DR 7.2: Scenario-based evaluation for multiple asynchronous sorties to assess a large-scale dynamic disaster area

Joachim de Greeff*, Tina Mioch†, Norbert Pahlke§, Robert Maul§ and the TRADR consortium

*TUD, Delft, the Netherlands
†TNO, Soesterberg, the Netherlands
§FDDo, Dortmund, Germany
⟨J.deGreeff@tudelft.nl⟩

Project, project Id:	EU FP7 TRADR / ICT-60963
Project start date:	Nov 1 2013 (50 months)
Due date of deliverable:	December 2015
Actual submission date:	February 2016
Lead partner:	TUD
Revision:	final
Dissemination level:	PU

In this deliverable we describe the scenario-based evaluations as carried out in Year 2 of TRADR. We discuss the creation of the use cases and the development of an evaluation methodology, along with reports on the organization and execution of the Year 2 exercises. In addition, feedback and observations from end-users are included, as well as an exploratory study on the ethics of robot-assisted search and rescue.

Executive Summary

This report documents milestone MS7.2: scenario-based evaluation year 2 for WP7. This milestone consists of task T7.2: to perform a *scenario-based evaluation for multiple asynchronous sorties to assess a large-scale dynamic disaster area.* Building upon the foundations created in Year 1, we further refined the scenario and use cases, which were subsequently used during three end-user evaluation exercises held in May, September (Dortmund, Germany) and December (Delft, the Netherlands). In addition, we started working towards a more structured evaluation methodology and have explored the ethical landscape of Robot-Assisted Search and Rescue.

Role of Scenario-based evaluation in TRADR

The goal of this workpackage is to perform a scenario-based analysis for multiple asynchronous sorties to assess a large-scale dynamic disaster area. Towards this end, the user-needs, the scenario and the use cases were refined for Year 2. In this year we have worked towards a more fundamental manner in which the scenario and use cases serve evaluation in the project. This methodology was applied during two field-experiments with the TRADR end-users. The TRADR Joint Exercise (T-JEx) serves as an exploratory field study during which new components, modules and use cases are tested, while the TRADR Evaluation (T-Eval) is a more formal evaluation with less emphasis on exploration and more on maturation of the system. A high level of end-user involvement remains important in TRADR; as such there is tight collaboration with the end-users in the creation of the scenarios and use cases to ensure realism and fidelity.

Contribution to the TRADR scenarios and prototypes

This workpackage is responsible for defining and creating the scenario and use cases. The overall scenario has not changed since Year 1, but the use cases have been further refined to include key elements for Year 2, such as: multiple synchronous or asynchronous sorties, larger-scale environment, dynamic events and occurring network loss. This was done by collecting input from all other workpackages, as well as through ongoing dialogue with the TRADR end-users.

Persistence

In the context of WP7 persistence means situation awareness reuse and further elaboration across teams in continued missions. The scenario is formulated in such a way that an adequate response requires accumulation of data and situation awareness through multiple sorties over potentially multiple days. Thus, by deploying such a scenario for the evaluation, the TRADR system is assessed on its ability to perform in a persistent manner. The requirement of persistence is reflected in the use cases. More specifically, Year 2 included use cases that required the system to deal with information gathered during previous sorties, e.g. "GUC 4: UGV[x] go to location X in (semi)autonomous mode" (which requires the creation of a map) and "EUC #13: UGV#1 and UGV#2 enter the site (semi)autonomously, using knowledge gathered on day 1". Also "GUC 10: TRADR team wraps up a mission" and "GUC 11: TRADR team continues a mission" subscribe to a notion of persistence over time. That is, a TRADR team continuing a mission (initiating a new sortie) entails access and effective presentation of information gathered during previous sorties. These use cases are described in more detail in section 1.3.1 and Annex [8] (Annex Overview 2.1).

1 Tasks, objectives, results

1.1 Planned work

For Year 2, the following work was planned:

- User needs analysis: the second year exceeds the first year in that the disaster area is dynamic; this can have implications on the end-users needs.
- Refinement of the socio-technical design rationale.
- Planning of scenario-based evaluation: based on the user needs analysis, the scenario needs to be defined to incorporate dynamic events.
- Defining methods and metrics: the scenario-based evaluation should incorporate dynamic events. Dynamic events are localized, such as contained fires, limited outpour of liquids, falling-over of small structures or objects (e.g. barrels, small containers). Sorties can be synchronous or asynchronous, and can involve one or more robots (1 UAV, up to 2 UGVs). The methods and metrics need to take into account the dynamic events that may occur.
- Assess with end-users: the assessment with the end-users needs to take into account the increased spatial complexity.

1.2 Addressing reviewers' comments

During last year's review, the following comment was made by the reviewers regarding this workpackage:

1. **An explicit process for the UI design and development and its evaluation should be included in a parallel and complementary way.** In Year 2 we have addressed TDS development in a more fundamental manner, to catch up with delays from Year 1. This process is described in more detail in TRADR Deliverable DR3.2, sections 1.3.1 and 1.3.2 (annex 2.1 and 2.2) [10]. Intrinsic to this development is also the evaluation with end-users, to test the effectiveness of the UI and collect feedback for further refinement and development. In Year 2 there have been three exercises during which end-users tested the UI, these were T-JEx, T-Eval and ITEX. Feedback collected during these exercises has subsequently been used as input to improve the TDS; thus development and evaluation have been more complementary compared to last year. Details of these are described in Section 1.3.3. By embedding the design of the TDS in an incremental develop-and-evaluation cycle, we believe we are addressing the reviewers' concerns.

1.3 Actual work performed

In this section we describe the actual work that was performed, and how it feeds into the overall objective for WP7 in Year 2. We first describe the use cases for Year 2 in Section 1.3.1, followed by a description of the evaluation framework that we started to develop in Section 1.3.2. We then describe how these aspects were applied during field experiments with the end-users (T-JEx and T-Eval) in Section 1.3.3, after which feedback from the end-users is provided in Section 1.3.4, along with an account from the end-user perspective (Section 1.3.5). Finally, we describe a study on the ethics of robot-assisted Search and Rescue is presented in Section 1.3.6.

1.3.1 Use cases Year 2 and TRADR unit

The creation of the Year 2 scenarios is part of the situated Cognitive Engineering (sCE) approach [26], which defines an iterative human-centred development process consisting of foundation, specification and evaluation (Fig. 1). In Year 2 we build upon the foundation as formulated in the socio-technical design rationale during Year 1 [34]. For Year 2, the claims and requirements remain largely the same as Year 1. For some subparts of the TRADR system (activity recognition and decision support) additional claims and requirements have been formulated, a description of which can be found in Annex 2.4 and 2.5 of TRADR Deliverable DR5.2 [20].

The scenario and use cases for Year 2 build upon those of Year 1. The scenario describes an emergency situation which provides a backdrop of the TRADR mission, for instance "fire and explosions reported at a chemical factory, with possible emission of gas and missing workers". The use cases describe in more detail specific functionalities that the system is able to perform; e.g. "the UGV explores the area and provides pictures from above", or "the UGV takes a sample using its arm". These scenario and use cases are just an example, a full description of the Year 2 scenario and use cases can be found in Annex [8] (Annex Overview 2.1).

In addition, in Year 2 we have stared with making a distinction between Generic Use Cases (GUC) which describe a certain functionality at an abstract level, and Evaluation Use Cases, which are tailored towards the particular Year 2 scenario, and use the GUCs as building blocks. The use cases have been formulated through ongoing discussion between TRADR partners, taking into account both what is technically achievable (and indeed, worth pursuing from a scientific viewpoint) and what is realistic given the disaster response domain. Particularly for the latter, input from the end-users has been invaluable.

TRADR unit We consider the TRADR unit to be a specialized unit, embedded within a larger first-response organization. Compared to Year 1,

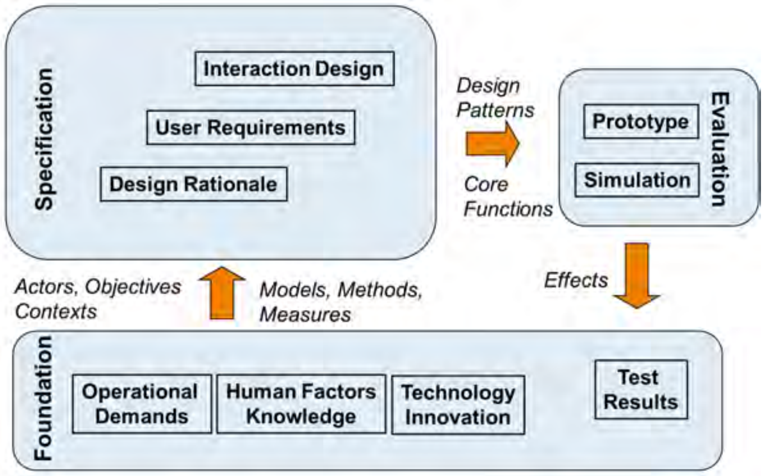

Figure 1: The situated Cognitive Engineering method with the different phases: foundation, specification and evaluation.

the TRADR team has expanded, as now two UGVs are used. The TRADR unit consists of the following:

- Four human team members in remote command post (TeamLeader (TL), UGV#1 operator, UGV#2 operator, UAV operator)
- One human Infield rescuer (IR)
- Two UGVs
- One UAV

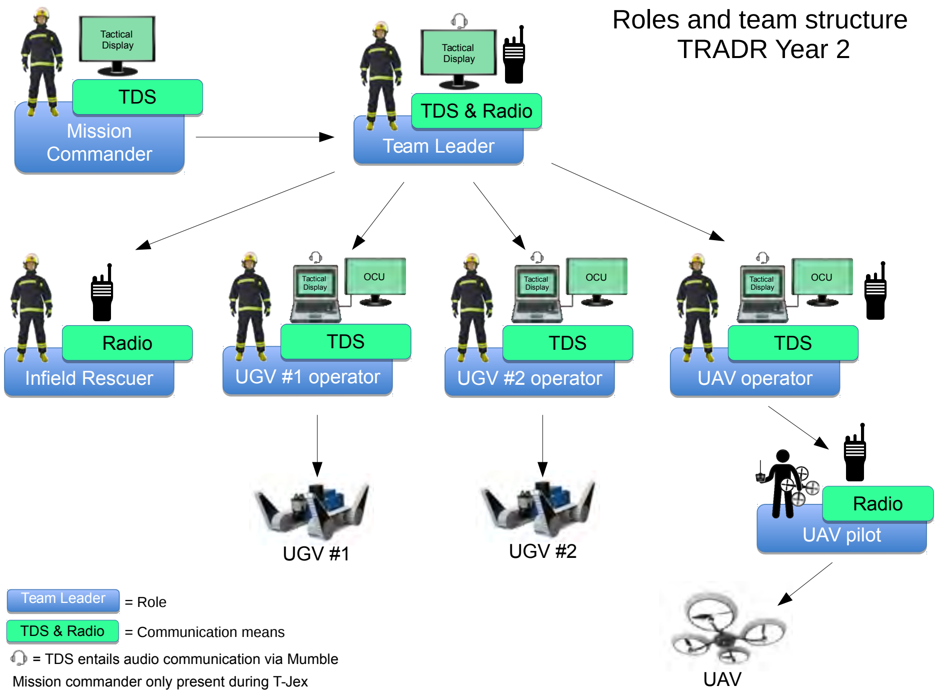

Figure 2: TRADR Year 2 roles and team structure.

Figure 2 shows the different roles and the team structure. In practice, we also had a human UAV pilot who was directly controlling the UAV, plus human safety officers for each robot. These people however, were not considered as part of the TRADR unit, as their role eventually would not be needed as technology advances.

Throughout the different exercises, experiences were gained regarding the constitution of the TRADR team. The end-users had indicated that the inclusion of a Mission Commander (MC) is more realistic, and in line with existing command structures. In a real mission, the MC would engage in high-level command and overview of multiple teams, the TRADR team being one of these. However, during the Year 2 exercises there were no other teams, and thus the MC's task only consisted of initiating the mission by briefing the TL, and keeping track of what was going on via TDS. So, for T-Eval it was decided that this role did not contribute much, and the MC was taken out of the scenario. The evaluation of this change constitutes lessons about where to delimit the scope for the experiments. This is an ongoing process, and for Year 3 the scope and team structure will be revisited, taking the experiences from Year 2 into account.

In a similar vein, the particular device that is used by the TL has also been the subject of exploration. Inherited from NIFTi, the TL was initially equipped with a big stationary screen in Year 1. However, the end-users indicated that the TL should actually be mobile. Thus, this was experimented with during T-JEx Year 2, where the TL used a portable tablet. However, this proved not satisfactory in practice (potentially due to preliminary nature of the interface), and hence the TL switched to a stationary laptop again during T-Eval. By experimenting with different setups, we have been collecting experiences in order to arrive at the most appropriate solution. Again this is an ongoing process to be continued in Year 3.

1.3.2 Evaluation methodology

Building upon the framework for scenario based evaluation that was developed through Year 1, in Year 2 we have continued the cycle of generation, evaluation, and refinement as per the sCE method. Particularly, regarding the evaluation, we have worked on a more fundamental, controlled and holistic methodology to evaluate the system at different levels of granularity. The aim has been to develop a framework for evaluation, which addresses both individual and (low-level) functional components of a large socio-technical system (such as TRADR), as well as its ability to support higher-level mission objectives as dictated by the scenarios. The methodology was first applied during T-Eval. Results appear to indicate the approach is useful. However, further refinement, particularly regarding which functional components are measured and how they are weighted in an overall score is needed. This will be an ongoing effort towards Year 3. Figure 3 illustrates the breakdown of the system into three different levels (system foundation, functional components and mission objectives), each with their own method of evaluation. For a detailed description of the methodology, see Annex [9] (Annex Overview 2.2).

1.3.3 Scenario-based evaluation exercises: T-JEx, T-Eval and ITEX

In Year 2, three events were organized during which scenario based evaluations took place. These were: the TRADR Joint Exercise (T-JEx) Year 2 in May 2015, the TRADR Evaluation (T-Eval) Year 2 in September 2015 and the Integration, TDS and end-user EXperience (ITEX) workshop in December 2015. T-JEx and T-Eval were organized in collaboration with the FDDo end-users, and took place at an old industrial complex of a former blast furnace “Phoenix West” in Dortmund, Germany. See Figure 4 and Figure 5 for an impression of the environment and a schematic overview respectively. The purpose for T-JEx and T-Eval was the overall evaluation of the TRADR system for Year 2, while ITEX was specifically aimed at the end-user experience through the newly developed TDS (TRADR Deliverable DR3.2 [10]). As such ITEX was not a field experiment, but instead

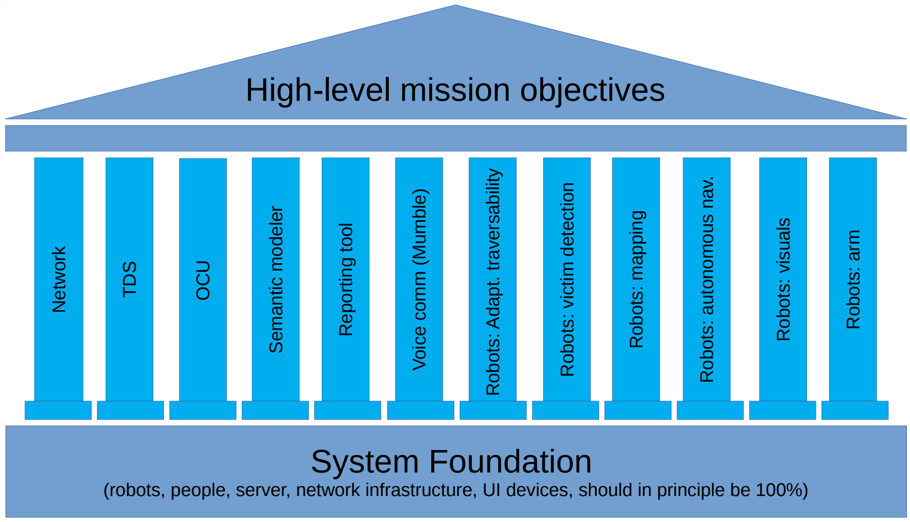

Figure 3: Overview of the evaluation structure as a pillar diagram.

used a simulated Search and Rescue environment. Even though this was less realistic than an actual field experiment, it was much easier to organize in terms of logistics, while it still allowed us to evaluate the new TDS and related TRADR infrastructure (e.g. databases).

The full report on the exercises is attached as Annex [11] (Annex Overview 2.3). Because the end-user feedback is fundamental to the research and development cycle, their feedback resulting from T-Eval is addressed explicitly in the next section (Section 1.3.4).

Figure 4: Impression of the "Phoenix West" former blast furnace, at which the T-JEx and T-Eval exercises took place.

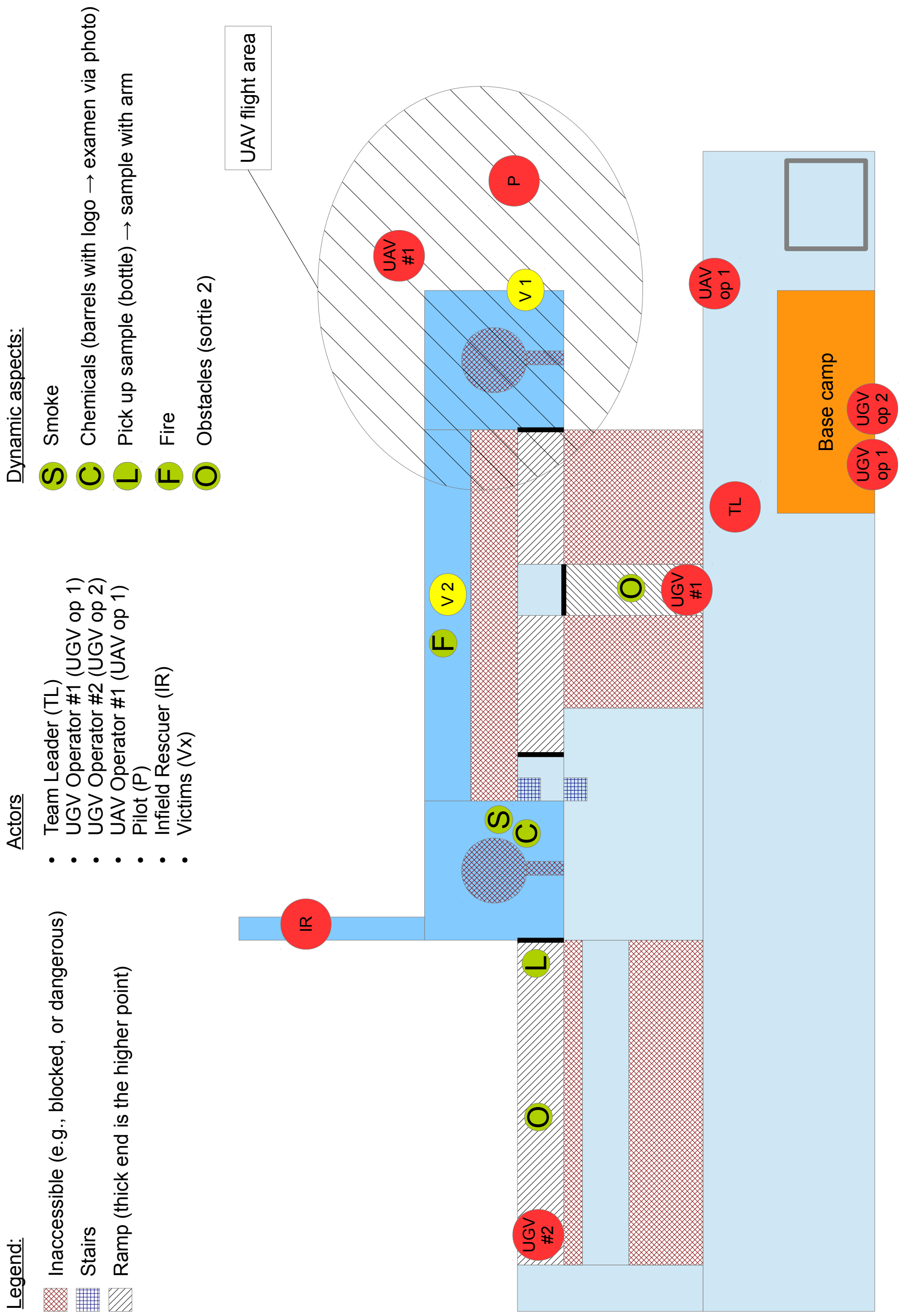

Figure 5: Schematic overview of the location at Phoenix West, Dortmund

1.3.4 Summary end-user discussion T-Eval

In this section, the discussions with the firefighters who participated during T-Eval are reported. We describe the general comments that were collected over the course of the exercises, as these are illustrative of the end-users experience during T-Eval. In addition, a large collection of comments specifically related to particular functionality of the TRADR systems were also obtained. The full list of these can be found in [25] (Annex Overview 2.4).

General comments of firefighters

- Very interesting day. It was very interesting to experience to get current pictures through the robots and to gain insight into the situation in this way.
- It is important to know the limits of the different systems. For example, the wind was too much, so that the zoom for the UAV was not possible; it was not possible to get details.
- The system is quite intuitive, it was possible to learn the basics in one morning, which is quite good.
- One day, the team leader forgot that he had an infield rescuer, so he got no task (just one at the end). The team leader later said that probably, he first should have sent a team into the building to check whether the robots needed to be sent.
- On the second day, it was not clear that the mission was part of a greater mission, and that the data in the database was still relevant and gave the situation as seen by the prior team. This was a misunderstanding during the briefing of the TL.
- Several participants said that they could image using the robots as a team leader/mission commander. They also agreed that they would not want to be the operator.
- It was a fun day.
- It is very important that the system stays easily operable.
- Very fascinating and exciting experience.
- Ground and air vehicles are at a different development level.
- Ground robots are not (yet) usable for a fire rescue mission.
- If all functionality that is currently implemented would actually work, and would be presented well, the system would be very useful already.

1.3.5 End-user reports on T-Jex/T-Eval

Additionally, the end-users created a report which describes the Year 2 exercises from their perspective. As is to be expected, there exists some overlap in the aspects they noted and the end-user comments that were collected as part of the exercise data-gathering. Nevertheless, their report is included because it provides a complementary description of the Year 2 exercises. The report be found in Annex [18] (Annex Overview 2.5).

1.3.6 Exploring the Ethical Landscape of Robot-Assisted Search and Rescue

Advancements in AI and robotics has fueled a growing realization that the ethics of human-robot interaction need to be addressed. For some domains – e.g. military, car industry, healthcare and education – ethical concerns regarding the application of robots have received a lot of attention [22]. However, in the field of Search and Rescue (SAR) it appears that ethics related to the use of robots has not so much attention. To address this gap, we explored the ethical robot-assisted SAR landscape by identifying and analyzing humans values (e.g. trust, autonomy and privacy) and value tensions. Value tensions refer to situations in which technology supports one value while at the same time hinders another; as such they are indicators of potential ethical dilemmas. Our approach is inspired on the Value Sensitive Design (VSD) methodology, which accounts for human values throughout the design process [15]. We conduct a series of three Value Assessment workshops with SAR workers – in this case firefighters – in which we make use of VSD methods to assess and analyze the stakeholders and their values in the SAR field. Using the workshop results, literature and experiences in the TRADR project, we identify key ethical concerns and dilemmas for the robot-assisted SAR field. The full study can be found in Annex [19] (Annex Overview 2.6).

1.4 Relation to the state-of-the-art

1.4.1 Scenario and use cases

The TRADR scenario and use cases are created in close collaboration with end-users. As such, the are realistic and in line with actual user requirements. Compared to existing robotics competitions, such as e.g. the Robocup Rescue League[1], the Darpa Robotics Challenge[2] and the euRathlon robotics competition[3], the TRADR scenarios tend to be less pre-determined in terms

[1] `http://www.robocuprescue.org`
[2] `http://spectrum.ieee.org/automaton/robotics/humanoids/drc-finals-course`
[3] `http://www.eurathlon.eu`

of specifying how a particular task (use case) should be accomplished. Typically, participating end-users are provided with a scenario and a description of circumstances, and are left relatively free to tackle the task as they see fit. It could be argued that the TRADR approach is more 'realistic' because of this, the upshot being that it is harder to quantify mission outcomes and comparison of performance between different years and missions is more challenging compared to more controlled scenarios. Towards this end, we have started developing a more structured evaluation methodology (see section 1.3.2).

1.4.2 Evaluation methodology

Various evaluation methods for HRI have been proposed, e.g. [30, 36, 38, 33]. Some of these are particularly relevant for Robot-Assisted Search and Rescue (RADR). For instance, [5] proposed the RADR Coding System, which aims to provide a standardized scheme for coding RADR missions. [27] provide generalizable metrics for human-robot teaming and apply these on a search-and-rescue task. Although highly relevant, this case concerns a single operator controlling multiple robots. Others have focused on particular RADR aspects, e.g. evaluation of map creation [3] or evaluation of interfaces [14].

In contrast, to evaluate a socio-technical system of the type that we focus on, we need a broader as well as a higher systems level of evaluation. So instead of isolating single HRI events (e.g. UGV-operator) or test variables (e.g. situation awareness) for evaluation, we aim to capture a holistic perspective and evaluate the socio-technical system at large, while also addressing individual functionalities.

Standardized approaches also exist, such as e.g. the NIST Reference Test Arena [31] and the ASTM set of standards for search and rescue robotics operations[4] which are used in Robocup Rescue League competitions. Inspired by the benchmarking approach of RoCKIn[5], the Eurathlon competition 2015 [1] includes a system-level (Task) benchmark and module-level (Functionality) benchmark in order to capture the performance of a system as a whole as well as the contributions and performance of individual modules.

Our approach is similar to Eurathlon in that we also aim to evaluate the system in a realistic scenario, at different levels, in order to gain more insight. But the difference is that we aim specifically at socio-technical systems with a focus on human-robot interaction as a key factor that co-determines the performance of such a system. The additional challenge we tackle is that we address human-robot collaboration by embedding robots into actual disaster response teams.

[4] `http://www.astm.org/COMMIT/SUBCOMMIT/E5408.htm`
[5] `http://rockinrobotchallenge.eu/`

1.4.3 The Ethical Landscape of Robot-Assisted Search and Rescue

Work specifically addressing the ethics of disaster response is hard to find, but some guidelines exists. For instance, the Council of Europe's "Ethical principles on disaster risk reduction and people's resilience" [28] specifically dictates how rescue workers should behave ethically, as well as specifying that rescue workers should have access to psychological assistance because they are exposed to the risk of developing psychiatric and post-traumatic stress disorders [16, 7, 4]. Other work has discussed the ethics of disaster management [17], drawing parallels with ethics of humanitarian aid, while others specifically address ethics of firefighters [29], comparing it with the ethics of the medical profession.

There is some work addressing robot ethics against a backdrop of the SAR domain. For instance, [21] propose a method of modeling accountability in human-robot teams, thus endowing artificial systems (robots) with some form of moral accountability. However, to the best of our knowledge there is relatively little work explicitly addressing the ethics of robot-assisted SAR. In order to get a better grip on the ethical concerns regarding SAR robots, we therefore discuss ethics surrounding the use of robots in the healthcare [37, 32, 12, 6] and military domain [2, 24, 35, 13, 23]. There is a considerable amount of work on roboethics in these domains, and both have links with the SAR domain (healthcare resembles victim care in SAR, and both military and SAR robots are used in rough and unknown terrains to perform reconnaissance and search for targets).

2 Annexes

This is the way to list annexes:

2.1 Joachim de Greeff "TRADR Scenario and Use Cases Year 2"

Bibliography Joachim de Greeff, "TRADR Scenario and Use Cases Year 2". Unpublished technical report, Interactive Intelligence, TU Delft, the Netherlands, 2015.

Abstract In this report we describe the scenario and use cases for TRADR Year 2. In this year we have started with making a distinction between Generic Use Cases (GUCs) which describe a certain functionality at an abstract level, and Evaluation Use Cases (EUCs), which are tailored towards the particular Year 2 scenario, and use the GUCs as building blocks.

Relation to WP This document describes the scenario and uses cases for Year 2. As such it is at the very core of WP7.

Availability Restricted. Not included in the public version of this deliverable.

2.2 Joachim de Greeff, Tina Mioch, Nanja Smets, Koen Hindriks, Mark Neerincx and Ivana Kruijff-Korbayová (2015), "TRADR Evaluation Methodology"

Bibliography Joachim de Greeff, Tina Mioch, Nanja Smets, Koen Hindriks, Mark Neerincx and Ivana Kruijff-Korbayová (2015), "TRADR Evaluation Methodology". Unpublished technical report, Interactive Intelligence, TU Delft, the Netherlands, 2015.

Abstract In this report we describe the evaluation methodology as further developed in year 2. Field tests, such as T-JEx and T-Eval are – due to their size, complexity and the large number of (experimental) components involved, as well as the inclusion of human users – non-trivial exercises to evaluate. Towards this end, we develop an evaluation methodology which takes into account multiple levels of granularity, including low-level system foundation, mid-level functional components and high-level mission goals for disaster response, combined with a strong focus on human factors. The methodology contributes to understanding interdependencies between various system components, and provides a comprehensive insight into the readiness of the TRADR system.

Relation to WP This document describes a further refinement of the methodology used for evaluation of the TRADR technology. As such it is of direct relevance to WP7.

Availability Restricted. Not included in the public version of this deliverable.

2.3 Joachim de Greeff and Chris Rozemuller "TRADR Scenario-based Evaluation Exercises Year 2"

Bibliography Joachim de Greeff and Chris Rozemuller, "TRADR Scenario-based Evaluation Exercises Year 2". Unpublished technical report, Interactive Intelligence, TU Delft, the Netherlands, 2015.

Abstract In this report we describe the scenario based evaluations as executed in Year 2. Specifically, three exercise are reported: T-JEx, T-Eval and ITEX.

Relation to WP This document reports on the three scenario-based evaluation exercises as conduced during Year 2. As such it is at the very core of WP7.

Availability Restricted. Not included in the public version of this deliverable.

2.4 Mioch (2015), "T-Eval Year 2 Overview Feedback End-users"

Bibliography Mioch, Tina. "T-Eval Year 2 Overview Feedback End-users", Unpublished technical report, TNO, December 2015.

Abstract In this report, we describe the feedback we have received from the end-users during T-Eval 2015. The feedback is organized by WP. The feedback list serves as input for the work plan for the upcoming period.

Relation to WP The report describes the feedback we have received from end-users during and after the use of the TRADR system, and how this feedback is used to improve the system. As such, it is at the core of WP7.

Availability Restricted. Not included in the public version of this deliverable.

2.5 Tonino Guerrieri, Roberto Perna, Salvatore Corrao, Guido van den Broek Humphrey, Norbert Pahlke and Robert Maul (2015), "Joint TRADR Year 2 End Users Report"

Bibliography Tonino Guerrieri, Roberto Perna, Salvatore Corrao, Guido van den Broek Humphrey, Norbert Pahlke and Robert Maul. "Joint TRADR Year 2 End Users Report", Unpublished technical report, FDDo, IFR and GB, December 2015.

Abstract This report presents the observations of the end-users in the TRADR project of year 2.

Relation to WP The report describes the experiences from the end-users' point of view. As such, it is at the core of WP7.

Availability Restricted. Not included in the public version of this deliverable.

2.6 Maaike Harbers, Joachim de Greeff, Ivana Kruijff-Korbayová, Mark Neerincx and Koen Hindriks (2015), "Exploring the Ethical Landscape of Robot-Assisted Search and Rescue"

Bibliography Maaike Harbers, Joachim de Greeff, Ivana Kruijff-Korbayová, Mark Neerincx and Koen Hindriks (2015), "Exploring the Ethical Landscape of Robot-Assisted Search and Rescue". In proceedings of the International Conference on Robot Ethics (ICRE 2015), Lisbon, Portugal.

Abstract As robots are increasingly used in Search and Rescue (SAR) missions, it becomes highly relevant to study how SAR robots can be developed and deployed in a responsible way. In contrast to some other robot application domains, e.g. military and healthcare, the ethics of robot-assisted SAR are relatively under examined. This paper aims to fill this gap by assessing and analyzing important values and value tensions of stakeholders of SAR robots. The paper describes the outcomes of several Value Assessment workshops that were conducted with rescue workers, in the context of a European research project on robot-assisted SAR (the TRADR project). The workshop outcomes are analyzed and key ethical concerns and dilemmas are identified and discussed. Several recommendations for future ethics research leading to responsible development and deployment of SAR robots are provided.

Relation to WP This work explores ethical concerns that may arise from the use of robots in the search and rescue domain. It is relevant to this workpackage, as such concerns could manifest themselves during scenario-based evaluations with end-users. Moreover, through this exploratory study, we conciser the potential (ethical) impact of projects such as TRADR in a wider societal context.

Availability Restricted, final version yet to be published. Not included in the public version of this deliverable.

References

[1] F. Amigoni, E. Bastianelli, J. Berghofer, A. Bonarini, G. Fontana, N. Hochgeschwender, L. Iocchi, G. Kraetzschmar, P. Lima, M. Matteucci, P. Miraldo, D. Nardi, and V. Schiaffonati. Competitions for benchmarking: Task and functionality scoring complete performance assessment. *Robotics Automation Magazine, IEEE*, 22(3):53–61, Sept 2015.

[2] R. Arkin. *Governing lethal behavior in autonomous robots*. CRC Press, 2009.

[3] B. Balaguer, S. Balakirsky, S. Carpin, and A. Visser. Evaluating maps produced by urban search and rescue robots: lessons learned from robocup. *Autonomous Robots*, 27(4):449–464, 2009.

[4] J. Bos, E. Mol, B. Visser, and M. H. Frings-Dresen. The physical demands upon (dutch) fire-fighters in relation to the maximum acceptable energetic workload. *Ergonomics*, 47(4):446–460, 2004.

[5] J. L. Burke, R. R. Murphy, D. R. Riddle, and T. Fincannon. Task performance metrics in human-robot interaction: Taking a systems approach. In *METRICS INTELL. SYST. WORKSHOP*, 2004.

[6] M. Butter, A. Rensma, J. v. Boxsel, S. Kalisingh, M. Schoone, M. Leis, G. Gelderblom, G. Cremers, M. d. Wilt, W. Kortekaas, et al. Robotics for healthcare: final report. 2008.

[7] C.-M. Chang, L.-C. Lee, K. M. Connor, J. R. Davidson, K. Jeffries, and T.-J. Lai. Posttraumatic distress and coping strategies among rescue workers after an earthquake. *The Journal of nervous and mental disease*, 191(6):391–398, 2003.

[8] J. de Greeff. TRADR scenario and use-cases year 2. Technical report, Interactive Intelligence, TU Delft, 2015. Unpublished.

[9] J. de Greeff, T. Mioch, N. Smets, K. Hindriks, M. Neerincx, and I. Kruijff-Korbayová. TRADR evaluation methodology. Technical report, Interactive Intelligence, TU Delft, 2015. Unpublished.

[10] J. de Greeff, N. Pahlke, R. Maul, J. van Diggelen, T. Mioch, T. Bagosi, M. Janíček, and the TRADR consortium. TRADR Deliverable DR 3.2: Layered SA for tactical decision making. Deliverable, EU FP7 TRADR / ICT-60963, 2015.

[11] J. de Greeff and C. Rozemuller. TRADR scenario-based evaluation exercises year 2. Technical report, Interactive Intelligence, TU Delft, 2015. Unpublished.

[12] M. Decker. Caregiving robots and ethical reflection: the perspective of interdisciplinary technology assessment. *Ai & Society*, 22(3):315–330, 2008.

[13] B. L. Docherty. *Losing Humanity: The Case Against Killer Robots.* Human Rights Watch, 2012.

[14] J. Drury, L. D. Riek, A. D. Christiansen, Z. T. Eyler-Walker, A. J. Maggi, and D. B. Smith. Evaluating human-robot interaction in a search-and-rescue context. In *Proc. of PERMIS*, 2003.

[15] B. Friedman, P. H. Kahn Jr, A. Borning, and A. Huldtgren. Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory*, pages 55–95. Springer, 2013.

[16] C. S. Fullerton, J. E. McCarroll, R. J. Ursano, and K. M. Wright. Psychological responses of rescue workers: fire fighters and trauma. *American journal of orthopsychiatry*, 62(3):371, 1992.

[17] S. K. Geale. The ethics of disaster management. *Disaster Prevention and Management: An International Journal*, 21(4):445–462, 2012.

[18] T. Guerrieri, R. Perna, S. Corrao, G. van den Broek Humphrey, N. Pahlke, and R. Maul. Joint TRADR year 2 end users report. Technical report, Openbaar Lichaam Gezamenlijke Brandweer, Netherlands (GB), the Corpo Nazionale dei Vigili del Fuoco, Italy (CNVVF) and Feuerwehr Dortmund, Germany (FDDO), 2015. Unpublished.

[19] M. Harbers, J. de Greeff, I. Kruijff-Korbayová, M. Neerincx, and K. Hindriks. Exploring the ethical landscape of robot-assisted search and rescue. In *International Conference on Robot Ethics, Lisbon*, 2015.

[20] Kasper, W. et al. and the TRADR consortium. TRADR Deliverable DR 5.2: Expectation management in common ground. Deliverable, EU FP7 TRADR / ICT-60963, 2015.

[21] G.-J. M. Kruijff and M. Janíček. Using doctrines for human-robot collaboration to guide ethical behavior. In *AAAI Fall Symposium: Robot-Human Teamwork in Dynamic Adverse Environment*, 2011.

[22] P. Łichocki, A. Billard, and P. H. Kahn Jr. The ethical landscape of robotics. *Robotics & Automation Magazine, IEEE*, 18(1):39–50, 2011.

[23] P. Lin, G. A. Bekey, and K. Abney. Robots in war: issues of risk and ethics. 2009.

[24] G. R. Lucas Jr. Industrial challenges of military robotics. *Journal of Military Ethics*, 10(4):274–295, 2011.

[25] T. Mioch. T-Eval year 2 overview feedback end-users. Technical report, TNO, 2015. Unpublished.

[26] M. A. Neerincx and J. Lindenberg. Situated cognitive engineering for complex task environments. In *Naturalistic Decision Making & Macrocognition*, pages 373–390. Ashgate Publishing Limited, 2008.

[27] P. Pina, M. Cummings, J. Crandall, and M. Della Penna. Identifying generalizable metric classes to evaluate human-robot teams. In *Proc. 3rd Ann. Conf. Human-Robot Interaction*, pages 13–20. Citeseer, 2008.

[28] M. Prieur. Council of Europe. European and Mediterranean Major Hazards Agreement (EUR-OPA). `http://www.preventionweb.net/english/\professional/publications/v.php?id=26384`, 2012. [Online; accessed 17-06-2015].

[29] P. Sandin. Firefighting ethics: Principlism for burning issues. *Ethical Perspectives*, 16(2):225–251, 2009.

[30] J. Scholtz. Theory and evaluation of human robot interactions. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, pages 10–pp. IEEE, 2003.

[31] J. Scholtz, B. Antonishek, and J. Young. Evaluation of human-robot interaction in the nist reference search and rescue test arenas. Technical report, DTIC Document, 2004.

[32] A. Sharkey and N. Sharkey. Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1):27–40, 2012.

[33] D. Y. Y. Sim and C. K. Loo. Extensive assessment and evaluation methodologies on assistive social robots for modelling human–robot interaction–a review. *Information Sciences*, 301:305–344, 2015.

[34] N. Smets, J. de Greeff, M. Harbers, S. Corrao, N. Pahlke, G. van den Broek Humphrey, M. Gianni, and the TRADR consortium. TRADR Deliverable DR 7.1: Scenario-based evaluation in a large-scale static disaster area. Deliverable, EU FP7 TRADR / ICT-60963, 2015.

[35] R. Sparrow. Killer robots. *Journal of applied philosophy*, 24(1):62–77, 2007.

[36] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 33–40. ACM, 2006.

[37] A. van Wynsberghe. Designing robots for care: Care centered value-sensitive design. *Science and engineering ethics*, 19(2):407–433, 2013.

[38] J. E. Young, J. Sung, A. Voida, E. Sharlin, T. Igarashi, H. I. Christensen, and R. E. Grinter. Evaluating human-robot interaction. *International Journal of Social Robotics*, 3(1):53–67, 2011.