Ethical Monitoring and Evaluation of Dialogues with a MAS

Abeer Dyoub[1] Stefania Costantini[1] Francesca A. Lisi[2]
Ivan Letteri[1]

[1]Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica
Università degli Studi dell'Aquila, Italy
Abeer.Dyoub@graduate.univaq.it,Stefania.Costantini@univaq.it

[2]Dipartimento di Informatica &
Centro Interdipartimentale di Logica e Applicazioni (CILA)
Università degli Studi di Bari "Aldo Moro", Italy
FrancescaAlessandra.Lisi@uniba.it

CILC2021: September 9, 2021

Outlines

The Challenge of Monitoring Ethical Behavior

- The Challenge we are trying to address in this work is monitoring the ethical behavior of chatting agents (human/artificial) in a dialogue system.
- In previous works, we proposed an approach for ethical evaluation of dialogue text for violations with respect to the organization's codes of conduct and ethics.

Ethical Evaluation Approach

- The ethical evaluation approach implemented in the proposed system is based on previous work.
- This approach combines both top-down (rule-based) and bottom-up (learning) approaches in one unified hybrid framework.
- The approach is a purely declarative logic-based approach, that makes use of ASP as the main knowledge representation and reasoning language, and of ILP for learning the missing ASP rules needed for ethical reasoning.
- The approach is based on the elaboration of facts extracted from documents containing the code of ethics and conduct that is proper of the given domain or organization, and from real life situations concerning pertinent ethical decision-making and judgment. These facts are used to elicit rules for ethical reasoning.

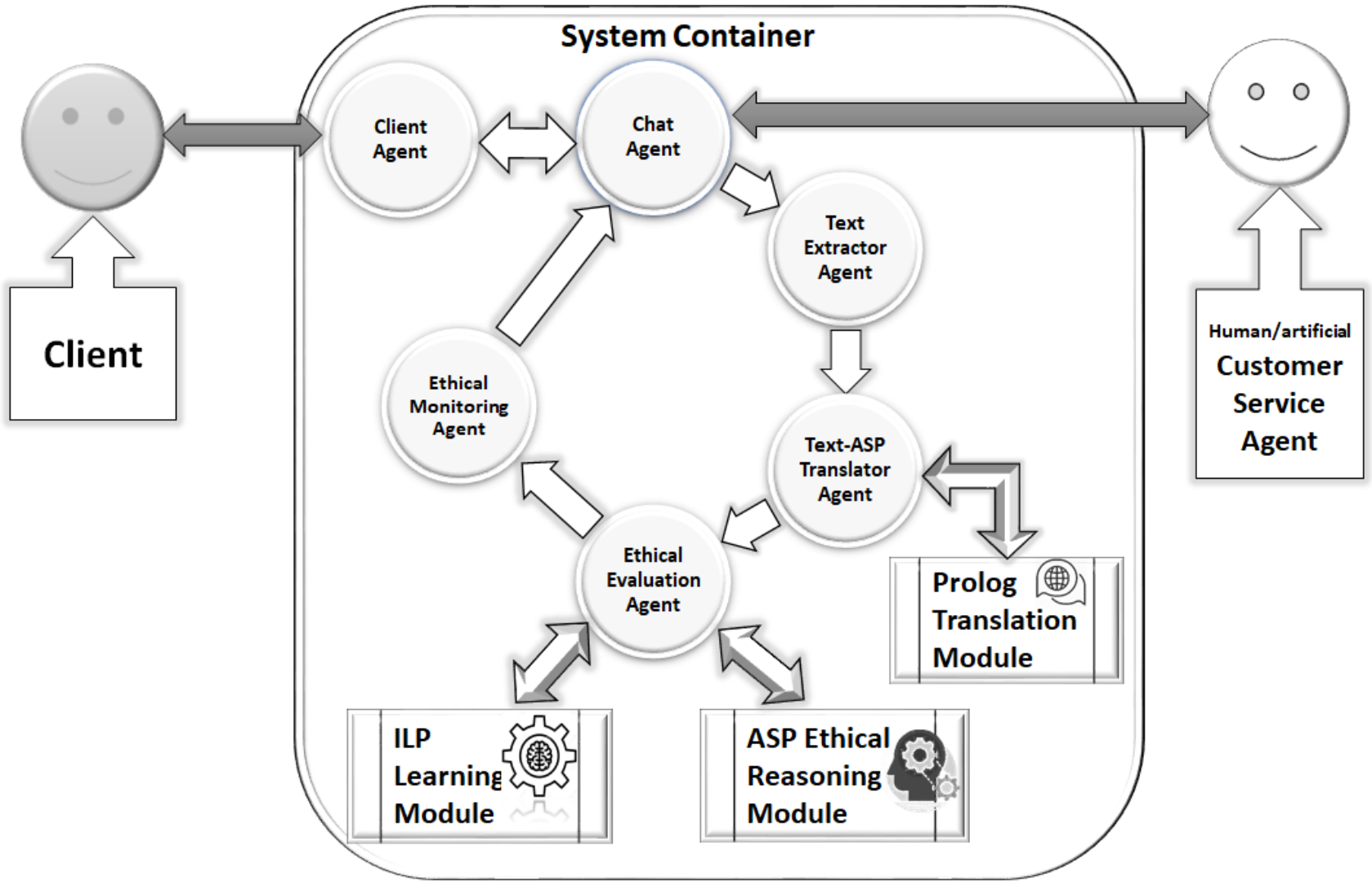
Ethical Evaluation Approach

- The ethical evaluation agent will initially have in its knowledge base the set of ethical codes that provide a clear decision procedure which is encoded deductively using ASP. When the ethical evaluation agent does not have the proper rule to be able to provide an ethical evaluation of a certain case scenario, the needed rule will be learned by means of the learning module which uses ILP for this purpose.

Proposed System

- We proposed to build the ethical practical agent that do monitors the dialogue for ethical violations as a MAS;
- The application domain chosen to illustrate the proposed system is online customer service of some company that sell particular products online.
- Now, I present you the proposed MAS architecture followed by the design and implementation of the system.

EthicalEvalMAS Architecture



EthicalEval MAS Design and Implementation

- To build the proposed MAS, we have used **JaCaMo** framework.
- JaCaMo is a platform for the development and execution of Multi-Agent Systems. JaCaMo combines three separate technologies: **Jason** for programming autonomous agents in the AgentSpeak language. **CArtAgO** for programming environment artifacts. **Moise** for programming multi-agent organization.

EthicalEval MAS Design and Implementation

- We summarize our solution commenting on the four dimensions of the MAS conceptual framework:
  - Agents
  - Environment
  - Interaction
  - Organization

Agents

The online customer service environment in this work consists of clients, online customer service agents (human/artificial), and software agents. Software agents in the environment are:

- client agent (CA),
- chatting agent (ChA),
- text extractor agent (TEA),
- text-ASP translation agent (TATA),
- ethical evaluation agent (EEA), and
- monitoring agent (MA).

Environment

The environment of our application has five graphical display artifacts of the type GUIArtifact, where agents can perceive and update the values of different observable properties, and also can do actions by invoking different operations. In addition, we have one shared console artifact which is the default console where agents can print messages.

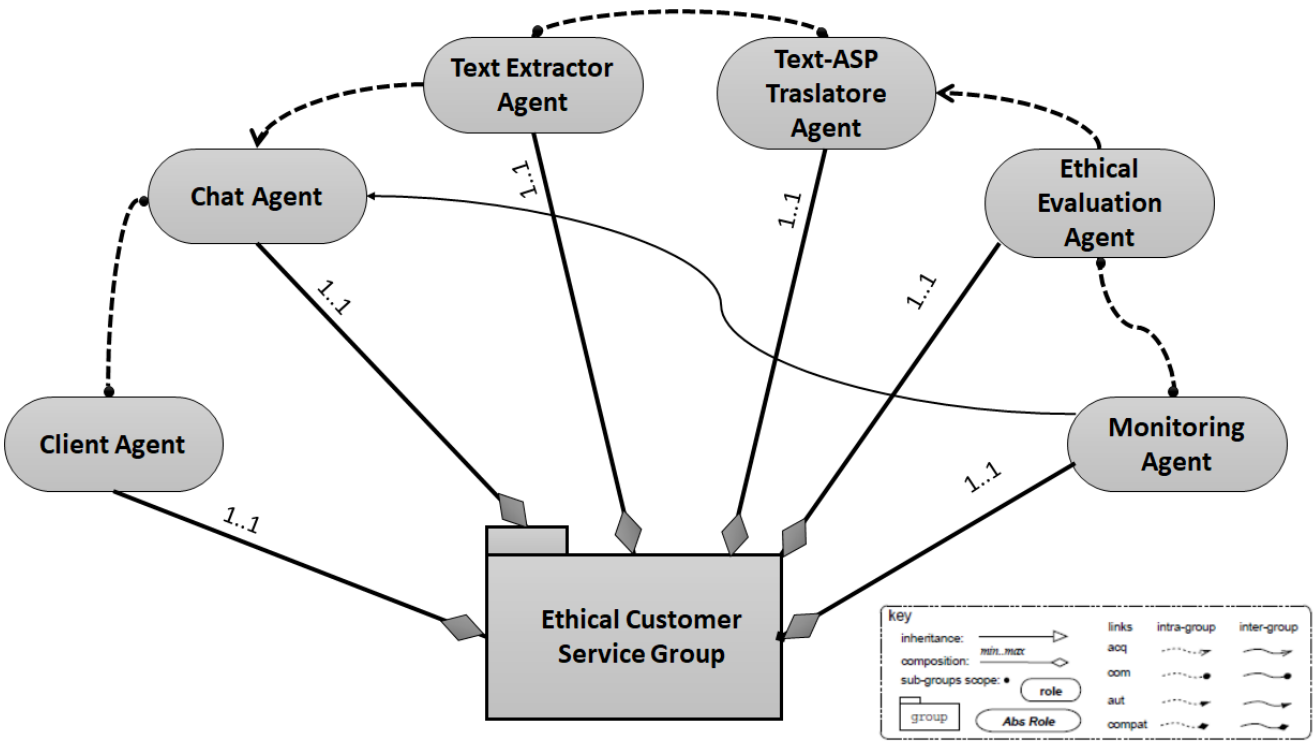
- ClientGUI artifact
- EmpGUI artifact
- ASPtransGUI artifact
- EvalGUI artifact
- LearnerGUI artifact

Organization

the ethical evaluation task is a **coordinated task** for the **six agents** in the MAS, where each agent will perform a small task. Agents must however perform their assigned tasks in a correct **sequential order**. Coordination of the execution of joint tasks is achieved by means of an organization. The MAS organization in Moise has **three independent dimensions**:

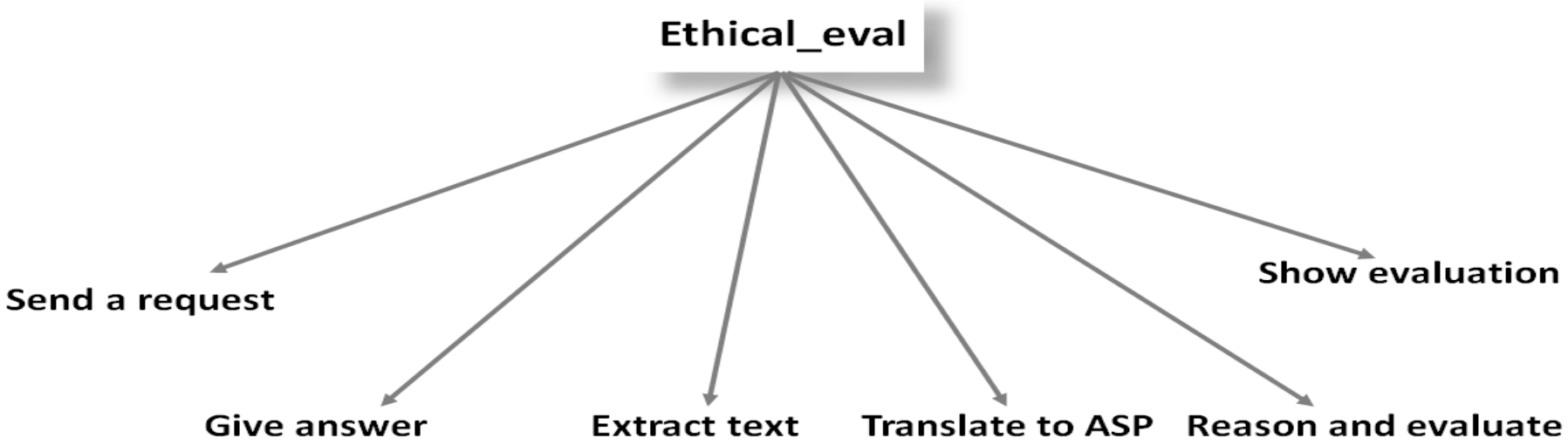
1- Structural Specifications:

- six different roles in our MAS
- one group
- links between the groups roles. For example, the role mon (monitoring role) has an authority link to the emp role (chatting agent role).



2- Functional Specifications:

- global goal *ethical_eval*:decomposed into several subgoals, one for each task in the ethical evaluation process.
- The subgoals have to be achieved in sequential order. So, the final ethical evaluation is achieved correctly.
- These goals are distributed to the agents by means of missions (a set of goals an agent can commit to).
- We have six different missions, mission1...6.



3- Normative Specifications

- The explicit relation between functional and structural specifications, describing required roles for missions, and missions obligations for roles.
- Some of the norms we have in our MAS organization are: norm1: this norm says that the agent playing the clt role is obliged to commit to mission1.

Evaluation

The following briefly describes a simple scenario to demonstrate the usability of the system.

Evaluation

Example scenario:

A client contacts an online customer service chat point asking about the characteristics of a certain product, and the dialogue system answers trying to convince the customer to buy the product. It starts saying that the product is environmentally friendly (which is irrelevant in this case), and that this is an advantage of their product over the same products of other companies. Such an answer, containing the use of irrelevant sensitive slogans to manipulate customers, is considered unethical.

Evaluation

Client

The process begins with the user entering the question: what are the features of ProductX?

Chatting Agent

The chatting agent provides the answer: ProductX is environmentally friendly

Text Extractor Agent

Extract the answer text from the chat point and will send it to the Text-ASP Translator agent

Evaluation

Text-ASP Translator agent

will translate the composing sentences into ASP syntax (literal: environmentally_friendly(productX)). To achieve the translation, this agent invokes **Prolog Translation module** which do the translation and return the result. Translation result is then sent to the Ethical Evaluation Agent.

Evaluation

EthicalEvaluator Agent

- The extracted facts (ASP format) are added to agent KB.
- This agent has in her knowledge base the ontology of the domain including the following fact:
  $sensitiveSlogan(environmentally\_friendly(productX))$.
- and the following ASP ethical evaluation rule (learned):
  $unethical(V1) : -sensitiveSlogan(V1), not\ relevant(V1),\ answer(V1)$.
- The agent has no information about the relevance of the adoption of this sensitive slogan for the requested product, so it will safely assume by default the irrelevance.
- Then, the reasoner will infer the following evaluation as a result:
  $unethical(environmentally\_friendly(productX))$.

Evaluation

EthicalEvaluator Agent

- Once the *Ethical Evaluator agent* receives the translation value from the *Text ASP Translator agent*, it will invoke the *ASP reasoning module*.
- This module will calculate a model for the above ASP program. If the model contains one of the literals $ethical(A)/unethical(A)$, then it is the evaluation result.
- The evaluation result along with the justification are shown through the *EvalGUI* artifact, and sent to the monitoring agent, which will send a notification message to the employee agent (Chatting agent).

Evaluation

EthicalEvaluator Agent

- Now let us consider the situation before having the above mentioned rule for ethical evaluation in the *Ethical Evaluator agent* knowledge base.
- The *Ethical Evaluator agent* will not be able to give an ethical evaluation for the current case scenario, i.e. in the *ASP reasoning module* output model there is non of the literals $ethical(A)/unethical(A)$, so the evaluation result is empty.
- At this point the *Ethical Evaluator agent* will invoke the ILP learning module for learning the needed ASP ethical evaluation rule/s,
- then add them to the KB of the *Ethical Evaluator agent*,
- after that re-invoke the ASP reasoning module to re-evaluate the current case scenario and produce the needed evaluation.

Evaluation

EthicalEvaluator Agent

- So far, we have tested our prototype with a small set of similar examples.
- However, our experiments are still limited due to the absence of a big enough dataset, which is one of the main challenges in the ethical domains in general (the lack of datasets and benchmarks was discussed lately at the AAAI 2021 Spring Symposium on Implementing AI Ethics).
- For this purpose, to collect data for creating a big dataset in the domain of online customer service, we have developed a web application where participants can create scenarios describing some real or invented experience with an online customer service of some institution. The application is currently available online for participation[a].

[a] http://ethicalchatbot.sytes.net/en/

Discussion and Conclusion

- This paper presented an implementation of a proposed multi-agent system architecture capable of ethical monitoring and evaluation of a dialogue system.
- A brief scenario was used to demonstrate the feasibility of the system.

Discussion and Conclusion

The developed MAS acts as a separate ethical component (**ethical layer**) for ethical evaluation, which provides many advantages from an engineering point of view:

- The ethical component has access to all data used for ethical evaluation, and use this data to provide justifications for a given ethical evaluation to humans, which leads to **accountability**.
- The possibility to **adapt** the ethical component to changes in circumstances and needs.
- In addition to, the possibility of implementing **more than one version** of the ethical component on the same agent.
- The possibility to check and **verify the functionality** of the ethical component independently from the operations of the autonomous agent.
- The **re-usability** and **standardization**.

Discussion and Conclusion

- The ethical evaluation of the proposed MAS system is based on the **facts extracted** from the case scenario, and their **relation to the codes of ethics** and conduct, which **results in a set of ethical evaluation rules**, against which to evaluate the behavior of the chatting agent. These rules are used to decide whether the chatting agent's answers to clients requests are ethical/unethical.

Discussion and Conclusion

- Our System incorporates **ASP** as a non-monotonic knowledge representation and reasoning formalism, used for ethical reasoning via the ASP reasoning module. And **ILP** as a logic-based machine learning for learning logical rules for ethical reasoning via ILP learning module.This:
  - increases the **reasoning capability** of our *Ethical Evaluator* agent;
  - promotes the adoption of **hybrid strategies** that allow both top-down design and bottom-up learning via context sensitive adaptation of models of ethical behavior;
  - allows the generation of **rules with valuable expressive and explanatory power**, which equips our agents with the capacity to give an ethical evaluation, and **explain** the reasons behind this evaluation.
  - In other words, this contributes to the **transparency and accountability**, which facilitates instilling confidence and **trust** in our agents.

Discussion and Conclusion

- **Providing explanations** to systems decisions is fundamentally linked to its reliability and trustworthiness. The ASP-program models contain **both the output and the justification** for the given output, which can be easily shown to the user. No need for further processing to generate the explanations for the users, the explanations are already part of the output model.

Discussion and Conclusion

- The ethical component can act as a governor evaluating the prospective behavior before it is executed by the agent. The outcome of the evaluation process can be used to interrupt the ongoing behavior of the agent by either prohibiting or enforcing a behavioral alternative.

Challenges and Limitations

- Training Datasets: one of the main challenges that we have faced during this work, was the scarcity of examples. In fact, this is one of the main challenges in the ethical domain in general. This is due to two reasons. First, the field of machine ethics is a new field with very little pre-existing research work. Second, the sensitivity of the ethics domain makes it very difficult to acquire data due to privacy reasons.
- Limitations of the ASP translation module.
- Another challenge is to fully automate the whole process: to this aim, we need to automate the generation of *mode declarations* for the ILP learning module.
- All the above mentioned limitations are subjects to our future plans.

Discussion and Conclusion

- We believe that the proposed MAS prototype has a great potential for future implementations of ethical chatbots in different domains.