents could thus become a means of influencing individuals for commercial or political purposes. But nudges are also often used to monitor our health or to improve our well-being (getting more exercise, drinking less alcohol, quitting smoking, etc.).

6.1 Are all nudges allowed? How can we distinguish between good and bad nudges?

6.2 Does the concept of free and informed consent still make sense when conversational agents nudge?

6.3 Free comments:

7) CHATBOTS AND FREE CHOICE.

During a dialogue, chatbots evaluate several possible answers and give one. In the case of recommendation systems, this single choice could limit the users' capacity to choose freely, by obscuring their view of the full range of available options. It also generates the risk of a filter bubble, a problem reinforced by the low level of configuration offered by the systems currently on the market.

For example, when asked to order a pizza, the chatbot suggests ordering from a particular pizzeria, which may be geographically closer, the top-rated on a given website, or one that has a commercial agreement with the chatbot's designer. However, the chatbot offers a single choice, while there are fifteen pizzerias in the neighborhood that offer the service requested. This single choice can pose an ethical problem related to freedom and discrimination.

7.1 In the example given, would you like the chatbot to explain all or several choices?

7.2 Do you think that transparent user information on the chatbot's criteria for recommendations is a satisfactory solution to the ethical problems of free choice and discrimination?

7.3 Free comments:

8) EMOTIONS OF CHATBOTS.

Mixed emotions are frequent in everyday life. The detection and identification of users' emotions therefore depend on a large number of contextual, cultural, and idiosyncratic factors. Affective computing has three main areas: recognition of human emotions, use of this information to modify the dialogue strategy, and generation of emotional expressiveness through language or nonverbal chat behavior.

For example, having recognized that the user is stressed, a conversational agent can simulate empathy and express understanding of the user's state.

8.1 Is it desirable to build chatbots that detect human emotions? Answer according to the context of use.

8.2 Is it desirable to build chatbots that simulate human emotions? Answer according to the context of use.

8.3 Free comments:

9) CHATBOTS AND VULNERABLE PEOPLE.

A chatbot can occupy a vulnerable person's full attention by replacing, as in autistic children, the difficult contact with other people. This often leads to polarized judgements: on the one hand, the person's well-being can be improved; on the other hand, this is at the expense of that person's "standard" human socialization.

For example, a child with autism may prefer the highly enriching and prolonged interaction with a chatbot to that with a parent or teacher. A young child might learn and imitate the emotional behaviors of the machine instead of those of humans. An older person may want to mourn or bury their chatbot if they are very attached to it and it is no longer functioning.

9.1 What purposes of interaction between a chatbot and a vulnerable person (monitoring, education, support, entertainment) are acceptable? Does the answer depend on the person's age (child, elderly) or status (patient, convalescent)?

9.2 Users, especially vulnerable people, are likely to become deeply attached to chatbots, which can lead to a lasting change in their lifestyle or social interactions. Is this a cause for concern? Why?

9.3 Free comments:

10) CHATBOTS AND MEMORY OF THE DEAD.

While the right to privacy ends when a person dies, post-mortem use of a chatbot's data, e.g. the person's voice, by a chatbot to "revive" that person may nevertheless infringe the principle of respect for the dignity of the human person.

An American journalist managed to create a chatbot, the "dadbot", from his memories of his father. He talks to this chatbot "as if" to his father.

10.1 Do you think chatbots can "give life" to a deceased person's memory or way of expression? Would such uses violate the principle of respect for the dignity of the human person?

10.2 How do you see the concept of death evolving with the possibilities offered by chatbots?

10.3 Free comments:

11) SURVEILLANCE BY CHATBOTS.

While some chatbots are parts of systems dedicated exclusively to human-machine interaction, others operate in shared environments. Chatbots capable of recording voices could monitor interactions around them, whether human or with other chatbots. This capability involves ethical and legal issues related to the protection of privacy, the use of personal data without consent, the risk of violation of personal or professional secrecy, and the introduction of security breaches. The disclosure by chatbots of content recorded without the knowledge of individuals may amount to denunciation.

For example, in the event of a deviation from the diet that a doctor has prescribed for a patient, the chatbot informs the doctor or even contacts the health care organization.

Another example is a chatbot that can monitor the behavior of vulnerable or elderly people and so "keep them company".

11.1 In the examples given, do you think the chatbot's behavior is justified? If so, how can users express their consent? What if chatbots are deployed in shared spaces?

11.2 Give other examples of situations in which chatbot monitoring seems justified.

11.3 If it is insulted by its user, should a chatbot inform a third party, its designer, for example?

11.4 Free comments:

12) CHATBOTS AND WORK.

Chatbots present opportunities and risks for companies, depending on the context in which they are used (evaluation, recruitment, entertainment, etc.). The introduction of chatbots in teams can induce organizational effects depending on the industrial sector, particularly in terms of information and emotional load, the temporality of work, the feeling of cohesion or isolation of workers, the effects of chatbots on employee morale, as well as the problems of equality and recognition of merit within companies.

For example, in the medical sector, assistance to human action (psychiatrists, general practitioners, nurses, emergency call center agents, etc.) provided by chatbots could have effects on the profession as a whole as well as on the well-being of patients and carers and on the relationship between them.

12.1 Are there professions or human practices in which the use of chatbots should be encouraged or prohibited?

12.2 How and on what time scale do you envisage the evolution of professions following the introduction of chatbots? Answer using one or more examples of usage.

12.3 By what means (legislative, code of conduct, etc.) should the use of chatbots be regulated?

12.4 Free comments:

13) LONG-TERM EFFECTS ON LANGUAGE.

In the medium to long term, the use of chatbots may have a lasting impact on human language and perhaps also on lifestyle habits.

For example, if chatbots respond with short, linguistically poor, impolite sentences, humans may imitate these language tics when speaking to other humans.

13.1 How do you envisage chatbots influencing the evolution of language? Can this influence be judged as good or bad?

13.2 What time scale can be envisaged for this evolution?

13.3 Free comments:

[75] James Vlahos. Talk to me, Amazon, Google, Apple, and the Race for Voice-Controlled AI. Random House, 2019.

II. ETHICAL FACTORS IN THE DESIGN OF CHATBOTS

14) SPECIFICATION PROBLEM.

Laws and rules of conduct in society are formulated in natural language. Their translation into a computer language requires a "specification ": definition of all terms in a formal framework. Often, complete specification is impossible: for example, the term "human" may include humans that would be easily identifiable by a learning computer system, but also humans that the system will not be able to identify as such because they are absent from the training data. Regardless of the learning base and the algorithm deployed, identification errors are inevitable: by nature, human language has multiple meanings.

For chatbots, the problem of specification translates, for example, into the difficulty of distinguishing, systematically and without error, the ironic or satirical use of a concept or expression from its standard indicative use.

14.1 Which mistakes made by chatbots would be acceptable and which would not? Answer according to the context (health, education, entertainment, after-sales service, etc.).

14.2 If a chatbot is not able to find an answer, must it say so explicitly?

14.3 What are the consequences for user behavior of the "I don't know" answer frequently given by current voice assistants? If you have had this experience, describe it.

14.4 Free comments:

15) METRICS AND EVALUATION FUNCTIONS.

In a conversational agent, the purposes intended by the designer result in the definition of a metric or evaluation function, which quantifies the measure of "correct response" or "adequate response" for the system. This metric is pre-encoded. A chatbot metric can also take into account factors that emerge during the conversation and which may otherwise cause disruptions in human understanding of system behavior. Often, the quality of the dialogue is measured by the user's level of engagement, i.e. willingness to continue the dialogue with the chatbot. The engagement metric uses the length of the exchanges as paralinguistic markers (laughing, smiling, hesitation, nodding, etc.) of satisfaction or interest. However, in the current state of research, it rarely takes into account the semantic content of the exchanges. This can disadvantage those who do not understand the conversational agent's evaluation process and, moreover, lead to manipulative behavior on the part of users.

By April 2016, Microsoft's Tay chatbot, which had the ability to continuously learn from its interactions with internet users, had learned how to make racist comments. Tay was quickly withdrawn by Microsoft.

Despite this experience, DeepCom, *another chatbot developed by Microsoft China in 2019 to comment on news on social media, was recognized by its designers themselves as likely to generate biased (e.g. discriminatory) content or even propaganda, following strong reactions in the research community. The first version of the publication postulated: "Given the prevalence of online news articles with comments, it is very interesting to set up a system of automatic news commentary with approaches built from data". In the revised version, the authors state: "There is a risk that individuals and organizations may use these techniques on a large scale to simulate comments from individuals for purposes of manipulation or political persuasion. "*

15.1 Should the user be informed that a chatbot's dialogue strategy can be adapted during a conversation?

15.2 As explained above, users can manipulate chatbot metrics for their own purposes. If they do so, should the designer share the possible responsibility for the results of this manipulation or be released from it?

15.3 Have you had personal experiences that you interpret as being related to particular chatbot metrics?

15.4 Free comments:

16) GOALS OF THE CONVERSATIONAL AGENT:

The chatbot's goals, i.e. the goals assigned to it, are defined by its designers, and the chatbot seeks to satisfy them from the outset. While this does not pose excessive problems for chatbots dedicated to one or more previously known tasks, the specification of goals can be complex for a general-purpose chatbot because they cannot all be enumerated at the time of design.

These goals can be very diverse: after-sales systems help to repair defective products, medical advisors seek to improve the patient's health, recruitment assistance services, etc.

Other systems have vaguer goals: some chatbots are designed to converse freely with the user on any topic. The fact that the perception of these goals or the judgment thereof may evolve does not remove this fundamental distinction between a conversational agent and a human, whose goal may be neither predetermined nor made explicit to others.

16.1 Should the purpose of a chatbot be revealed to the user? If so, when and in what form? If not, why not?

16.2 Should it be accepted that a chatbot capable of interactive learning (e.g. a general-purpose conversational agent) can be directed to a particular goal through intentional or unintentional user influence (e.g. encouraging the person to make a donation or purchase a particular product)? Answer according to the context (health, education, entertainment).

16.3 Free comments:

[76] https://www.vice.com/en/article/d3a4mk/microsoft-used-machine-learning-to-make-a-bot-that-comments-on-news-articles-for-some-reason

17) TRAINING BIAS.

A system learns from data selected by a "coach" (human agent responsible for their selection). Bias in training data is a major source of ethical conflicts, particularly through ethnic, cultural, or gender discrimination.

For example, recorded speech data may contain only adult voices, whereas the system is supposed to interact with children as well, or a body of text may use female pronouns statistically more frequently than male pronouns.

The system will then reproduce these biases from a training corpus, unless it is equipped with specially designed tools to correct them, which already presupposes knowledge of possible biases. However, some biases may not be known in advance.

17.1 Do you consider that a conversational agent should be unbiased? Is this possible? Answer according to the context (health, recruitment, after-sales service, education, security, domestic voice assistant).

17.2 Do you think chatbots should mimic human biases or correct them?

17.3 Free comments:

18) TRAINING INSTABILITY.

Errors are inevitable when a learning system classifies data that do not resemble, or falsely resemble, those contained in the corpus used during its training. In the case of conversational agents, this includes homophones, homographs, homonyms, or other examples of linguistic ambiguity.

A simple case is that of spelling mistakes: the chatbot's behavior in this case differs completely from that of a human being. For example, the human user recognizes a word even if it contains several errors, whereas, because of instability, an algorithm stops correctly recognizing a word containing one or two spelling mistakes.

18.1 Since chatbot learning is unstable, it sometimes induces obvious mistakes. Are you willing to tolerate these errors more than human errors? Answer according to the context.

18.2 Do chatbots' mistakes elicit different feelings or reactions than human mistakes? Which ones?

18.3 Free comments:

19) EXPLAINABILITY AND TRANSPARENCY.

The transparency of a system means that its operation is not opaque or incomprehensible to humans. It relies in particular on the traceability of the responses selected by a conversational agent. Explainability means that a user can understand the chatbot's behavior. Problems of transparency and explainability are caused by various factors, notably that, unlike a human being, a computer system does not understand the meaning of the sentences it generates or perceives.

For example, a chatbot, which has no representation of the world, is likely to formulate phrases that do not correspond to any reality ("black milk"), to answer without taking into account the context ("How are you?" - "It's sunny"), or to use an unpleasant or prohibited lexicon.

The immediate effects of such a dialogue on the user can be significant (strong emotional reaction, break in understanding, abandonment of the dialogue, or disconnection from the system). The question of responsibility then arises with regard to the designers and trainers of conversational agents. Is the aesthetic dimension (some words may be strange but beautiful) enough to free the chatbot from the need to always imitate human speech?

19.1 What reaction can be expected from a user in a situation where there is a lack of understanding in a dialogue with the chatbot? Answer according to the chatbot's purpose and the context (e.g. health, general-purpose voice assistant, entertainment, recruitment).

19.2 When the user spontaneously gives meaning to unclear responses by the chatbot, is this a playful attitude or does it pose an ethical problem?

19.3 Free comments:

20) IMPOSSIBILITY OF RIGOROUS EVALUATION.

A conversational agent provides an answer by applying dialogue strategies that depend on interpretation. The most advanced models use large bodies of data to learn.

The evaluation of this inherently dynamic dialogue system is difficult in at least two ways: a) predicting user-generated input is often not possible; and b) the vagaries of learning contribute to the difficulty of replicating the system's behavior.

Uncertainty in theory and practice goes hand in hand with the learning techniques that give systems their high efficiency.

20.1 Is it acceptable for a chatbot to utter "incongruous" phrases, which no human being has ever used and which might influence the user?

20.2 Should a chatbot be limited to a predetermined set of phrases or, conversely, should it generate them freely? Answer according to the context (entertainment, after-sales service, education, general-purpose voice assistant).

20.3 Free comments:

THANK YOU FOR YOUR CONTRIBUTION!

YOU CAN SEND IT TO THE ADDRESS CNPEN-CONSULTATION-CHATBOTS@CCNE.FR

ANNEX 3 : MEMBERS OF THE WORKING GROUP

Laurence Devillers and
Alexei Grinbaum, co-rapporteurs.
Gilles Adda
Raja Chatila
Caroline Martin
Serena Villata
Célia Zolynski

Also contributed:
Eric Germain
Christophe Lazaro
Félicien Vallet (CNIL)
Camille Darche (secretary)

ANNEX 4 : METHODOLOGY OF THIS WORK

Interviews :

- Pr Pierre Philip (PU PH), University of Bordeaux. USR SANPSY (sleep, addiction, neuropsychiatry) USR3413, CNRS - University Bordeaux 2 - 24/02/20
- Julia Velkovska, sociologist at EHESS and at Orange - 04/20/20
- Mickaël Cabrol, (www.easyrecrue.com) - Easyrecrue CEO - Arthur Guillon, Senior Machine Learning Engineer - Léo Hemamou, PhD student, thesis in automatic detection of social signals - 15/05/20

Consultations : https://www.ccne-ethique.fr/en/actualites/cnpen-ethical-issues-conversational-agents
https://www.ccne-ethique.fr/en/actualites/cnpen-ethical-issues-conversational-agents

- The committee launched a consultation on ethical issues specific to chatbots from June to October 2020. It identified ethical issues and proposed scenarios for different chatbot applications.
- The call for contributions was intended to allow the stakeholders and the public to express their views on ethical issues related to chatbots. Each contributor was invited to answer all of the questions. It was indicated that contributors' comments would be anonymized and they would not be cited by name.
- The committee received the opinions of around 100 respondents (individuals, public and private institutions).
- The committee did not wish to analyze the results quantitatively; three working sessions were devoted to studying all of the responses and stimulating the collective reflection.

The National Pilot Committee for Digital Ethics (CNPEN) was established in December 2019 at the request of the Prime Minister under the auspices of the National Consultative Ethics Committee for health and life sciences. It brings together figures from the academic, industry, and policy areas. Experts of the digital and other technologies, law, economics, philosophy, linguistics, logic, and medicine join forces for an ethical reflection that has become inevitable in the face of digital innovation. It also serves to inform public debates. Previous opinions of CNPEN were focused on Ethical issues regarding digital tools at the lifting of the lockdown (May 2020) and Ethical issues regarding "autonomous vehicles" (May 2021).

LES MEMBRES DU COMITÉ NATIONAL PILOTE D'ÉTHIQUE DU NUMÉRIQUE

Gilles Adda
Raja Chatila
Theodore Christakis
Laure Coulombel
Jean-François Delfraissy
Laurence Devillers
Karine Dognin-Sauze
Gilles Dowek
Valeria Faure-Muntian
Christine Froidevaux
Jean-Gabriel Ganascia
Eric Germain
Alexei Grinbaum
David Gruson
Emmanuel Hirsch
Jeany Jean-Baptiste
Claude Kirchner - directeur
Augustin Landier
Gwendal Le Grand
Claire Levallois-Barth
Caroline Martin
Tristan Nitot
Jérôme Perrin
Catherine Tessier
Serena Villata
Célia Zolynski