Arisa Robot
Shinjuku

Lovot video1 video2

About Chatbots

Some Ideas about Ethics and AI

{Jean-Marc.Vincent}@univ-grenoble-alpes.fr

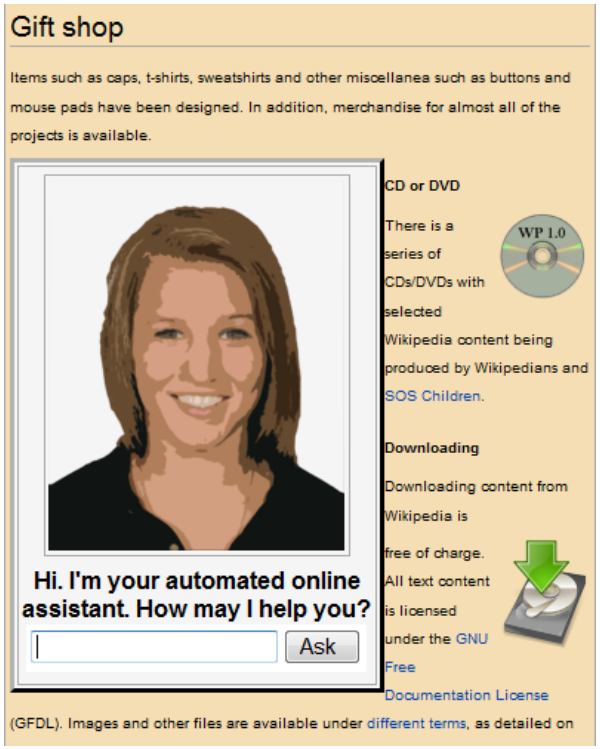
Université de Grenoble-Alpes, UFR IM^{2}AG
M2R-MOSIG Scientific Methodology

December 2024

ABOUT CHATBOTS

CHATBOTS

A Personal Assistant



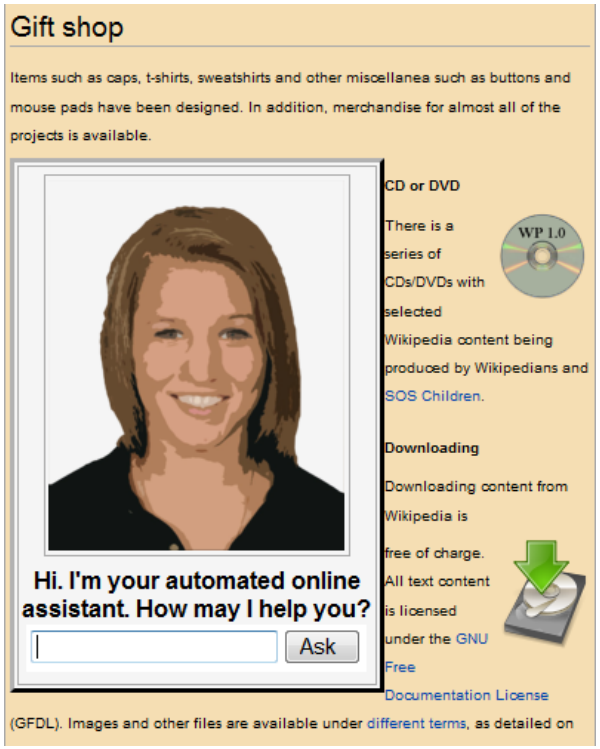
Wikipedia

Human-Computer Interaction

- Chatbot : Voice and/or Video
- Human : Voice and/or Video
- Natural Language and expressions
- one-to-one dialog

CHATBOTS

A Personal Assistant



Wikipedia

Human-Computer Interaction

- Chatbot : Voice and/or Video
- Human : Voice and/or Video
- Natural Language and expressions
- one-to-one dialog

Exercise

1. Find some examples of Chatbots and choose one of them
2. Describe the technical characteristics of the chatbot
3. Describe the context of the chatbot and the objective of the chatbot

ABOUT CHATBOTS

ETHICS POINT OF VIEW

Exercise

1. Which ethical problems appear in this context ?
2. Who is concerned ?
3. What should be done to avoid these problems ?
4. Try to classify these ethical problems ?

ETHICS POINT OF VIEW

Exercise

1. Which ethical problems appear in this context ?
2. Who is concerned ?
3. What should be done to avoid these problems ?
4. Try to classify these ethical problems ?

Translation (approximate) Comité National Pilote d'Éthique du Numérique
Recommendation 3 *Agents conversationnels : Enjeux d'éthique* texte français - in english

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.

ABOUT CHATBOTS

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

Fields for ethical questions (Principles)

- Proportionality and Do No Harm

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

Fields for ethical questions (Principles)

- Proportionality and Do No Harm
- Safety and security

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

Fields for ethical questions (Principles)

- Proportionality and Do No Harm
- Safety and security
- Fairness and non-discrimination

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

Fields for ethical questions (Principles)

- Proportionality and Do No Harm
- Safety and security
- Fairness and non-discrimination
- Right to Privacy, and Data Protection

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

Fields for ethical questions (Principles)

- Proportionality and Do No Harm
- Safety and security
- Fairness and non-discrimination
- Right to Privacy, and Data Protection
- Human oversight and determination

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

Fields for ethical questions (Principles)

- Proportionality and Do No Harm
- Safety and security
- Fairness and non-discrimination
- Right to Privacy, and Data Protection
- Human oversight and determination
- Transparency and explainability

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

Fields for ethical questions (Principles)

- Proportionality and Do No Harm
- Safety and security
- Fairness and non-discrimination
- Right to Privacy, and Data Protection
- Human oversight and determination
- Transparency and explainability
- Responsibility and accountability

AI RECOMMENDATION OF UNESCO

Fields for ethical questions (Values)

Draft text of the Recommendation on the Ethics of Artificial Intelligence

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

Fields for ethical questions (Principles)

- Proportionality and Do No Harm
- Safety and security
- Fairness and non-discrimination
- Right to Privacy, and Data Protection
- Human oversight and determination
- Transparency and explainability
- Responsibility and accountability
- Multi-stakeholder and adaptive governance and collaboration