National Pilot Committee for Digital Ethics (CNPEN)

Ethical issues of conversational agents

Call for participation

~~open till September 30th, 2020, at midnight~~

New date:

October 31th, 2020, at midnight.

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.

This response was coordinated by **Trilateral Research** (Nicole Santiago, Anaïs Resseguier, Rowena Rodrigues, and Mistale Taylor) **within the EU-funded SIENNA Project**. SIENNA (Stakeholder-Informed Ethics for New technologies with high socio-ecoNomic and human rights impAct) is looking into ethical, legal and human rights issues and is developing ethical guidelines for human genomics, human enhancement and AI & robotics. It has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.

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[1]. 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[2]. 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[3], 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).

---

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

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

[3] 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.

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[4], 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.
- 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?

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

[4] 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.

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[5].

[5] 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[6] 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[7].

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.)?

We recommend reducing as much as possible the confusion of a chatbot with a human being. Therefore, the user should always be informed when the interlocutor is a machine. This is especially true in cases where the potential impact is significant and adverse. Information about the nature of the chatbot should be easily accessible to the user, and the user should be given the option to readily access more information (e.g., name of designer/developer). Chatbots that interact with vulnerable populations (e.g. children) must be specifically adapted and appropriate, both in the information given and way it is presented.

As a general note throughout this submission, we encourage more social sciences studies and impact assessments on human-chatbot interaction to better understand and address the potential and actual impacts more concretely.

1.2. Do you think that in Europe we should adopt a legislative framework comparable to that of the State of California?
Yes, it is essential that people are informed of when they interact with a machine. However, unlike the California legislative framework, the user should always be informed when interacting with a chatbot, not just in the context of a commercial transaction or elections. A regulatory gaps assessment would be necessary to determine consistency with existing requirements under GDPR

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

7 High-Level Expert Group on Artificial Intelligence | Shaping Europe's digital future

1.3 Free comments: n/a

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?
To reduce the potential for anthropomorphization and emotional attachment to a machine, we recommend assigning, as a default, non-human identifiers to chatbots. If a user wants to re-assign a human name to a machine as an identifier, the option could be available, but this should remain an option for the user, not a default setting. Setting non-human identifiers would also help avoid gender, cultural and racial stereotyping of chatbots. Chatbots with human names might lead users to be more trusting of what might turn out to be a dubious service or product on offer and might lead to serious risk of harm.

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?
See above response to question 2.1.

2.3 Free comments: n/a

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?
What is morally reprehensible or not is highly contested, debated, and culturally specific. We do not condone reprehensible behaviour of any sort, towards humans or chatbots, but in this particular case, there is no reason to take a stance. However, if the question is whether insulting a chatbot should be an illegal act, the answer is no. A chatbot is not a human being.

3.2 Should a chatbot who is insulted be able to respond by insulting the user in turn?
No. A chatbot insulting a human would add no value to an interaction. This type of response will also lead to chatbots being rejected and used less and potentially cause harm to the user.

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.
The question is not whether this type of interaction constitutes abuse towards women. Rather, the critical issue is whether the design of the chatbot and the responses generated perpetuate stereotypes towards women (such as subservience) and whether those contribute to minimising issues of sexual harassment. Users are more likely to harass a chatbot than another human, particularly if the chatbot presents as female. There are many studies examining this, such as :

- Fessler (2017) on chatbots shows that, too often, chatbots tend to be "promoting stereotypical passivity, dismissiveness, and even flirtation with abuse." ('We tested bots like Siri and Alexa to see who would stand up to sexual harassment', *Quartz*, https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/).

- **Brahnam and De Angeli** (2012) also raise the issue that "[p]eople attribute negative stereotypes to female-presenting chatterbots more often than they do to male-presenting chatterbots, and female-presenting chatterbots are more often the objects of implicit and explicit sexual attention and swear words." (Sheryl Brahnam, Antonella De Angeli, 'Gender affordances of conversational agents', *Interacting with Computers* 24(3):139–153).

-**Silvervarg et al**. (2012) confirmed that chatbots with female voices "are more prone to be verbally abused than male agents, but also show that the visually androgynous agent was less abused than the female although more than the male agent." ('The Effect of Visual Gender on Abuse in Conversation with ECAs', Lecture Notes in Computer Science book series (LNCS, volume 7502 https://www.researchgate.net/publication/262242099_The_Effect_of_Visual_Gender_on_Abuse_in_Conversation_with_ECAs).

Therefore, we recommend using androgenous 'voices' and non-human identifiers in the design of chatbots to minimise the perpetuation of stereotypes towards women. This will also help inclusivity.

3.4 Free comments: For more discussion, please see:

Aditya Singh, *"Siri, talk dirty to me"- The Ethics of Conversational AI*, Trilateral Research, July 2020, https://www.trilateralresearch.com/siri-talk-dirty-to-me-the-ethics-of-conversational-ai/.

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?
In human-human interaction, if one person does not know the answer to a question, we turn to another person. Similarly, if a chatbot is unable to answer the question, there should be the possibility to turn to another source of information, preferably a person. A chatbot should inform the user that the answer to an inquiry is unknown or could not be found, but that should not end the interaction. A chatbot should also not modify an answer to provide misleading or incorrect information. A user should always be given the opportunity to interact with a human when requested.

4.2 In order to gain trust, can the chatbot introduce itself as the user's "assistant / advisor / friend"?
We recommend the use of neutral terms, such as “assistant”. It should be left open to the user to decide how to characterise the relationship with a chatbot. This should not be pre-determined by the designer/developer or company offering the service, as they are not privy to the user’s perspective of use. Furthermore, where chatbots are misleadingly sold as ‘advisors’ and ‘friends’, potentially in the future this would open their manufacturers up to lawsuits where harms occur on the basis of this relationship (e.g., for breach of trust, negligence)

4.3 Free comments: n/a

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.)?
Chatbots should not be trained to lie, misrepresent, or provide misleading information. They should not be equipped, or designers should not pretend that they are equipped, with the nuances of human conversational abilities or abilities to adapt quickly and apply knowledge. There are an infinite variety of human social interactions and types of conversations (intimate, argumentative, apologetic, persuasive etc.) and a chatbot should not pretend to be able to display such diversity of conversational abilities, including that of concealment or lying. If the appropriate response requires nuance or concerns a sensitive topic (like the example provided concerning health), then a chatbot should not be used in the context. Chatbots should direct users to second opinions from human experts where required and especially in cases where harms are likely to occur. A lie told by a chatbot which results in harms could invoke legal liability.

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

5.3 Free comments: n/a

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?
Not all nudges should be allowed. In determining criteria to distinguish between good and bad nudges, a central consideration should be whose interest is furthered by the nudge: that of the user or that of the developer/deployer? Another consideration should be who profits from the nudging. Additional considerations include whether the nudging is obvious, and the consequence of the nudging.

6.2 Does the concept of free and informed consent still make sense when conversational agents nudge?
As a general note, the principle of free and informed consent is always important and there are many contexts where this is a right that must be guaranteed (e.g., health). Chatbots should not undermine this right in any way, including through the use of nudging. As part of informed consent, the user must be informed of the chatbot's limitations and conflicts of interest.

6.3 Free comments: n/a

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?
A chatbot should provide as much information as reasonably useful; this should always include disclosure when there are other choices/options excluded and not presented. The user should also be given the option to access more information. Personalisation of choices should be explicitly noted and explained, and limits on personalisation criteria established.

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?

This is the minimum required; the user should be presented with an easy to understand explanation and option to learn more.

7.3 Free comments: n/a

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.
No, it is not desirable. Considering the infinite variety of human emotions, it is hard to imagine a system that would be effective at this detection. If some accuracy can be reached, we can reasonably expect that it would work for a particular population whose data would be used to train the system, as it has been the case so far with data over-representing white males. Hence, there is a high risk of high level inaccuracy for the category of the population not represented in the data. Errors in such detection systems, particularly when the results are advertised as accurate, could cause significant negative harms (e.g., arrest or detention as a system interprets 'guilty' emotional behaviour). Furthermore, assuming such detection systems could be accurate, there is incredible potential for misuse.

8.2 Is it desirable to build chatbots that simulate human emotions? Answer according to the context of use.
This is also not desirable, as it anthropomorphizes the machine and undermines human-human contact and interaction. For example, a chatbot that simulates human emotions could be employed in the context of elderly care for companionship, and we acknowledge that such use may have benefits. However, we do not see the potential benefits outweighing the harms; this application perversely facilitates and encourages neglect of the elderly and vulnerable, as it may be considered sufficient to replace human-human contact and care.

8.3 Free comments: n/a

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)?
The answer to this depends on context; in all cases the purposes of the interactions must be specific and limited. Governments have legal obligations to protect vulnerable people, including under human rights law. Any time a chatbot's use might impact vulnerable groups, an assessment of the impact must be carried out by developers/deisgners and/or deployers. This should include consideration of the unintended ways a chatbot may be used by different vulnerable groups. All vulnerable groups are not the same. Reliability of the information communicated through a chatbot is a particular concern, especially if a vulnerable person is relying on communication via a chatbot for provision of basic services. Furthermore, the chatbot must be sufficiently tested before deployment in these settings to identify risks and harms. There are contexts where red lines should be established; one such potential red line is the use of chatbots and AI-enabled toys for children.

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? It is a cause of societal concern and should be taken into consideration. We should make sure not to bring about a world in which people only interact with machines. It is essential that we maintain human-human interaction and contact. This is even more so for vulnerable people who tend to be moved out of the social environment, such as the elderly or the disabled people. Where acute attachment to chatbots is likely to occur, these situations should be monitored and assessed for risks (e.g., disassociation from society, loss of human contact, neglect, adverse impact on family relations.

9.3 Free comments: n/a

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[8]*. 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?
No, chatbots cannot 'give life' to a deceased person; they can only generate communications based on established patterns. We don't think that, *in principle*, such uses would violate the principle of respect for the dignity of the human person. However, clear conditions and limits should be put in place before such possibility is offered. It should be the choice of the person (before death) whether they want communications based on their expressions generated post-mortem. There should be parameters (decided before death) about the purpose of the chatbot, the length of time it will be available, who would have access to it, and the communications must be clearly marked as computed generated. Grieving people are a vulnerable group and such chatbots must not be used to feed on their grief or exploit them, if people see such chatbots as an acceptable means of carrying on their loved ones memories or relationships with them.

10.2 How do you see the concept of death evolving with the possibilities offered by chatbots?
We do not see the concept of death as evolving with the possibilities offered by chatbots and we should not pretend it is.

10.3 Free comments: n/a

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.

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

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?
The chatbot should inform the patient of its actions or must have had prior consent to the sharing of the information. A chatbot could be used to monitor activities, but only if the monitoring is set up by a consenting adult and for a limited and specific purpose. The adult can choose the type of activities to monitor and who the information is shared with. Monitoring and sharing of information via a chatbot (or other technology) should never be a prerequisite for access to essential goods and services (e.g., an individual should never be prevented from seeing a doctor and receiving medical care because they are unwilling or unable to track and share information).

11.2 Give other examples of situations in which chatbot monitoring seems justified.
Chatbot monitoring may be appropriate if the application is very limited (e.g., asking an elderly person if medicine has been taken). In all situations, promoting and protecting autonomy should be paramount. Chatbot monitoring should only ever be a complement, never a replacement for human care.

11.3 If it is insulted by its user, should a chatbot inform a third party, its designer, for example?
As discussed in our response to question 3, a chatbot cannot be insulted, therefore we disagree with the premise of the question.

11.4 Free comments: n/a

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?
In any context where a chatbot is used, the quality of the chatbot is paramount. At the moment, many chatbots already in use, such as in public administration or consumer services, are unable to respond meaningfully or efficiently to many requests, leading to high degree of frustration for users. In these situations, the use of chatbot should be discouraged or quality should be improved before being implemented. As discussed in the response to question 4, when a chatbot cannot provide an accurate answer, the user must always have the opportunity to speak to another human.

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.
The response will be different for every profession and context. We must consider whether we want to have chatbots introduced, taking into account public opinion and stakeholder engagement. Proper testing and evaluation should be implemented as well.

12.3 By what means (legislative, code of conduct, etc.) should the use of chatbots be regulated?
As with the governance of any emerging technology, chatbots must be regulated through law that protects against harms and guarantees fundamental rights. Legislative frameworks could be complemented with industry initiatives and civil society engagement, but that is not sufficient to protect against harms. There should be adequate access to complaint and redress mechanisms.

12.4 Free comments: n/a

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? No response.

13.2 What time scale can be envisaged for this evolution? No response.

13.3 Free comments: n/a

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.).
See responses to questions 4 and 5.

14.2 If a chatbot is not able to find an answer, must it say so explicitly?
Yes. See responses to questions 4 and 5.

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.
See responses to questions 4 and 5.

14.4 Free comments: n/a

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*[9]*. 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?
Yes, part of transparency for the user is knowing how the chatbot is generating, and will use its information. However, transparency is not enough; we must consider the values and risks of these systems. Just because we have the ability to develop certain systems and applications, doesn't mean we should deploy and use them.

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? Yes, the designer, developer and/or deployer should share responsibility, if aware that the manipulation is possible, the possibility of misuse is reasonably foreseeable, and the designer/developer took inadequate steps to test, mitigate or warn against the misuse.

15.3 Have you had personal experiences that you interpret as being related to particular chatbot metrics? No response.

15.4 Free comments: n/a

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.

[9] Microsoft Used Machine Learning to Make a Bot That Comments on News Articles For Some Reason

16.1 Should the purpose of a chatbot be revealed to the user? If so, when and in what form? If not, why not?
Yes, and in an easily understandable way (i.e., non-technical terms). The user should have the option to learn more if interested.

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).
See responses to questions 6 and 15.

16.3 Free comments: n/a

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).
Chatbots are made by humans using human data, as such they cannot be unbiased. Users must be made aware of this and developers must be honest about limitations/biases of the technology. Therefore, efforts to identify and minimise bias are still needed on an on-going basis. It is also essential to conduct impact assessment on the short-, medium-, and long-term.

17. Do you think chatbots should mimic human biases or correct them?
Chatbots should not mimic human bias, however they could be used to make visible certain biases (e.g., AI systems revealing human bias in Twitter facial thumbnails).

17.3 Free comments: n/a

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.
See response to question 4.

18. Do chatbots' mistakes elicit different feelings or reactions than human mistakes? Which ones?
See response to question 4.

18.3 Free comments: n/a

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).
See response to question 4.

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?
The answer to this question depends on the context and nature of the chatbot and should be a case-by-case consideration.

19.3 Free comments: n/a

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?
While this may be 'acceptable', the real question is whether it is desirable, given that such a response would undermine the efficiency and usefulness of the chatbot.

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). No response.

20.3 Free comments: n/a

Thank you for your contribution!

Send it to: cnpen-consultation-chatbots@ccne.fr