Improving the Explainability of Autonomous and Intelligent Systems with Conversational Interfaces

Nico Hochgeschwender

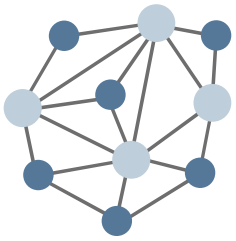
Simulation and Software Technology

Intelligent and Distributed Systems

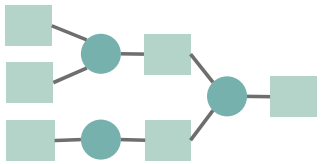
Intelligent Software Systems

German Aerospace Center (DLR)

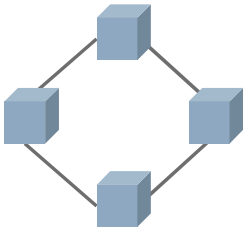
Intelligent and Distributed Systems – Research Topics



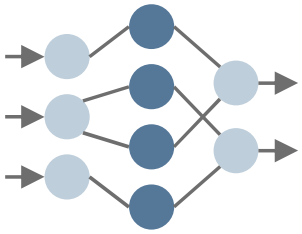
Distributed Systems



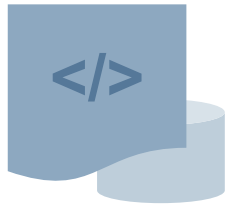
Workflows & Provenance



Blockchains



Machine Learning



Software Engineering

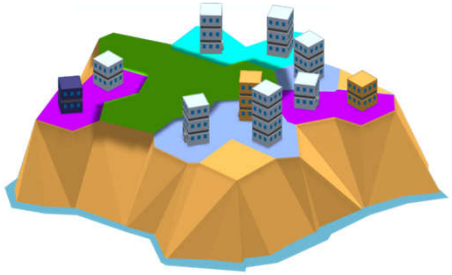
Software Analytics

Improving the Understandability of Software Architectures

Island Metaphor for Visualizing OSGi-based Software Architectures in Virtual Reality

Classes
Multi-storey **buildings** with a new storey for every n lines of codes

Packages
Continuous **regions**



Bundles
Islands with multiple regions; each island with distinct shape

M. Misiak, D. Seider, S. Zur, A. Fuhrmann, and A. Schreiber, "Immersive Exploration of OSGi-based Software Systems in Virtual Reality," 25th IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR 2018), Reutlingen, Germany, 2018.

Improving the Explainability of an Autonomous Office Assistant

Autonomous Office Assistant*

Autonomous mobile service robot capable of performing a **wide range** of tasks over **a long-period** of time in everyday environments.

Examples of tasks are:

- *Patrolling* the building for checking WiFi signal strenght,
- *Escorting* visitors to meeting rooms, and
- *Informing* visitors about the institute.

Task should be carried out preferably **24/7**

We build up on **long-term autonomy abilities** developed in related projects

* Collaboration with University of Luxembourg

Making the Case for Transparent and Explainable Robots

- What do we mean by transparency and explainability for autonomous systems?
- *A system is considered to be transparent if it is possible to discover why it behaves in a certain way, for instance, why it made a particular decision while explainability defines to which extent the internal state etc. accessible to the user through provision of e.g. human-like language.**

Example: Explainable Autonomous Office Assistant

Nico:	**What actions did you consider when escorting Bob today?**
Robot:	*To escort him to your office or to show on a map where your office is.*
Nico:	**Which choice did you make and why did you make it?**
Robot:	*To escort him by, because Bob has never been to your office.*
Nico:	**What did you expect to happen after you reached my office?**
Robot:	*...*

"After-action review" inspired by Pat Langley [4]

*IEEEs WG P7001 in the context of the initiative for Ethical Considerations in AI and Autonomous Systems

Making the Case for Explainable Robots (cont.)

- Why is transparency important?
  - Humans need to understand what robots are doing.
  - Without this understanding humans will not **trust** robots.
- **Standards** are required to foster the development of transparent robots
  - P7001 Transparency for Autonomous Systems*

Basic proposition in P7001 "…it should be always be possible to understand why and how an autonomous system made a particular choice…[6]".

To this end, P7001 aims to:

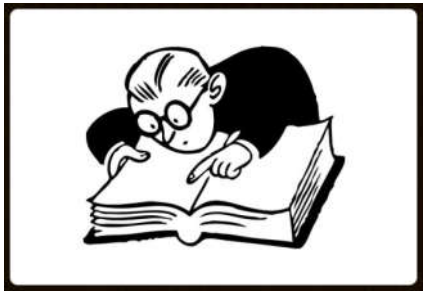
- Identify *stakeholders* (e.g. users, lawyers, certification agencies) and their transparency requirements, and
- Describe *measurable*, *testable levels of transparency*, so that autonomous systems can be objectively assessed and levels of compliance determined.

*IEEEs Initiative for Ethical Considerations in AI and Autonomous Systems

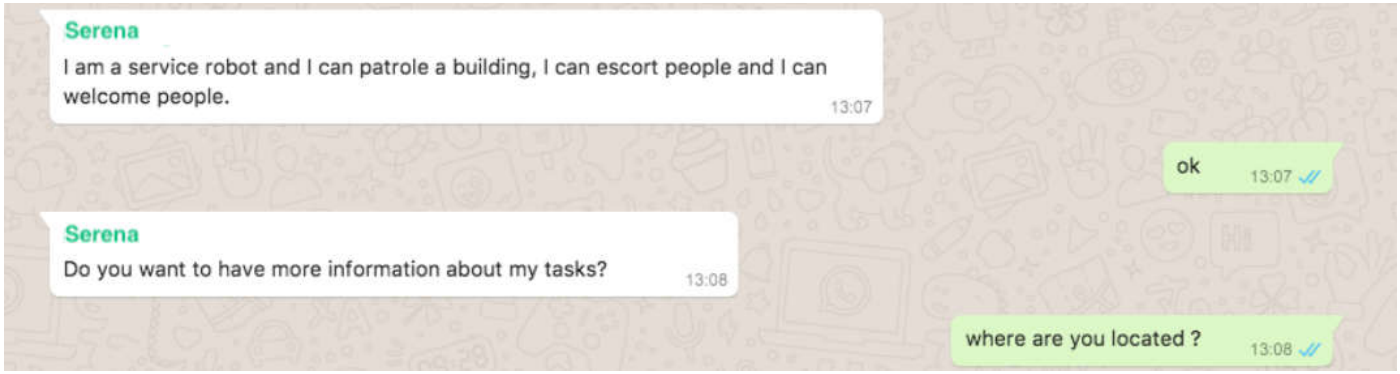
Research Challenges for Explainable Autonomous Systems

- To implement explainable robots the following constituents are required:
  - **Representations:**
    - e.g. about the world, robot abilities and their limitations and their planning and decision making process
  - **Reasoning:**
    - e.g. methods deriving the reasons for selecting decisions over alternatives.
  - **Human-Robot Interfaces:**
    - e.g. means to communicate the decision making process in a way that makes contact with human concepts.

Conversational Interface for Autonomous Office Assistant



- What can the autonomous office assistant actually *do* for me?
  - Reading an *instruction manual* is not an option as:
    - new tasks are added to the robot,
    - execution of tasks depends on the dynamic environment, etc.
- **Idea:** Develop a conversational interface (e.g. WhatsApp chatbot) which users can exploit to learn about the abilities of the autonomous office assistant.
  - Advantage: A chatbot is accessible anytime and everywhere



Research Design and Methodology

- **Research objective:** To derive a *functional specification* for a conversational interface which makes the abilities of a mobile service robot transparent.
- Research questions:
  - **RQ1:** Do different users ask the chatbot similar questions for solving their task at hand?
  - **RQ2:** What types of questions did users ask during the conversation?
- *Wizard-of-Oz* experiment using WhatsApp:
  - 12 participants (researchers and non-scientific stuff).
- We employed Conversation Analysis (CA) [7] to analyze the dialogues.

Excerpt of the Results

- Users perceive the chatbot **not** as a separate speaker:
  - Interaction can be analyzed as a dialogue and not as a three-party conversation.
- Speech-exchange system of the chats can be considered as an interview:
  - 77% of the users turns are questions, but many of the are not related to the robot abilities. For example:
    - “Do you follow the WorldCup?”
    - “Can you tell me how the weather will be tomorrow?”
- Questions can be categorized along four categories, namely questions related to the robot’s
  - *abilities* and properties which do not change,
  - *past experiences* or tasks,
  - *current tasks*, and
  - *future plans* and activities.

Excerpt of the Results (cont.)

- Transparency vs. Privacy/Security
  - The office assistant is embedded in the real-world, thus there are also questions about other entities such as people and facilities. For example:

User:	**Is Bob in his office right now?**
Bot:	*I do not know.*
User:	**Can you check this for me?**
...	
User:	**Are you able to take a picture of his office and send it to me?**

User tries to exploit the robot as a spy

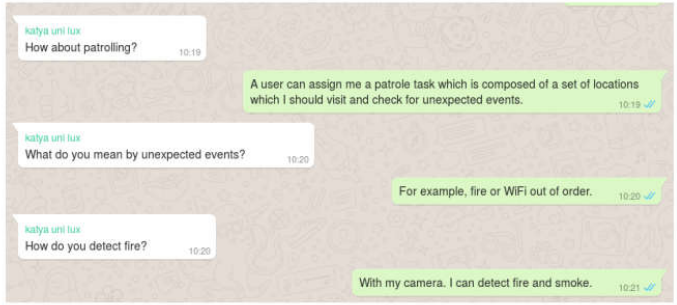
- Experiments revealed the necessity to carefully investigate security/privacy risks:
  - **Asset:** Personal right of Bob
  - **Threat:** Picture will be sent to User
  - **Vulnerabilities:** Wrong access management to sensor data for this particular task

G. Cornelius and N.Hochgeschwender et al., "A Perspective of Security for Mobile Service Robots" Advances in Intelligent Systems and Computing. 2017.

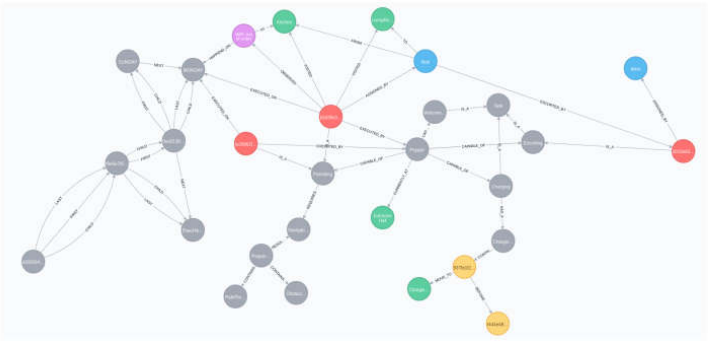
Current Implementation

- IBM Watson for natural language understanding (intent classification):

Intent	Utterance
explain_patrolling	*How about patrolling?* *What is patrolling?* ...
robot_experience	*Have you ever observed fire?* ...



- Knowledge representation: Graph-based knowledge base of robot abilities and experiences.



N.Hochgeschwender et al., "Graph-based Software Knowledge: Storage and Semantic Querying of Domain Models for Run-Time Adaptation" IEEE Conference on Simulation, Modeling and Programming for Autonomous Robots. 2016.

- Reasoning: Semantic queries on the graph taking into account contextual information of the current dialogue (e.g. conversation history, stored variables, etc.)

Conclusion and Future Work

- Research Challenges and Opportunities:
  - Transparency and Explainability vs. Privacy/Security
    - How do we build transparent and explainable systems which at the same time fulfill privacy and security requirements?
  - User-driven approach
    - How does a system fulfill explainability requirements for different stakeholders?
  - Group chats for robots
  - Conversational approaches seems to be a feasible interface also for other use cases such as inspecting large amount of log-files generated e.g. by robots over a long-period of time.

Potential Links with Participants

- Francisco J.C. Garcia: *Conversational Interfaces*
- Christopher Gerking: *Security*
- Narges Khakpour*: Security*
- Andreas Wortmann: *Robotics*
- Simos Gerasimou: *Robotics*
- *…*

Thank You!

Questions?

Nico.Hochgeschwender@dlr.de
www.DLR.de/sc/ivs | @nico_roboticist

DLR