Deliverable: D5.1

Functional requirements, interactions and constraints

Consortium

University of Twente (UT)
Imperial College London (ICL)
University of Passau (UP)
Institute of Mathematics of the Romanian Academy (IMAR)
University College London (UCL)
Serbian Society of Autism (SSA)
Autism Europe (AE)
IDMind (IDM)

Grant agreement no. 688835

Research and Innovation Action
<table>
<thead>
<tr>
<th>Document Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deliverable number:</td>
</tr>
<tr>
<td>Deliverable title:</td>
</tr>
<tr>
<td>Date due:</td>
</tr>
<tr>
<td>Actual submission date:</td>
</tr>
<tr>
<td>Work Package number and title:</td>
</tr>
<tr>
<td>Lead partner:</td>
</tr>
<tr>
<td>Participating beneficiaries:</td>
</tr>
<tr>
<td>Authors:</td>
</tr>
<tr>
<td>Reviewers:</td>
</tr>
<tr>
<td>Dissemination level:</td>
</tr>
<tr>
<td>Nature:</td>
</tr>
<tr>
<td>Draft/final:</td>
</tr>
<tr>
<td>No. of pages including cover:</td>
</tr>
<tr>
<td>Keywords:</td>
</tr>
</tbody>
</table>
Table of Contents

1. Summary........................................................................................................................................4

2. Scenario Definition of DE-ENIGMA..........................................................................................5
   2.1. Robots and ASC..................................................................................................................5
   2.2. User Requirements..............................................................................................................6
   2.3. DE-ENIGMA Methodology..................................................................................................8
   2.4. Identification of ethical issues..............................................................................................11

3. System Dependencies..................................................................................................................12
   3.1. Shared Vision Workshop Results.......................................................................................12
   3.2. Main Modules......................................................................................................................19

4. System Technical Specifications..................................................................................................25
   4.1. System Overview................................................................................................................25
   4.2. Hardware..........................................................................................................................25
   4.3. Perception.........................................................................................................................27
   4.4. Reasoning..........................................................................................................................28
   4.5. Human Robot Interaction....................................................................................................28
   4.6. Integration Framework..........................................................................................................28
   4.7. Data Collection System Architecture...................................................................................29

5. Methodology of Development.....................................................................................................31

References.........................................................................................................................................32
1. Summary

The DE-ENIGMA project aims to create and evaluate the effectiveness of a robot-based technology, developed to support autistic children in their learning. The Consortium will design effective and user-adaptable styles of robot behaviour for autistic children, leading to more personalised and effective therapies than previously available and deliver a therapy-based interactive game for robot assisted therapy.

During the first months of the project the partners have been concentrating their efforts on the definition of the scenario related with the data collection experiments (WP1), the envisioned interactions between the different tasks and the required technology to accomplish it.

This deliverable addresses the general specification of DE-ENIGMA taking into account the performance needs in terms of sensing, actuation, communications, processing and energy, to address the project’s scientific and technological challenges. Its main objective is to serve as a guide for the developments that are taking place in the different work-packages.
2. Scenario Definition of DE-ENIGMA

Work on the scenario for the DE-ENIGMA robot assisted therapy-based interactive game begins in M12 in Task 4.2: Design and Realization of interactive game for training socio-emotional skills – the first prototype of which will be delivered in M18. This will develop out of the observations of therapies in the data collection experiments in WP1 and the study being carried out on Robot-Child interaction in Task 4.1.

Currently several therapy approaches are used to teach social skills to children with Autism Spectrum Conditions (ASC). Curricula such as Do Watch Listen Say (Kathleen Ann Quill) for example may help facilitate developing individualized programs to improve social success. Moreover Early Start for Your Child with Autism by Sally Rogers, Geraldine Dawson, and Laurie Vismara provides caregivers with many techniques to aid in the development of social-communication and learning through everyday activities. The current project uses an approach by Howlin, Baron-Cohen and Hadwin to teach children with ASC social skills.

2.1. Robots and ASC

ASC are characterized by social communication difficulties and restricted and repetitive behaviour patterns. Alongside their difficulties, individuals with ASC tend to have intact and sometimes superior abilities to comprehend and manipulate closed, rule-based, predictable systems, such as robot-based technology. Over the last couple of years, this has led to several attempts to teach skills to individuals with ASC, using humanoid robots. Using robotic intervention has been shown to improve motor performance and imitation of actions in children with ASC for example (Srinivasan et al., 2015). A recent systematic review of social robotics in ASC also highlights that children with ASC tended to perform better with a robot compared to another human, showed reduced repetitive behaviours and robots seemed to elicit social behaviours (Penissi et al., 2016). Taken together, even though in its infancy, using robots for therapy has been shown to be very effective as an integral part of the psychoeducational therapy for children with ASC. The main reason for this is that humanoid robots are perceived by children with autism as being more predictable, less complicated, less threatening, and more comfortable to communicate with than humans, with all their complex and frightening subtleties and nuances. Also, children with autism tend to develop a far better relationship with lifelike robots than with screen-based technology, like tablets and touch-screens, because robots act in a more human way in 3D rather than being 2D forms or shapes (Farr et al, 2010; Pop et al., 2013).

Humanoid robots have also been used to improve the emotional skills of children with autism, in particular their ability to recognise emotions in others. For example, the Milo robot was used to enhance children's ability to recognise happiness, sadness, anger, and fear, which Milo would portray with facial expressions and the child would be asked to identify the correct emotion by selecting the answer from multiple choices on an iPad3.

Furthermore, Zeno (Milo's ‘brother’) was used in another study to investigate the effect of incorporating bodily gestures to the emotion expression prediction accuracy of typically developing (TD) children and children with ASC, and it was shown that the gestures could significantly impact the prediction accuracy of both ASC and TD children in a negative or
positive manner depending on the specific expression, evidencing the relevance of the use of gestures for conveying emotional expressions by a humanoid robot in a social skill therapy (Salvador et al, 2015). The Kaspar robot is another similar example (Wainer et al., 2014a; Wainer et al., 2014b), as well as other less anthropomorph robots that do not have a lot of facial manipulation possibilities like Nao, Tega, Probo, Reetie and Popchilla. However, virtually all of these robots studies were based on remote-control-based HRI. Audio-visual technologies that could robustly and accurately track and recognise children’s facial, bodily, and vocal behaviour and naturalistic interactions ‘in the wild’, and react appropriately based on the observed child’s behaviour, as observed by a webcam integrated into a humanoid robot and used in unconstrained recording conditions, have not been available so far, which has made it difficult to facilitate the use of humanoid robots for assessment of children’s ability to express appropriate behaviour and improve their emotional and social imagination skills. Furthermore, the main focus of these robots in therapy of children with ASC is on enhancing their ability to recognise emotions in others but not on their ability to respond to these emotions appropriately. This latter skill is rarely targeted in the experimental studies and it involves a higher-level understanding of emotions in context.

The DE-ENIGMA project aims to create and evaluate the effectiveness of such a robot-based technology, directed for children with ASC.

2.2. User Requirements

One of the main aims of the DE-ENIGMA project are to test whether autistic children’s ability to learn socio-emotional (facial, bodily, vocal) skills within a direct-instruction programme is better during a robot-assisted than a therapist-only session, when performed in unconstrained recording conditions and when using all three modalities (face, body, voice).

The need is to match the teaching procedure across robot-delivered and therapist-delivered sessions and to integrate facial and bodily expressions.

In order to be able to do so, we will use a robot that mimics an articulated human face and is complemented with realistic gestures and movements.

Furthermore, autistic children perceive differently, seeing details rather than a whole. Some children cannot focus both on facial expression and gestures, so for their learning process with the robot, it is important to integrate these two. More specifically, because autistic children are so different from one another, some of them pay greater attention to facial expressions while others focus on gestures (this can depend on their interests, motor skills etc.)

The Consortium has considered different off-the-shelf robot options for the project, some of them are depicted in Figure 1. But, for the aforementioned reasons, we opted for Robokind’s Zeno R25, also known as Milo (see Figure 2).
Figure 1. Considered robot options: NAO by Aldebaran (top left); Q.Bo by TheCorpora (top right); Darwin-OP by ROBOTIS (bottom left); and reeti by Robopec (bottom right).

Figure 2. Robot platform chosen for the project: Zeno R25 (*aka* Milo) by Robokind.
2.3. DE-ENIGMA Methodology

During the first months of the project data will be collected to use to develop novel technologies for automatic audio-visual analysis of human behaviour to be integrated into the actual exploitable product: the fully automatic, context-sensitive, humanoid-robot setup (to use with the interactive game) for teaching autistic children emotional and social imagination skills. Obviously the final product will evolve from the data collection setup as the same input channels will be necessary for automatic analysis in the robot assisted therapy.

The methodology has been defined for the data collection experiments. For these data collection sessions, the robot interactions are based on a Wizard-of-Oz setup. This methodology will be adapted according with the robot prototype developments during the project.

Overall aim: To enhance autistic children’s social and emotional skills through human-robot interaction.

Overall procedure: To examine autistic children’s responses to the robot- and therapist-led sessions, we have followed Howlin et al.’s (1998) approach to teaching perception, expression, understanding, and social imagination related to 4 affective states: happiness (positive valence), sadness (negative valence, low arousal), anger (negative valence, high arousal) and fear (negative valence, high arousal). In this basic scenario, we therefore focus on direct-instruction strategies (i.e., directly teaching emotional displays), rather than on implicit learning (i.e., indirectly teaching about emotions through play) targeted by WP4-5. For example, now [in M6] in Belgrade, 26 children have been administered (and recorded) during the following six phases of teaching these social and emotional skills (in 10-15 minutes daily sessions), which become progressively more challenging, moving from simple matching to identification to more complex social understanding.

There are multiple trials in each session within each phase. Table 1 lists the results completed by each child during the 1st data collection experiments.

The exact duration of each phase will differ for each child, depending on his/her individual learning rates, which can vary widely. Children will move on to the next phase once they have reached criterion (see below) and is maintained across at least 2 days. In those cases in which a child cannot reach criterion, he or she will not take part in further phases. The six different steps are listed below have been tried out for all children enrolled in the first round of testing in Belgrade.
Table 1. First round of participants in Belgrade. *Last session of repeating the last achieved step is also calculated here. For number of sessions needed to reach the last step, subtract 1.

Teaching Procedures, step-by-step

**STEP 1**

**Aim:** The aim here is to teach emotion recognition and matching of emotions. This step will teach children the ability to recognise facial expressions, such as happy and sad, from photographs.

**Procedure:** *Matching across same static emotional images.* The robot displays facial expressions of target affective states and the child is asked to choose identical images (of the robot) from the set of images. Children are asked to identify four photographic robot facial expressions of emotions (placed in a randomized order). Photos of the robot are placed on table. Then the child is asked to match them.

If a child identifies the emotion correctly, positive feedback and encouragement is provided (e.g., “Well done! That is correct!”). Then the therapist activates the WoZ and the robot cheers the child. If there are errors, the child is given feedback by the therapist by correcting him/her. The training will move to the next step if the child can 75% of the emotions, in other words if the child can match 3 out of 4 emotions correctly in a trial.
STEP 2

**Aim**: Matching across different static emotional images. This step teaches the ability to identify correct expression from an emoticon.

**Procedure**: The robot displays facial expressions of target affective states and the child is asked to choose appropriate emoticons. The feedback is the same across steps.

STEP 3

**Aim**: Matching from dynamic ‘real’ emotional displays to static images.

**Procedure**: The robot displays an affective state using facial, bodily and vocal expressions and the child is asked to choose appropriate emoticons. The therapist would for example activate the WoZ and then the robot displays an emotion such as happiness. Then the therapist tells the child to look at Zeno and point to a card which shows how Zeno feels.

STEP 4

**Aim**: Identifying dynamic ‘real’ emotional displays and expressing that emotion.

**Procedure**: The robot displays an affective state using facial expressions and the child is asked to identify the affective state and to display the same affective state in the way s/he usually displays that affective state. Imitation will not be judged, and the training will continue even if the child does not imitate correctly.

Advanced Scenarios

STEP 5

**Aim**: Identifying dynamic ‘real’ emotional displays and expressing that emotion *in the same way*.

**Procedure**: The robot displays an affective state using facial expressions and the child is asked to identify the affective state and to display the same affective state in the way the robot displayed that affective state.

STEP 6

**Aim**: Understanding own/others’ emotional states (following Howlin et al., 1998), divided into 5 parts.

**Procedure**: The therapist tells a simple story in which *robot is the main character*, aiming for a certain emotion. The child is then asked to guess the emotion. For example a sad story could be the following: ‘The robot likes to play with children. You don’t want to play with him. How does robot feel when you don’t want to play with him?’ After reading the story, the therapist offers 2-4 emotion photographs of robot and asks how Zeno feels.

In the next advanced step *the child is the main actor*, being in one of the target affective states, and the child is asked to guess and display that affective state. In the last scenario another child is the main character. The feedback is the same as in the other steps; if a child identifies emotion, provide positive feedback and encouragement. If there are errors, the therapist will provide feedback.
2.4. Identification of ethical issues

DE-ENIGMA ethical issues are fully covered in the Data Management Plan (D6.2) and in the Ethical Clearances (D1.1). Ethical standards and guidelines of Horizon 2020 are being applied, regardless of the country where the research is being carried out.

Sensitivity to children’s needs

Extensive efforts will be made to ensure that the research is sensitive to children’s needs. In all of our research, we go to great lengths to ensure that our tasks and methods are enjoyable and developmentally appropriate for the age range of children being assessed. In all previous studies, children and young people have shown great enthusiasm in taking part and have enjoyed the one-to-one attention they receive during the testing session. We will ensure that the same care is taken with the assessment sessions in the current project.

Researchers will monitor closely participants’ well-being and breaks will be taken if deemed necessary by the researcher, or if they are requested by the participant. Children will be given as many breaks as necessary during the testing session. All children will be given positive reinforcement throughout the various parts of the sessions.

Risks to participants

This project and the nature of the tasks have been designed to be enjoyable and entertaining for children. Although existing studies have repeatedly shown that autistic children generally enjoy interacting with a human-like robot, probably because its behaviour is predictable and interpretable, it is nevertheless possible that some children might experience the robot’s behaviour as unexpected, which could cause some discomfort. Should children show an unusual level of unease during the interaction with the robot, our researchers will stop the session, switch into a fun, unrelated activity and consider whether testing can be resumed later on.
3. System Dependencies

This section presents the dependencies between the key technical/scientific WPs. It starts by presenting the results from a workshop about the partners' shared vision about the project and it will then summarize the inputs and outputs from each WP based on the DoA. The results show clearly the mixture of disciplines in the consortium – each of these with their own "requirements" for the final system.

3.1. Shared Vision Workshop Results

The project kick-off meeting included a workshop about DE-ENIGMA Shared Vision. The workshop's main goal was to explore the connections between the partners and their work for the DE-ENIGMA project. The workshop consisted of the following parts:

- Welcome and introduction to goal of workshop
- Warming-up with a hidden word assignment
- Discussion: explore the design space
- Brainstorm: explore the differences
- Information: Understand autism
- Scenario play of robot-therapy
- Brainstorm: find critical connections between partners
- Define shared vision and DE-ENIGMA slogan
- Closing the workshop

The results of the brainstorm parts of the workshop will be presented in the following subsections.

3.1.1. Brainstorm: explore the differences

In this part of the workshop, all partners were invited to write on post-it notes their main goals within DE-ENIGMA project. These goals were clustered under specific themes. Below the defined themes and pointed goals are presented.

Technical

- Reliable solution with fault tolerance
- An integrated working robot

Fundamental

- Understanding the extent to which robots can attune to the specific behavior of autistic children
- Understanding the similarities and differences in educational practice for autistic children in Serbia and UK
- Understanding the nature + extent of autistic + typical children responses to a robot

**Awareness**
- Make speech analysis robust in the wild, including typical speech
- Personalized emotion-sensitive interaction
- A working RTS system for action recognition human pose reconstruction
- Reliable human detection, pose estimation and action recognition localization
- Evaluation and development of child specific vs child independent models for affect recognition
- Large scale optimization of learning methods for human pose estimation
- Robot capable of dealing with unexpected events
- Robust ML and CV models for context-sensitive (personalized) affect-sensing from a) faces and b) audio-body cues

**Interaction**
- Interaction sessions where robot is experienced as a social actor
- Ongoing dialogue comfortable to child
- Ongoing interaction leading to improved social skills

**Evaluation**
- Evaluating the effectiveness of the final HRI system relative to HHI
- Improvement comparable to traditional therapy
- Help children with autism overcome barriers to understand better social interaction
- Enhance support for children with autism in a fun/entertaining way
- Find easily transferable tools to cater for the variety of needs across EU
- A tool that we can exploit in the end

**Ethical/outreach**
- Sensitization of parents and professionals about the use of assistive technology in therapy
- Analysis of cultural differences for affect recognition
- Understanding the views + perspectives of parents, therapists + children of using robots as educational tools

### 3.1.2. Connections between partners

One person from each project partner stated their main contribution for the DE-ENIGMA project. Further they explained what type of input they expect from other partners and what output they will provide to other partners. This information is collected in the Connection
scheme (see Figure 3). Figure 4 systematizes the notes made on the whiteboard during the brainstorm sessions which served as input for the Connection scheme.

**IDM**

**Main contribution:**
- Fully integrated hardware and software system

**Inputs from other partners:**
- Short term: defined scenario, specification of components (sensors etc.), initial hardware onboard and eventually off board (note: keep it as simple as possible!)
- Long term: perception node, vision node, speech recognition node, behaviour node.

**Outputs to other partners:**
- Operating system to work with. Preferably Linux, using ROS to operate the robot (might change if for example cloud services will be used to transfer data).

**IMAR**

**Main contribution:**
- People recognition (body posture and gestures) in challenging set-ups

**Inputs from other partners:**
- Understand the set-up with kids and therapist, how difficult they are. How sensors will be placed to best view the scene.
- Data to annotate, recognize some behaviours
- Similar recording setup in UK and Serbia
- Similar rooms (to create similar datasets)

**Outputs to other partners:**
- Synchronize needs with needs of other partners, such as face annotation data and sound
- Input to the user model
- Input to DE-ENIGMA DATABASE
- Interpretation of data

**ICL**

**Main contribution:**
- Faces and facial expressions of the child and interpretation of these expressions

**Inputs from other partners:**
- Similar recording setup in UK and Serbia

**Outputs to other partners:**
- Input to the user model
• Input to DE-ENIGMA DATABASE
• Interpretation of data

UT
Main contribution:
• System deciding on what action to take
• Define scenarios for final system
• Develop the human-robot interaction
Inputs from other partners:
• User model (Interpreted data on facial expressions, body posture and gestures and sound)
• Understanding the course of therapy sessions and how a therapist would react to specific situations
• Learning about autistic children and how to help them
• Understanding sensing and actuation capabilities of the robot
Outputs to other partners:
• Repertoire of behaviour for the robot (if-then-else)
• Strategy for therapy session with robot
• Encode behaviour for system (e.g., state machine)
• Recommendation to “sensing” partners to put focus on specific things

UP
Main contribution:
• Provide audio analysis (of emotion, laughter)
• Recognize audio from other people (parents/therapist)
Inputs from other partners:
• Sound data from robot
• Extra microphone in room or on child or therapist (to learn how to improve sound data from the robot)
• Learned models for context transfer
• Specification of system
• Similar recording setup in UK and Serbia
Outputs to other partners:
• Input to the user model
• Input to DE-ENIGMA DATABASE
• Interpretation of data
• Feed to scenario what is interesting for them to know

SSA and UCL
Main contribution:
• Access to autistic children and therapy sessions
• Inputs from other partners:
• Questions about what is technically wise possible
• Requirements for recordings/sensing partners
• Information on types of data that will be released (for the ethical clearance)

Outputs to other partners:
• Access to children
• Data and settings are comparable

AE
Main contribution:
• Project dissemination

Inputs from other partners:
• Active disseminating via own channels
• Published articles
• Respond to information request

Outputs to other partners:
• Newsletter
• Dissemination of the project

Final DE-ENIGMA slogan

DE-ENIGMA: playfully empowering autistic children
Figure 3. Partners’ connection scheme.
Figure 4. Organization of contributions by thematic areas.
3.2. Main Modules

DE-ENIGMA scientific and technological developments are organized as depicted in Figure 5. The following subsections describe the interfaces between the main modules.

![Figure 5. Overall relation between components.]

3.2.1. DE-ENIGMA Database (WP1)

DE-ENIGMA uses generic detectors of facial landmarks and gestures, bodily posture and gestures, vocal and verbal cues, as well as generic audio-visual (continuously valued) estimators of subjects’ target affective states (happiness, anger, sadness, surprise), affective dimensions (valence and arousal), and mental states (interest and stress), which are all trained using annotated examples stored in the DE-ENIGMA database. Also, the examples of rapport establishment between children with ASC and the therapist/robot, used by the rapport recognition tools are stored in the DE-ENIGMA database. (Note: As said above, the generic detectors are personalised in other DE-ENIGMA modules by means of context-sensitive and context transfer learning frameworks.)

The DE-ENIGMA database will contain fully-synchronised and annotated audio, 2D and 3D recordings of interactions between autistic children and (a) the robot, (b) the researcher and (c) their parents made in structured teaching settings. A total of 128 children on the autism spectrum will be recorded, half (n = 64) from London and South East of UK, and the other half from Serbia. In each culture group, 32 children will be aged between 5 and 8 years and the other 32 between 9 and 12 years. The children from the two cultures will take part in identical experiment settings. Namely, for each culture, half of the children will be involved in robot-led teaching and the other half will be involved in researcher / clinician-led teaching. During the experiment, children within each age group will be randomly assigned to either robot-led or researcher / clinician-led teaching intervention, which will be implemented across multiple short sessions (10-15 minutes long) every 1-2 days for a maximum period of 3 weeks. We will follow Howlin et al.’s (1998) approach to teaching perception, expression,
understanding, and social imagination related to four affective states: surprise, happiness, anger and sadness.

In the data collection experiment, all robot-led sessions will be facilitated through a “Wizard of Oz” (WoZ) setup. Namely, the robot will be controlled directly by the researcher / clinician using a small keypad hidden from the child’s view. Nevertheless, the robot will also perform a set of idle animations autonomously, such as eye-blinks, head-turns, and minor hand movements, to achieve a more “life-like” appearance.

All teaching sessions will be recorded in 3 modalities: audio, 2D video, and 3D video. For each modality, the following devices are used:

1. Audio: 4 professional omnidirectional microphones will be used as the main data sources. Among these microphones, two are mounted close to the child and the researcher / clinician respectively. Another one is mounted on the ceiling of the room directly above the experiment setup. And the last one is a wireless microphone carried either by the child or the researcher / clinician (in case if the child is unwilling or unsuitable to carry the microphone). In addition to these professional microphones, we will also use the 2D and 3D video cameras’ built-in consumer-grade microphones to make extra audio recordings. Each 2D camera has 2 built-in microphones (except the one mounted on the robot’s chest that only has 1), and the 3D camera (Microsoft Kinect) has 4 built-in microphones. Therefore, for each session, a total of 18 (in researcher / clinician-led sessions) or 19 (in robot-led sessions) distinct audio recordings will be made.

2. 2D Video: 5 (in researcher / clinician-led sessions) or 6 (in robot-led sessions) 720p HD webcams will be used to make video recordings at approximately 30 frames per second. The placement of these cameras is as follows: 2 cameras are mounted at the opposite corners of the room to record from overview perspectives; 3 cameras are placed close to the researcher / clinician and the child to capture their facial expressions (1 facing the researcher / clinician, 1 facing the child, and 1 facing both); and, in robot-led sessions, 1 camera will be mounted on the robot to capture the scene from the robot’s perspective. In addition, the 2D video captured by the 3D camera will also be recorded. These add up to 6 (in researcher / clinician-led sessions) or 7 (in robot-led sessions) 2D video recordings per session.

3. 3D Video: We will use one Microsoft Kinect to record 2D and 3D video and sound data. We will record the monocular image, the registered depth field of the scene, and the 4-channel sound, at a sample rate of approximately 30 Hz.

The sensor placement is further illustrated in Figure 6. Note that the two overview webcams at the corners and the microphone on the ceiling are not visible in the picture.

All recorded data streams will be time-stamped and synchronised. Specifically, the internal clock of all data capturing machines will be synchronised to universal time coordinated (UTC) using network time protocol (NTP). These clocks will then serve as the reference clock to time-stamp all recorded data on either per-frame basis (for 2D and 3D video data) or per-buffer basis (for audio data).

The DE-ENIGMA database will also include annotations of the recordings in terms of facial landmarks and gestures, body postures and gestures, vocal and verbal cues, continuously valued emotion dimensions, and rapport behaviours. The data will be annotated in an
iterative fashion, starting with a sufficient amount of examples to be annotated in a semi-automated manner and to be used to train the algorithms in WP2-WP4, and ending with a large database of annotated facial and bodily behaviour recorded in the wild.

Figure 6. Sensor placement in the experiment setup for robot-led sessions. *This setup is also the starting point for the final robot-assisted therapy setup to facilitate the interactive game.

**Input:**
Raw audio, 2D and 3D video data, and annotations in terms of facial landmarks and gestures, body postures and gestures, vocal and verbal cues, continuously valued emotion dimensions, and rapport behaviours.

**Output:**
The DE-ENIGMA database that will be used to train the algorithms in WP2-WP4. The database will also made available to academic researchers, thus to become a benchmark multilingual dataset of annotated atypical facial, bodily, vocal and verbal interactive behaviour recordings made in naturalistic settings.
3.2.2. DE-ENIGMA Perception Module (WP2)

This module processes audio-visual input from the children with ASC and the therapist/robot, to obtain low and mid-level acoustic and visual features by robust, context sensitive, and real-time machine analysis of facial, bodily, and vocal cues. These enable (in WP3) reasoning about affective states/dimensions and mental states, and rapport, as shown by ASC children in real-life conditions (including semi-dark and noisy rooms, with dynamic change of room impulse response, lightning conditions, and distance to sensors, and behaviours that are atypical and abruptly changing as characteristic for ASC children) and in both languages English and Serbian. This is based, in part, on existing technology, such as models for automatic speech recognition (ASR) (e.g., Weninger et al., 2014b), readily available face and facial landmark detectors/ trackers (e.g. Tzimiropoulos et al., 2014, Asthana et al, 2015), and models for automated analysis of 2D/3D human kinematics (e.g., Ionescu et al. 2014a, Bo & Sminchisescu 2010), and, in part, on novel models devised in the DE-ENIGMA project with the aim of addressing the task-specific challenges typically encountered in real-life conditions, as described above.

**Input:**
Robust, and context-sensitive features are extracted from facial, bodily, and vocal cues in unconstrained recording conditions including semi-dark and noisy rooms, with dynamic change of room impulse response, lightning conditions, and distance to sensors, and behaviours that are atypical and abruptly changing as characteristic for ASC children. Input signals are acquired directly from the modality-related sensors, such as microphones, 2D and 3D cameras.

**Output:**
WP2 will deliver speech, acoustic, and visual (facial and bodily) features that enable (in WP3) recognition of rapport and affective states including valence, arousal, interest and stress as shown by ASC children in real-life conditions and in both languages English and Serbian.

3.2.3. DE-ENIGMA Reasoning Module (WP3)

Based on the acoustic and visual features extracted in WP2, this module achieves fully context-sensitive and robust continuous intensity estimation of affective states (happiness, anger, sadness, surprise), affective dimensions (valence and arousal), and mental states (interest and stress), and recognition of rapport. These are used further in the human-robot interaction module in WP4. The target models are based upon existing technology, such as audio-visual affect recognizer developed during the FP7 SEMAINE project (Schroeder et al., 2012), the state-of-the-art audio-visual affect predictors such as those proposed by Nicolaou et al. (2011, 2013, 2014a, 2014b), and context-sensitive framework for affect intensity estimation (e.g. Rudovic et al., 2014), and extended to account for more complex contexts (across space and time, and including meta-data about recorded subjects) and to achieve context-transfer in these models (e.g., within and between cultures). DE-ENIGMA project also devises in this WP novel models for a fully context-sensitive and cross-modal (including facial, bodily and verbal/vocal modalities) rapport recognition, based on extensions of existing models for unsupervised segmentation and classification of sequential data (Liwicki et al., 2012, Walecki et al., 2015).
Input:
Speech, acoustic and visual (facial and bodily) features produced by the WP2 module.

Output:
WP3 module will output the estimation of the child’s affect (happiness, anger, sadness, surprise), continuously-valued valence and arousal level, level of interest, and stress level. The module will also recognise the display of rapport between the child and the robot. All the recognition results will be accompanied by meaningful confidence measures.

3.2.4. DE-ENIGMA Human-Robot Interaction Module (WP4)
This WP concerns understanding the context of use for the DE-ENIGMA system and the design and development of the DE-ENIGMA interactive game to engage autistic children in developing socio-emotional skills. This WP builds upon the previous WPs: WP1’s data collection by analysis of the recordings from T1.3; WP2’s perception by developing a constant feedback-loop between perception and action possibilities of the robot in the game; and WP3’s high-level reasoning by offering concrete sets of behaviours for execution and planning. The outcomes of WP4 will to a great extent inform WP5 which involves the integration and evaluation of the integrated DE-ENIGMA solution.

Input:
Decision on what behaviour the robot should display at a certain point in time will be based on the input from the WP3 reasoning model, which include affective states, affective dimensions, mental states, and recognition of rapport. Additionally, input from WP2 will be used to decide whether the behaviour selected by the dialogue model can be successfully performed by the robot.

Output:
WP4 will deliver a set of verbal and non-verbal behaviours, to accommodate both the children from the United Kingdom and Serbia, and decide which behaviours to execute at a certain point in time. The selected behaviour will be translated to control primitives, which are used in WP5 to translate the control primitives to actual movements and speech.

3.2.5. DE-ENIGMA Integration & Evaluation (WP5)
WP5 uses the output from WP4 to perform integration of the DE-ENIGMA modules into the robot and to address the key objectives: (1) to assess the impact of Human-Robot Interaction (HRI) compared to Human-Human Interaction (HHI) on autistic children’s socio-emotional (facial, bodily, vocal) skills, (2) to determine the extent to which culture (UK vs. Serbian) moderates the impact of HRI vs. HHI; and (3) to test, using a randomised controlled trial, whether ASC children’s learning is better attained during HRI (based on facial, bodily, and vocal modalities) with (a) explicit learning (direct instruction) vs. (b) implicit learning through games. It also targets to determine the feasibility of implementing Human-Robot Interaction (HRI) in a school-based context (including acceptability to children, parents and teachers).
Input:
The WP5 module receives behaviours (encapsulating sets of primitives) from the WP4 module, which are then translated into real robot actuations of movement and speech.

Output:
WP5 delivers the framework which supports the communication between the different hardware and software modules. It will also convert WP4’s behaviours into real robot actions which will be used to assess the system performance and evaluate its results.
4. System Technical Specifications

DE-ENIGMA hardware system will be centred in an off-the-shelf robot called Zeno R25, produced by Robokind and depicted in Figure 8. Moreover, the system will comprise external sensors and processing units. The general technical specifications are now presented.

4.1. System Overview

DE-ENIGMA system is based in a hardware layer composed by the robot and some external sensors. Over this hardware layer, the system will be organized in three software layers with different levels of abstraction. Figure 7 summarizes the system organization.

![System Overview Diagram](image)

**Figure 7.** DE-ENIGMA system overview.

4.2. Hardware

The hardware layer will be composed by the robot platform, 2D and 3D cameras, microphones and computing resources.

4.2.1. Robot Platform

The project is making use of an off-the-shelf robot called Zeno R25, produced by Robokind and depicted in Figure 8. Zeno R25 is a humanoid robot with a total of 21 degrees of freedom (DOF). The head comprises 7 DOF which gives it a great capacity of expression. Its sensing ability includes an array of 8 microphones and a 5MP camera. The robot internal processing is based on an OMAP 4460 dual core 1.5GHz processor. The full robot specifications are listed in
Table 2. The accessibility and quality of the built-in sensors will be investigated in the coming months.

Figure 8. Robokind’s Zeno R25 robot (aka Milo).
4.2.2. External Sensors

The system will include some external sensors to complement the sensing capabilities of the robot, namely:

- A RGB-D camera
- HD web cameras
- Microphones

4.3. Perception

The lower level processes of perception need to run continuously. Features are extracted at a high-rate from facial, bodily, and vocal cues. Machine analysis of vocal cues will rely on Zeno’s built-in acoustic sensors, and/or additional microphone placed on the final robot prototype. These processes might be running on independent machines with different OS.
4.4. Reasoning

Understanding the context in which the robot interacts is a key issue in human-robot interaction. The project proposes a user-centred situation assessment system which will grow in terms of capabilities during the time-span of the project. The reasoning layer of software works as an intermediary between low-level continuously running perception processes and decision making high-level layer. More precisely it will classify perceived visual and vocal features into emotional states.

4.5. Human Robot Interaction

This layer decides on the different actions to be executed. It encapsulates a state machine which will evolve according with the context analysis provided by the reasoning processes and the defined interaction scenarios. This module will be sending actuation commands directly to the robot controller.

4.6. Integration Framework

Software integration within the DE-ENIGMA project will be accomplished with ROS\(^1\). ROS provides the user with a middleware which can be used to connect different machines. The main characteristics of the middleware are listed below.

- Great versatility of types of communication:
  - *data-centric* communication, in which publisher and subscribers of different data topics can be created.
  - *service-centric* communication, in which servers can provide clients with services.
  - *action-oriented services*\(^2\).
- The transmission is carried out only through TCP when processes from different computers are connected.
- Acts as a nameserver, which is used to generate TCP connections. Once the different modules are connected, the communication does not need to pass through the nameserver node.
- Multiplatform. It can be used in all platforms where ROS can be installed.
- Provides a distributed parameter system for configuring the ROS nodes.

The high-level HRI decision layer machine will be running over ROS. It will embed the ROS Master which will provide naming and registration services to the rest of the nodes in the ROS system. It tracks publishers and subscribers to topics as well as services. The role of the Master is to enable individual ROS nodes to locate one another. Once these nodes have located each other they communicate with each other peer-to-peer.

ROS is developed and tested for Linux-based platforms, namely, Ubuntu. Although the execution of the Master ROS node in Windows platforms is not yet fully possible, a ROS node

\(^1\) [http://wiki.ros.org/](http://wiki.ros.org/)
\(^2\) [http://wiki.ros.org/actionlib](http://wiki.ros.org/actionlib)
can be written for any system that supports TCP/IP communication. Figure 9 depicts a preliminary architecture for DE-ENIGMA system. To ease the process of development, we are considering a PC for each partner. At a later stage, this will be optimized to reduce as much as possible the number of processing machines.

Figure 9. Preliminary system architecture.

4.7. Data Collection System Architecture

DE-ENIGMA comprises a set of data collection sessions. During the months of June and July 2016, partner SSA hosted the first data collection meeting. The results are presented in Table 1. This meeting served as a first approach to the hardware that will be used in DE-ENIGMA system. The architecture of the data collection setup is depicted in Figure 10. Figure 6 shows the sensor placement for these experiments. More details about the data collection can be found in D6.2.
Figure 10. Data collection setup.
5. Methodology of Development

DE-ENIGMA will be developed in an iterative way. The system will be displaying incrementally improving “emotional intelligence” in terms of the abilities to interpret and react appropriately to stress, affect, interest and rapport shown by autistic children. More specifically:

- The 1st prototype of DE-ENIGMA robot (M18) will be a Wizard-of-Oz HRI application for direct-instruction learning of socio-emotional skills and will be tested for general acceptance by children with autism.

- The 2nd prototype of DE-ENIGMA robot (M30) will be able to interpret and react appropriately to stress, affect, and interest shown by autistic children involved in the direct-instruction learning of socio-emotional skills.

- The 3rd prototype of DE-ENIGMA robot (M36) will be able to interpret and react appropriately to all stress, affect, interest and rapport shown by autistic children involved in the implicit learning approach of socio-emotional skills and will be thoroughly tested through a range of trials.

We will follow a step-by-step approach for addressing integration:

- Feature boxing grouping of features (i.e. a set of features that can be grouped together for a given purpose).

- Concept matching making sure that concepts between feature groups match and relate well with each other.

- Information compatibility checking message structure and semantics, translation of information

- Data timing matching frequency, latency, data ordering

- Resource allocation check which and how much resources are needed for all the groups of features that need to run. Resources can be computing resources (CPU time, memory, network access) and/or domain specific resources (e.g. power, control of actuators, tasks allocated to the robot)
References


