Dipl.-Psych. Marija Bertović

Human Factors in Non-Destructive Testing (NDT): Risks and Challenges of Mechanised NDT

BAM-Dissertationsreihe · Band 145
Berlin 2016

Die vorliegende Arbeit entstand an der Bundesanstalt für Materialforschung und -prüfung (BAM).

Impressum

**Human Factors in
Non-Destructive Testing (NDT):
Risks and Challenges of Mechanised NDT**

2016

Herausgeber:
Bundesanstalt für Materialforschung und -prüfung (BAM)
Unter den Eichen 87
12205 Berlin
Telefon: +49 30 8104-0
Telefax: +49 30 8104-72222
E-Mail: info@bam.de
Internet: www.bam.de

Layout: BAM-Referat Z.8

ISSN 1613-4249
ISBN 978-3-9817502-7-0

Human Factors in Non-Destructive Testing (NDT): Risks and Challenges of Mechanised NDT

Vorgelegt von
Dipl. -Psych.
Marija Bertovic
geb. in Ogulin, Kroatien

von der Fakultät V – Verkehrs- und Maschinensysteme
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktorin der Philosophie
-Dr. phil.-

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. phil. Manfred Thüring
Gutachter:	Prof. Dr. phil. Dietrich Manzey
Gutachter:	Dr. rer. nat. et Ing. habil. Gerd-Rüdiger Jaenisch

Tag der wissenschaftlichen Aussprache: 1 September 2015

Berlin 2015
D83

Abstract

Non-destructive testing (NDT) is regarded as one of the key elements in ensuring quality of engineering systems and their safe use. A failure of NDT to detect critical defects in safety-relevant components, such as those in the nuclear industry, may lead to catastrophic consequences for the environment and the people. Therefore, ensuring that NDT methods are capable of detecting all critical defects, i.e. that they are reliable, is of utmost importance.

Reliability of NDT is affected by human factors, which have thus far received the least amount of attention in the reliability assessments. With increased use of automation, in terms of mechanised testing (automation-assisted inspection and the corresponding evaluation of data), higher reliability standards are believed to have been achieved. However, human inspectors, and thus human factors, still play an important role throughout this process, and the risks involved in this application are unknown.

The overall aim of the work presented in this dissertation was to explore for the first time the risks associated with mechanised NDT and find ways of mitigating their effects on the inspection performance. Hence, the objectives were to (1) identify and analyse potential risks in mechanised NDT, (2) devise measures against them, (3) critically address the preventive measures with respect to new potential risks, and (4) suggest ways for the implementation of the preventive measures.

To address the first two objectives a risk assessment in form of a Failure Modes and Effects Analysis (FMEA) was conducted (Study 1). This analysis revealed potential for failure during both the acquisition and evaluation of NDT data that could be assigned to human, technology, and organisation. Since the existing preventive measures are insufficient to defend the system from identified failures, new preventive measures were suggested. The conclusion of the study was that those preventive measures need to be carefully considered with respect to *new* potential risks, before they can be implemented, thus serving as a starting point for further empirical studies.

To address the final two objectives, two preventive measures, i.e. human redundancy and the use of automated aids in the evaluation of NDT data, were critically assessed with regard to potential downfalls arising from the social interaction between redundant individuals and the belief in the high reliability of automated aids.

The second study was concerned with the potential withdrawal of effort in sequential redundant teams when working collectively as opposed to working alone, when independence between the two redundant individuals is not present. The results revealed that the first redundant inspector, led to believe someone else will conduct the same task afterwards,

invested the same amount of effort as when working alone. The redundant checker was not affected by the information about the superior experience of his predecessor and—instead of expected withdrawal of effort—exhibited better performance in the task. Both results were in contradiction to the hypotheses, the explanations for which can be found in the social loafing and social compensation effects and in the methodological limitations.

The third study examined inappropriate use of the aid measured in terms of (a) agreement with the errors of the aid in connection to the frequency of verifying its results and in terms of (b) the overall performance in the task. The results showed that the information about the high reliability of the aid did not affect the perception of that aid's performance and, hence, no differences in the actual use of the aid were to be expected. However, the participants did not use the aid appropriately: They misused it, i.e. agreed with the errors committed by the aid and disused it, i.e. disagreed with the correct information provided by the aid, thereby reducing the overall reliability of the aid in terms of sizing ability. Whereas aid's misuse could be assigned to low propensity to take risks and reduced verification behaviour because of a bias towards automation, the disuse was assigned to the possible misunderstanding of the task.

The results of these studies raised the awareness that methods used to increase reliability and safety, such as automation and human redundancy, can backfire if their implementation is not carefully considered with respect to new potential risks arising from the interaction between individuals and complex systems. In an attempt to minimise this risk, suggestions for their implementation in the NDT practice were provided.

Zusammenfassung

Die zerstörungsfreie Prüfung (ZfP) wird als eines der wichtigsten Qualitätssicherungsmaßnahmen für technische Systeme und deren sichere Anwendung betrachtet. Wenn die ZfP kritische Defekte in sicherheitsrelevanten Anlagen, wie z.B. in der Kerntechnik, nicht entdeckt, kann dies zu katastrophalen Folgen für die Umwelt und den Menschen führen. Deshalb muss gewährleistet sein, dass die Verfahren der ZfP zuverlässig sind, d.h. dass sie alle kritischen Defekte entdecken können.

Die Zuverlässigkeit der ZfP wird von menschlichen Faktoren beeinflusst, die jedoch bisher in diesem Feld selten betrachtet wurden. Durch den verstärkten Einsatz von Automatisierung beispielsweise bei der mechanisierten Prüfung (automatisierungsunterstütze Prüfung und die zugehörige Datenbewertung) wurde die Erreichung eines höheren Zuverlässigkeitsniveaus erwartet. Menschliche Faktoren sind trotz der Automatisierung immer noch bedeutsam für den gesamten Prüfprozess. Die Risiken der stärkeren Automatisierung der Prüfungen sind nicht vollständig bekannt.

Das generelle Anliegen der Autorin dieser Arbeit ist die erstmalige Feststellung der Risiken der mechanisierten ZfP und das Aufzeigen von Möglichkeiten, diese zu verringern. Die konkreten Ziele dieser Arbeit sind dementsprechend (1) die potenziellen Risiken bei der mechanisierten Prüfung aufzuzeigen und zu analysieren, (2) präventive Maßnahmen für diese Risiken abzuleiten, (3) diese präventiven Maßnahmen kritisch hinsichtlich neuer Risiken zu beleuchten sowie (4) Umsetzungsvorschläge aufzuzeigen.

Für die ersten zwei Ziele wurde eine Risikoabschätzung mit der Fehlzustandsart- und -auswirkungsanalyse (FMEA) durchgeführt (Studie 1). Diese Analyse ergab Fehlermöglichkeiten während der Datenaufnahme und –bewertung bei der mechanisierten ZfP, die dem Menschen, der Technik und der Organisation zugeordnet werden können. Weil die vorhandenen präventiven Maßnahmen unzureichend für die Vermeidung der identifizierten Fehler waren, wurden neue präventive Maßnahmen vorgeschlagen. Die Schlussfolgerung der Studie zeigt, dass vor der Umsetzung präventiver Maßnahmen eine sorgfältige Betrachtung hinsichtlich *neuer* potenzieller Risiken erfolgen muss. Dies war der Ausgangspunkt für die weiteren empirischen Untersuchungen.

Für die letzten beiden Ziele wurden zwei präventive Maßnahmen untersucht: die menschliche Redundanz und die Anwendung automatisierter Assistenzsysteme bei der ZfP-Datenbewertung. Im Fokus lagen potenzielle Schwachstellen, die aus sozialer Interaktion der redundanten Individuen und aus dem Vertrauen in die hohe Zuverlässigkeit der automatisierten Assistenzsysteme entstehen können.

In der zweiten Studie wurde die potenzielle Reduzierung der Anstrengung in sequentiellen redundanten Teams untersucht, indem die gemeinsame Aufgabenbearbeitung in Teams der individuellen Aufgabenbearbeitung gegenüber gestellt wurde. Die Ergebnisse zeigten, dass der erste redundante Prüfer, dem mitgeteilt wurde, dass ein anderer Prüfer die Prüfaufgabe nach ihm durchführen wird, die gleiche Anstrengung investierte wie der individuelle Bearbeiter. Der zweite redundante Prüfer (*redundant checker*) wurde durch die Information, dass sein Vorprüfer die höherwertige Erfahrung besitzt, nicht hypothesenkonform beeinflusst—anstelle der erwarteten Rücknahme der Anstrengung—zeigte er eine bessere Leistung bei der Durchführung der Aufgabe. Beide Ergebnisse stehen in Wiederspruch zu den Hypothesen und können durch *social loafing* und *social compensation* Effekte sowie durch methodische Aspekte erklärt werden.

In der dritten Studie wurde die unangemessene Nutzung eines automatisierten Assistenzsystems untersucht operationalisiert als (a) die Übereinstimmung mit Fehlern des Systems verbunden mit der Überprüfungshäufigkeit seiner Ergebnisse und (b) die Leistung bei der Aufgabe. Die Ergebnisse zeigten, dass die Information über die hohe Zuverlässigkeit des Systems die Wahrnehmung der Systemleistung nicht beeinflusste und folglich keine Unterschiede in der tatsächlichen Nutzung des Systems zu finden waren. Die Probanden nutzten jedoch das System nicht angemessen: sie stimmten den Fehlern des Systems zu (*automation misuse*) und sie lehnten korrekte Informationen des Systems ab (*automation disuse*). So reduzierten sie die Gesamtzuverlässigkeit des Systems, zumeist bei der Fehlergrößenbestimmung. Während *misuse* mit einer niedrigen Risikobereitschaft und eingeschränkten Überprüfungsverhalten auf Grund des *automation bias* erklärt werden kann, wird *disuse* dem möglichen Missverstehen der Aufgabe zugeordnet.

Die Ergebnisse dieser Studien haben das Bewusstsein dafür erhöht, dass Methoden zur Erhöhung der Zuverlässigkeit und Sicherheit sowie Automatisierung und menschliche Redundanz versagen können, wenn die potenziellen Risiken ihrer Umsetzung aufgrund der Interaktion zwischen Mensch und Technik nicht bedacht werden. Um diese Risiken bei der Anwendung präventiver Maßnahmen zu minimieren, wurden Vorschläge für die ZfP-Praxis erarbeitet.

Table of Contents

1. Introduction

Increasing demands for safety in our daily life and the infrastructure around us have led to the development of appropriate risk management and life prediction tools. Quantifiable non-destructive testing (NDT) methods are a key in providing substantial information about the integrity of materials, components, and systems by evaluating their properties without affecting their state in any way (e.g. V. Deutsch, Platte, Schuster, & Deutsch, 2006; Erhard, 2013; McGonnagle, 1975; Raj & Venkatraman, 2013). Non-destructive examination (NDE), non-destructive inspection (NDI), and non-destructive evaluation (NDE) are commonly used synonyms for NDT.

NDT is defined as "*an examination, test, or evaluation performed on any type of test object without changing or altering that object in any way, in order to determine the absence or presence of conditions or discontinuities that may have an effect on the usefulness or the serviceability of that object*" (Hellier, 2013, p. 1.1). According to the German Society for Non-Destructive Testing, NDT is one of the most important methods of safety monitoring (V. Deutsch et al., 2006), as it is an "*essential part of quality control of engineering systems and their safe use*" (Erhard, 2013, p. 161). Failure of NDT to detect a critical defect in the material—even though it is just a first step in technical diagnostic—might lead to catastrophic events and endanger society and the environment (Erhard, 2013).

Determining whether NDT is *capable* to find all *critical* defects is expressed in terms of reliability. It is known that the capability of NDT to find critical potentially structure-breaking defects in the materials depends not only on the technical capability of the equipment, but also on the conditions under which the NDT inspection is carried out and the human and organisational factors (Müller et al., 2013). Whereas technical capability is typically addressed using quantifiable methods, i.e. the Probability of Detection (POD) curves[1], human factors are difficult to include in the quantitative assessments, and are hence, largely neglected in the reliability assessment.

Addressing human factors in NDT plays a vital role in ensuring safety of organisations with high reliability and safety demand. It is best illustrated on examples of accidents and events that had happened due to too little attention being dedicated to human factors. One such example is the crash of the United Airlines flight 232 in Sioux Gateway Airport (Sioux City, Iowa) en route from Denver to Philadelphia. In the accident report following the crash, the

[1] POD is used to ascertain whether a defect of a certain quality, for example the size, will be detected with 90% probability and 95% confidence.

National Transportation Safety Board (1989) summarised the accident and the cause of the crash:

> On July 19, 1989, at 1516, a DC-10-10, N1819U, operated by United Airlines as flight 232, experienced a catastrophic failure of the No. 2 tail-mounted engine during cruise flight. The separation, fragmentation and forceful discharge of stage 1 fan rotor assembly parts from the No. 2 engine led to the loss of the three hydraulic systems that powered the airplane's flight controls. The flight crew experienced severe difficulties controlling the airplane, which subsequently crashed during an attempted landing at Sioux Gateway Airport, Iowa. There were 285 passengers and 11 crewmembers onboard. One flight attendant and 110 passengers were fatally injured.
>
> The National Transportation Safety Board determines that the probable cause of this accident was the inadequate consideration given to human factors limitations in the inspection and quality control procedures used by United Airlines' engine overhaul facility, which resulted in the failure to detect a fatigue crack originating from a previously undetected metallurgical defect located in a critical area of the stage 1 fan disk […] [Executive summary, p. *v*].

In a more recent event, U.S. Nuclear Regulatory Commission (2012) issued a report on a failure of an NDT in-service inspection (ISI) in 2009 to identify five Primary Water Stress Corrosion Cracking (PWSCC) indications[2] in the steam generator (SG) safe-end weld. The two 100% through-wall and three partial through-wall indications exceeding the acceptance criteria were detected in a subsequent ISI in 2012. A post-event evaluation suggested that it was very likely that the indications existed in 2009, that the indications were within the inspectors' "*ability to foresee and correct*" (p. 11), and that the "*examinations performed on SG safe-end weld did not provide assurance that the structural boundary of the reactor coolant system remains capable of performing its intended safety function*" (p. 11). A root cause evaluation yielded the following conclusion:

> The evaluation determined that the root causes included that the site NDE organization did not adequately implement their responsibility to ensure effective application of the examination procedure by supplemental examination personnel and that the on-site briefing conducted with NDE technicians was not adequate to ensure successful execution of the examination (p. 10).

Even though this event did not lead to an accident, it illustrates that NDT can fail and that the underlying causes can extend beyond the technical capability.

The motivation for addressing human factors in NDT came from a number of inspections, during which significant variations in the individual performances were observed but could not be overcome by physical or engineering methods (e.g. Fücsök & Müller, 2000; McGrath, Worrall, & Udell, 2004; McGrath, 1999; Nichols & Crutzen, 1988; Nockemann, Heidt, & Thomsen, 1991).

The observed variability between the inspectors—shown to decrease NDT reliability—is typically addressed by improving of the inspectors' technical ability, by assuring only those that display their ability in a blind performance demonstration can be sent out to inspect (applies typically to the U.S.A.), by improving the inspection procedures, by introducing human redundancy, and by increased automation of the inspection process.

However, focusing only on the individual as a source of failure and applying methods such as more strict procedures, supervision, training and introducing automation to replace the human

[2] Indication is a representation of a signal from a discontinuity in the material in the format typical for the method used.

operators as much as possible is an approach that is slowly being abandoned, as looking into the interactions between all the present systems, i.e. the individual, the team, the technology, the organisation, and the extra-organisational environment, takes over the attention in the contemporary human factors and human error research (e.g. Badke-Schaub, Hofinger, & Lauche, 2012; Reason & Hobbs, 2003).

As many other domains, NDT is experiencing a rise in the use of automation. Automation-assisted inspection (in NDT—and hereafter—referred to as the *mechanised* inspection) is seen as beneficial, not only due to expected higher reliability and the advantage of storage of data for later use or re-evaluation, but also as a means of decreasing human variability observed during manual inspection. In this process, NDT inspection personnel is not replaced by automated systems and asked to monitor them, but the inspector is still actively involved in the setup of the measurement system, and, most importantly, in the evaluation of the collected data (even though not automated or mechanised *per se*, evaluation is, in this thesis, embedded in the term mechanised testing).

Contemporary human-automation interaction research suggests that increased automation is not only related to benefits, but also to costs—a paradox frequently dubbed as the *automation ironies* (Bainbridge, 1987) or *automation surprises* (Sarter, Woods, & Billings, 1997). Those ironies and surprises refer to those elements in the interaction between human users and automation that were not considered by the automation designers, but which can fundamentally change the responsibilities of the human operators of systems and the nature of the cognitive demands, e.g. the need for new skills, retention of old skills for problem solving, loss of situation awareness, different nature of workload, reliance on automation, etc. (e.g. Manzey, 2012; Parasuraman & Riley, 1997; Sarter et al., 1997; Sheridan & Parasuraman, 2005).

The costs of automation use in NDT were never investigated. Therefore, the overall aim of this study was to explore those risks and to find ways of preventing them. This was achieved by identifying the risks, their causes and consequences, and by suggesting preventive measures. However, as with introducing automation, implementing preventive measures can also lead to *new* risks. The two singled out preventive measures, i.e. the human redundancy and the use of automated decision aids, have yielded a substantial amount of attention in the human factors field. Even though both are associated with benefits for the increase of NDT reliability—the former as a measure of error recovery, and the latter as a defect detection aid—findings from the literature suggest that the lack of independence between the redundant elements (Clarke, 2005) and over-trust in automated aids (Parasuraman & Riley, 1997) may actually counteract the expected benefits. Hence, the objective of this study was to investigate for the first time the potential problems that can arise from the implementation of these measures in NDT.

Addressing risks associated with mechanised NDT and of implementing preventive measures is a relevant step in the long-term endeavour of ensuring that NDT inspections are performed reliably, thereby contributing to the overall safety of the organisation they serve to protect and maintain.

The relevance of addressing this issue is especially high in domains, in which failures can lead to catastrophic consequences. Nuclear industry and aviation have been, thus, the front-runners in human factors research. A new emerging field that is of safety concern—and will be, even after all the nuclear power plants stop producing energy—is the final disposal of spent nuclear fuel. No permanent repositories for spent nuclear fuel exist currently in the world, and the operation of the most developed programmes—that of Finland and Sweden—is scheduled for the near future. This application carries unknown risks, but at the same time,

the highest safety and reliability standards need to be achieved as the spent nuclear fuel is expected to be safely stored for at least the next 100,000 years (the extent of the safety assessment). The work presented in this dissertation has been conducted within the scope of this challenging new field and concentrates, hence, primarily on the nuclear domain.

The work is organised in the following manner: Following this introduction, the second chapter will present the theoretical underpinnings used as the motivation for the empirical work. The chapter will introduce the field of NDT, outline the role of human factors in NDT, present the current state of the art in the research field, and close with identifying the challenges of the field and the aims of the study.

The subsequent chapters are concerned with the empirical work conducted within the scope of this dissertation. The three conducted studies will each begin with a theoretical background, followed by the aims of that particular study, the hypotheses, the method, the results, and their individual discussions.

The first study will be presented in the third chapter. In the theoretical part, relevant aspects of human error and the fundamentals of risk management will be outlined, as the aim of this study is to identify risks associated with mechanised NDT. In the empirical part, will present the risk assessment approach chosen to identify the risks, the methodology, and the results will be presented, followed by the discussion of the findings. A special emphasis will be given to the identified potential preventive measures and the implications of their implementation. The chapter will conclude by raising questions about the optimal implementation of two preventive measures, i.e. the application of human redundancy and of the use of automated aids in the evaluation of NDT data, which lay the foundations for further empirical work.

The following two studies will be outlined in chapters 4-6. Taking into account the similarities in the applied method in both empirical studies, the fourth chapter will present the design of the experimental task and the recruitment of the participants.

The fifth chapter is concerned with the application of human redundancy in the evaluation of data acquired with mechanised systems. It will commence with the theoretical underpinnings of human redundancy and its potential downfalls when applied in complex industrial environments. Considering a particular scenario in which human redundancy can be implemented in NDT, i.e. sequential redundancy, the roles of both inspector roles will be scrutinised with respect to a) expectation one's work will be subdued to human redundancy, and b) the effects of familiarity between the inspectors on the performance. After presenting the method and the results of two experimental studies, this chapter will conclude with a discussion of the results and with suggestions for the implementation of human redundancy in NDT.

The interaction between the inspectors and an automated detection aid will be explored in the sixth chapter. The focus will be put on the belief in high or low reliability of the aid, propensity to take risks, differences in verification behaviour, and their potential consequences on inappropriate automation use in a data evaluation task. As in the previous study, after presenting and discussing the results, recommendations for overcoming problems that can arise from working with automated aids in NDT will be given.

In the seventh chapter, the results of all three studies will be jointly discussed with emphasis on their implications for the reliability of NDT.

2. Human factors in non-destructive testing

"The reconstruction of mindset begins not with the mind.
It begins with the circumstances in which the mind found itself"

Dekker, 2002, p. 50

The purpose of this chapter is to provide with general understanding of non-destructive testing, present the existing state of the art in the study of human factors in NDT, and identify research that is still missing in the field. This will lead to establishing the aims and the objectives of the study, which will be explored in the following chapters.

Everyone who had ever been examined by a doctor with an ultrasound or an x-ray machine had in fact been non-destructively examined, i.e. the doctors could learn about the condition of our internal organs and the functioning of our organism without surgically opening to see. Non-destructive methods are also widely used to inspect materials, without damaging them. To understand how these methods are used in materials' testing, this chapter will begin with addressing some fundamentals of non-destructive testing (NDT), with respect to different methods, applications, and reliability (section 2.1).

Human performance and its influence on reliability and safety in complex organisations, such as the nuclear power plants, have been a topic of research in the field of human factors since its beginnings. The obtained knowledge is continuously used to identify and prevent risks stemming from the interaction of humans with ever-developing technology. Some of that knowledge, relevant to this study, will be presented, including the sources of variability in the inspection performance, models of human performance in NDT, and a review of existing literature on human factors in NDT (section 0). Finally, the chapter will conclude with the challenges and the aims of the study (section 2.3).

2.1. Non-destructive testing

In our daily life, we rely on structures around us, such as the bridge we walk on, the plane we fly with or the nearby power plant, to hold and function as intended. One of the tasks in ensuring those structures will hold and withstand different conditions over the planned life cycle is to ensure that materials, components, or structures do not contain discontinuities—

deemed critical for the component—that could endanger their structural integrity and have an impact on their functionality. This is achieved by means of non-destructive testing.

NDT can be achieved using almost any kind of energy and letting it interact with the component, the oldest of which is using the naked eye to look for changes on the surface. Generally, NDT methods aim at understanding the interaction of some form of energy (in form of rays or waves) that is being sent through or into the material, and the material itself (Raj & Venkatraman, 2013). Each NDT method acts upon the material through its physical ability. For example, *ultrasonic testing* (UT) sends sound waves through the material, *eddy current testing* (ET) is based on induction of currents, *radiographic testing* (RT) penetrates the material with electromagnetic waves with high frequency (x-ray), and in *visual testing* (VT) the objects is illuminated and then inspected with an eye or with other photosensitive devices, such as a camera. The choice of the method to be used depends on the material, geometry, defect type and location, applicability, accessibility, and suitability (e.g. Raj & Venkatraman, 2013). For example, UT and RT are best suitable to look for discontinuities in the volume of the component, ET for discontinuities at or near to the surface, and VT for discontinuities at or open to the surface. Other frequently applied methods include thermography, liquid penetrant testing, magnetic particle testing, etc.

In addition to the physical principles of each method and the location of the discontinuities that need to be found, the method with which NDT is applied depends on the degree of involvement of the inspector and the technology. With that respect, NDT methods can be divided into manual, semi-automated (or automation-assisted), and fully automated methods. In manual UT, for example, an inspector typically goes on-site with a hand-held device, prepares the component and the equipment, conducts the inspection by manually displacing the probe (the sender and the receiver of the energy, i.e. ultrasound, being sent through the material) along the component, and interprets the signals.

In semi-automated testing, i.e., *mechanised testing*, the testing equipment is mounted on a manipulator and the data acquisition is carried out by automatically displacing the probe along the inspection area. Still, the inspector, or a team of inspectors, is involved in the preparation of the equipment and the setup of the software, monitoring of the automatic scanning process, and—most importantly—they are involved in the evaluation of the collected data. Even though only the process of acquisition can be mechanised, the corresponding data evaluation bears characteristics that makes it significantly different from that of evaluating signals in manual inspection and rather specific to mechanised inspection. E.g., in comparison to some manual NDT methods (e.g. UT), in which only the raw signal amplitudes are instantaneously evaluated, evaluation of data acquired by mechanised methods benefits from the possibility of using several views and tools for visualisation of the collected data during the analysis. Thus, in this thesis, the term "mechanised testing" includes both acquisition and evaluation of data.

Fully automated NDT is carried out without direct involvement of the human operator, where data collection, data evaluation, and even the decisions are performed automatically.

Absolutely homogeneous material does not exist. All materials contain discontinuities, but not all of them threaten the integrity of the structure. NDT is used to record all signals of discontinuities in the material, to report all those that exceed a predetermined *reporting threshold* and, finally, to identify those discontinuities in the material that are large enough to cause concern of structural failure, i.e. the *critical* defects (Ali, Balint, Temple, & Leevers, 2012). A defect is defined as "*a component discontinuity that has shape, size, orientation, or location, such that it is detrimental to the useful service of the part*" (Hellier, 2013, p.2.24).

There are different kinds of defects (e.g. cracks, delaminations, pores, inclusions) and they occur during the material's forming process, component manufacturing (both prior to its use), or develop in the material during the operational period (in-service) (McGonnagle, 1975). In industries, in which in-service inspection (ISI) is used, e.g. in the nuclear industry and aviation, structures with defects can be put to further use, provided that the defect size is regularly monitored before the defect grows to a critical size that could result in structural failure. This is known as the *damage tolerant design* (Raj & Venkatraman, 2013). Judgment about the materials being free of critical defects will, thus, allow those materials, components and systems to be used (V. Deutsch et al., 2006). If faulty, it could lead to devastating effects for the people and the environment.

2.1.1. NDT task

The primary role of the inspector is to detect (i.e. identify, find) and interpret signals received from the equipment (i.e. determine the size, shape, orientation and, frequently, the type of found indications).

In doing so, the inspector is guided by a set of inspection procedures and instructions, which determine how to find and analyse the signals the equipment is providing. Inspection procedures and instructions are some of the most important tools in everyday life of an NDT inspector. An inspection procedure is a written description of all essential parameters that need to be applied when carrying out an inspection. NDT instruction, on the other hand, is a precise, written description of the steps that must be followed during testing (DIN EN ISO 9712:2012). NDT procedures and instructions are typically written by certified personnel in accordance with standards, codes, or specifications.

The NDT process is embedded into an organisational context that goes beyond the organisation carrying out the inspection. Hence, the inspector (or a team) carrying out the task is influenced by the organisation it serves, but also by the service provider, regulators, and the international and national laws and regulations (Figure 1).

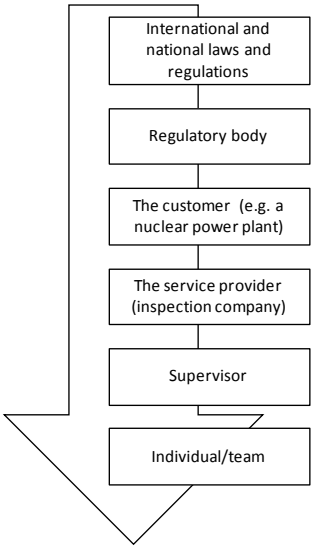


Figure 1: Organisational structures affecting the NDT inspections

Having interviewed about 200 German inspection companies, and having carried out a series of workshops, expert panel discussions, and five case studies, the German Institute for Vocational Education and Training *f-bb* (*Forschungsinstitut Berufliche Bildung*) described the overall NDT duties (Zeller, Küfner, & Neumann, 2012), as depicted in Figure 2.

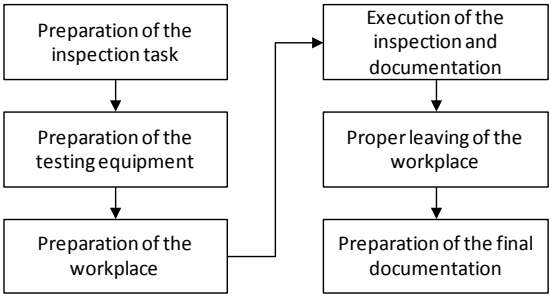


Figure 2: Typical NDT inspection working cycle (adapted and translated from "*Untersuchung zu neuen und modernisierten Berufsprofilen und einem Berufsgruppenprinzip für prüftechnische Berufe.* Abschlussbericht" [unpublished], by B. Zeller, C. Küfner, and F. Neumann, 2012, Nürnberg: Forschungsinstitut Betriebliche Bildung (f-bb) GmbH)

A successful NDT inspection requires careful planning, choice of appropriate NDT inspection methods and techniques, quality inspection procedures, and qualified and trained inspection personnel. Whereas these decisions are made by the service provider and the customer in accordance with the prescribed laws and procedures, the NDT inspection task (execution of the inspection + documentation) can be broken down into the three main steps. First, the inspector needs to *prepare*, i.e. familiarise himself with the inspection procedure, prepare the equipment and the inspection area or the component, and adjust the equipment to the appropriate sensitivity to find the desired discontinuities in the structure (a process also known as the calibration). Second, the inspector needs to *inspect* the structure. This process is highly dependent on the NDT method, as the tools one uses (hardware and software) differ. In any case, during the inspection one searches for defects. If they are at the surface, one can decide by looking at them or by touching them whether they are defects that need to be reported. If they are situated in the component's volume, they can no longer be seen, which makes the inspection task more demanding. In this case, the inspector does not identify and interpret defects (discontinuities that might be harmful for the component), but rather the signals received by the equipment. These are called *indications*. Not all indications necessarily reflect defects in the component, i.e. signals can also come from the geometry of the component, the equipment, the material's structure, etc. The task of the inspector is to distinguish between signals coming from discontinuities or other sources and further analyse those they are required to (this requirement depends on the magnitude of the signals above a predetermined reporting threshold, which can be determined, e.g. by the fracture mechanics). For the process of indication interpretation, the signals are analysed to that extent that the inspector is estimating the defect's size, position, the orientation and, if possible, the type. The final step is to *report* (document) all the findings. (How the findings of the inspection are dealt with goes beyond the task of the NDT inspector and will not be discussed in this thesis.)

NDT task is, hence, a very complex task, during which the inspectors have to rely on their sensory, perceptual, cognitive, and motor skills. It is described as signal detection, information processing, and as a decision-making task. It requires high vigilance, because inspections are

often carried out over long periods of time, on large components, with attention shifting from the object to some kind of a display; the task being frequently monotonous, tedious, and even boring. Moreover, the operating personnel require skill, knowledge, training, experience, and official qualification to be able to carry out the task. The task is greatly affected by the environment in which it is carried out. It frequently includes high noise, vibration, high temperatures and humidity, poor lighting, restricting working place, in nuclear industry (of interest for this thesis) high radiation and protective equipment, all of which may significantly degrade performance. The inspected object (with its shape, surface roughness, and accessibility) and the equipment are known to affect further the quality of the inspections.

When employed to detect and estimate the size of defects in thick metal components or the welds, NDT can be a demanding cognitive task. A successful result requires the application of knowledge about the inspected component, understanding of available NDT techniques, and interpretation of a large variety of signal characteristics. Reaching a correct decision depends on the accuracy of the information obtained and the effectiveness with which the information is interpreted and weighed (Harris & McCloskey, 1990).

Under regular circumstances, the task is manageable for an average qualified and experienced inspector. However, as the conditions get harder and the line between an acceptable and an unacceptable discontinuity becomes harder to draw, the demands of the task rise and sharper skill and judgment by the inspector are required (Dickens & Bray, 1994; Enkvist, Edland, & Svenson, 1999).

2.1.2. NDT reliability

NDT methods are not capable of finding all defects. Establishing whether and to what degree an NDT method can detect the critical defects is usually expressed in terms of reliability. Reliability of NDT generally refers to "*the degree that an NDT system is capable of achieving its purpose regarding detection, characterization and false calls*" (Taylor & Nockemann, 1999, p. 7).

Ideally, that what is reported after an NDT inspection in the inspection report should correspond to the true state of the component. In reality, these two sometimes differ due to a number of factors that can influence the NDT inspection results.

NDT reliability assessments focus primarily on the technical aspects of reliability, resulting in further advancements in technology. Supporting this statement is the fact that the reliability is typically expressed in terms of probability of detection - in short, the POD (Berens, 1989). With the help of POD curves reliability is expressed in terms of defect detection dependency on some selected parameter, such as the defect size, depth, orientation, etc. This method assumes that a property of the defect—most frequently the defect's size—is the most important determinant of whether this defect is going to be found. However, measuring the overall NDT performance using only the POD has been deemed unsuitable because, among other things, POD does not take the inspectors' ability to discriminate between defect and non-defect indications and the inspectors' decision criteria into account (Spanner Sr., 1986).

According to the *Modular Reliability Model*, developed by the NDT community during the first European-American Workshop on Reliability of NDT (Nockemann & Fortunko, 1997), NDT reliability depends not only on the intrinsic capability of the NDT system, i.e. its technical capability, but also on the application parameters, and the human factors. Application parameters refer to the factors reducing the capability of an NDT system (e.g. the material surface conditions or access to the component) and human factors to the factors furthermore reducing the system's capability or effectiveness. Human factors were more thoroughly defined as "*the mental and physical make of the individual, the individual's training [,] and experience, and*

the conditions under which the individual must operate that influence the ability of the NDE system to achieve its intended purpose" (Taylor & Nockemann, 1999, p. 8).

In light of new research illustrating the effects of organisation on the reliability (Bertovic, Gaal, Müller, & Fahlbruch, 2011; Gaal et al., 2009), the model was further expanded by Müller et al. (2013) by embedding the three influencing factors (intrinsic capability, application parameters, and human factors) into an organisational context (Figure 3).

Supporting this model is the classification of the factors having the biggest influence on reliability into 10 classes (Ali et al., 2012, p. 104):

- Arising from the technique itself, capability even in the best of circumstances;
- Setting up and calibrating the equipment;
- Poorly written, or absent procedures, and the inspector experiencing difficulty applying the procedure;
- Human factors, which enhance the opportunity for errors, such as fatigue due to long shifts or un-adjusted shift work;
- The aims of the inspection (is it a purpose designed inspection or a general purpose one, is it being used simply to satisfy regulatory requirements);
- The inspectability of the component (access, surface finish, etc.);
- Defect characteristics (especially if they differ from those the NDT is designed for);
- Management of the inspection (clear information, communication, breaks, commercial pressures, etc.);
- Data processing and classification (may be more appropriate in automated systems);
- Reporting (sometimes even critical defects might be found, correctly sentenced but unreported).

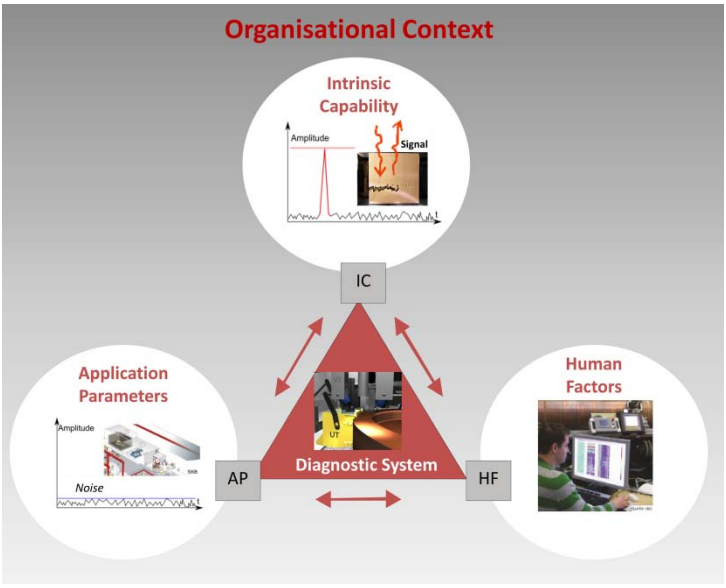


Figure 3: Modular reliability model (From "Paradigm Shift in the Holistic Evaluation of the Reliability of NDE Systems", by C. Müller et al., 2013, *Materials Testing*, 55(4), p. 264. Reprinted with permission of the Carl Hanser Verlag)

There is no doubt that inspectors play a vital role in an NDT inspection. The inspector is there to operate the equipment, interpret its signals, appropriately use the procedures, and do all that in frequently very unfriendly environment. Even after the equipment and the procedures are adequate, the largest variation comes from the inspecting personnel. When asked about the performance in the field, the study participants—all experienced inspectors—state that different inspectors frequently produce different results (e.g. Wheeler, Rankin, Spanner, Budalmente, & Taylor, 1986).

However, the consideration of human and organisational factors in the estimation of NDT reliability has not deserved sufficient attention. Whereas methods for measuring the equipment's capability have been excessively developing, the causes of human variability in NDT and ways to mitigate them are still in the background.

2.1.3. Types of errors in NDT

The variability in NDT is typically associated with differences between the inspectors in detecting and interpreting defects. This difference can lead to a defect not being detected or its size being underestimated or overestimated. The different types of errors will be addressed in this section.

2.1.3.1. Errors in defect detection

NDT is, above all, a signal detection task. In signal detection, the task is to determine whether a stimulus is present or not, i.e. the aim is to distinguish a signal from a background interference or noise (Swets, 1996). Thereby, two correct or two false responses to the stimulus, i.e. a defect in the material, can be made: one can correctly or falsely accept or reject a signal.

If there is a defect in the material reflecting a signal strong enough to be distinguishable from the background noise, the inspector will most probably detect it (true positive). However, other signals coming from the material, the surface of the component or the interference of the equipment might mask the actual signal and cause that signal to be missed (false negative).

Even when no defect is present, an inspector receives a number of signals from the interference of the equipment, from the material, or from the surface of the component (all part of the background noise). If interpreted correctly, one has made a correct rejection (true negative), i.e. one has rejected a supposed signal as being part of the background noise. However, if misinterpreted for a signal, the inspector will report a false alarm (false positive). Figure 4 shows all possible responses to an existing or not existing defect in the material.

Two of the four responses are the correct ones: the correct acceptance and the correct rejection (i.e. true positive and true negative). Of concern are the misses and the false alarms.

Misses and false alarms do not raise the same concern for the ultimate safety of the structure. Note that a miss is *critical* only when the defect is of a size that might raise concern of the structure's integrity. Whereas false alarms can lead to unnecessary repair or replacement of the component and—consequently—to financial cost, a missed defect can raise the possibility of structural failure and pose a threat to safety, economics, reputation, and the environment. It is in interest of every organisation to keep both at a minimum or, ideally, to completely avoid them. From the safety side, however, misses are significantly more important and of main concern in the reliability assessment.

		Presence in the material	
		Defect	No defect
Response	Defect	HIT True positive	FALSE ALARM False positive
	No defect	MISS False negative	CORRECT REJECTION True negative

Figure 4: Defect detection as a signal detection paradigm

A critical defect might inadvertently be left in a structure because of several reasons. First, an NDT method might not detect the defect. Second, once detected, the inspector may incorrectly sentence the defect (make an incorrect decision about the criticality of the defect). Some other examples include the inspector failing to report the defect, or the repair shop failing to correctly repair the component (Ali et al., 2012).

2.1.3.2. Errors in defect interpretation

Another issue of concern, apart from the detection, is the interpretation of defects, e.g. the assessment of the defect's size. Accurate sizing requires skill and experience. A statement about the size of the defect often leads to the assessment of the criticality of that defect.

It is well known that every repeated measurement produces different values, often normally distributed. I.e. if repeatedly measuring a defect of a specific size that defect can, on occasion, be undersized or oversized. Underestimating the size of the indication and, thereby, the criticality of the defect, can pose a threat to safety. Overestimating the defect's size will raise concern and—similarly to false alarms—may result in the dismissal of a component that is fit for purpose and, thus, lead to an unnecessary financial cost (e.g. Ali et al., 2012). Other errors include wrong positioning, orientation, and the wrong determination of the defect type.

In simple cases, where the signals to be found are big enough and not ambiguous, it can be assumed that the majority of the inspectors would be unified in their assessment of the signal. As the conditions become more difficult (e.g. the signals hardly distinguishable from the background, poor accessibility to the inspection area, difficult working environment, etc.) the results of different inspectors begin to vary.

2.2. Human performance in NDT

Perfect, flawless performance is impossible. Performance per se cannot be exclusively "good" or "bad", as it is not a static entity. Performance is a "*highly dynamic interactive process, between the individual and the context in which he operates*" (Enkvist et al., 1999, p. 13). As such, the human performance in complex organisations with high safety and reliability demand, such as the nuclear power industry or aviation, is determined by a number of factors, which have been a topic of a decades-long endeavour in the field of human factors.

The overall aim of human factors research and engineering is a more effective system performance (Harris & Chaney, 1969). This is achieved by identifying human performance problems; by applying human knowledge to the design of the system (maximising the effects of human capabilities, and minimising those related to human limitations); by reducing system costs (people, equipment and information); by developing and maintaining human resources (e.g. high morale and job satisfaction); and, finally, by utilising the gained knowledge to other similar situations that currently exist or will occur in the future. Human factors discipline is seen as an area where psychology and engineering intersect and is, as such, multidisciplinary in nature. Goals of human factors are to avoid the negative effects of the interaction between humans and technology, or to decrease them in order to increase the well-being of people, functioning of the systems, and safety (Badke-Schaub et al., 2012).

However, the role of human factors in various industrial applications was not always straightforward. According to Reason (1993), only after the technological revolution of the last century has provided us with reliable technological systems and solutions, has it become clear that the human operator plays a crucial role in safety. This was followed with a period, in which not only the technology or the human are seen as main error causes, but rather the interaction between different subsystems (socio-technical approach), e.g. between the technology and the operator, or the operator and the organisation. Wilpert & Fahlbruch (1998) expanded this view by stating that the modern complex, often conflicting, settings require not only the intra-organisational, but also the inter-organisational approach (taking relationships between organisations into account).

The approach to human factors *in NDT* has not undergone a similar development as its original discipline. The focus is still mainly on the inspector and the prevention of human errors. In NDT literature, the "human factor"—referring to the inspector—has been frequently identified as the main source of error or variability in the results, even though the influences of the working conditions and the inspection procedure are generally acknowledged. This can be observed in the definition of human factors in NDT as:

> the mental and physical make of the individual, the individual's training and experience, and the conditions under which the individual must operate that influence the ability of the NDE system to achieve its intended purpose (Taylor & Nockemann, 1999, p. 8).

In contrast, the contemporary understanding of safety and reliability in high-risk occupations and organisations consider a much broader perspective by including environmental and organisations factors. Moreover, they emphasise the importance of looking not only into the single elements of the socio-technical system, but also into the interaction between system components, i.e. the individual, the team, the technology, the organisation, and the environment (Badke-Schaub et al., 2012; Fahlbruch & Wilpert, 1999; Reason, 1997; Wilpert & Miller, 1999). Giesa and Timpe (2002) elaborate this by saying that environment, organization, technology and the individual are so interconnected that saying that a cause of a failure is either technology or the human does not do them justice.

That looking only into the person—rather than into the entire system—is a limited approach to ensuring safety and reliability was extensively elaborated in human error literature and in reference to prevention of organisational accidents (e.g. Reason, 1990, 1997; Woods, Dekker, Cook, Johannesen, & Sarter, 2013). It is also reflected in the widely accepted definition of human factors by the Health and Safety Executive (HSE):

> Human factors refer to environmental, organisational and job factors, and human and individual characteristics which influence behaviour at work in a way which can affect health and safety (HSE, 1999, p. 5).

This definition emphasises three relevant aspects that have an effect on people's health and safety-related behaviour:

- The job, i.e. matching the job to the person in terms of workplace and working environment design, and the individual's requirements for information and decision-making, as well as task and risk perception.
- The individual, i.e. individual's personal attitudes, skills, habits, and personalities, which—depending on the task demands—can turn into strengths or weaknesses and have an effect on task performance.
- The organisation, i.e. work patterns, the culture of the workplace, resources, communications, and leadership and so on.

By comparing the two definitions—the general one and the definition of human factors by the NDT community—it appears that the consideration of human factors in NDT is rather limited because of its primary focus on the inspector. Even though this may be a prevalent definition in the NDT circles, there have been various studies in the field taking into account the broader perspective of human factors, as coined by the HSE and the experts alike.

In the following sections, the motivation for addressing human factors in NDT, i.e. the variability in the inspection results, will be addressed. This will be followed by models of human performance in NDT and a literature review.

2.2.1. Variability in NDT

As mentioned in the introduction, the motivation for addressing human factors in NDT came from a number of inspections, during which significant variations in the individual performances were observed but could not be overcome by physical or engineering methods (e.g. Fücsök & Müller, 2000; Nockemann, Heidt, & Thomsen, 1991). For example, in their study of radiographic weld inspection and the evaluation of radiographic film images, Nockemann et al. (1991) found differences in performance (expressed in terms of Receiver Operating Characteristics, ROC [Swets, 1996]) between professional inspectors and scientists (as both were represented in the study), which they assigned to the difference in experience. However, even equally experienced individuals varied in their performance, reason for which remained unknown. The Nordtest NDE Programme (1976-1990) revealed differences among inspectors in accepting and rejecting a defect. The results of the two phases of the Programme for the Inspection of Steel Components (PISC)—PISC I (1976-1979) and PISC II (1981-1984)—furthermore highlighted the variability in human performance. This was illustrated by a large scatter in the inspection results and a tendency to either oversize or undersize defects (McGrath, 1999). The study performed by the Netherlands Institute of Welding (NIL) between 1991 and 1995 resulted in detection reliability of only 50% and substantial sizing deviations, which the authors attributed to manual testing. Thus, they suggested that mechanised NDT methods should be preferred, if high detection reliability is to be expected (McGrath, 1999).

These and other similar observations initiated a number of studies and research projects with the aim of determining causes of the variability. Considering the potential safety implications, most of the research had been carried out in the nuclear industry (followed by aviation, and substantially less accounts can be found related to oil and gas, and railway). The specific conditions encountered in the nuclear industry can make the NDT task especially demanding, which is why investigating human factors in this application received more attention, e.g.:

- The effects of an unreliable inspection can be catastrophic, i.e. leakage of highly radioactive nuclear waste into the environment.
- The components to be inspected have complicated geometries, and consist of materials and welds often difficult to inspect (e.g. austenitic materials).
- Inspections are carried out under unfriendly working conditions, i.e. radiation, wearing of the protective equipment, humidity, surrounding noise, heat originating from the material being inspected and the surroundings, time pressure caused by the heat and economic pressure, poor accessibility to the testing object, and so on.

Task complexity and the working conditions, typical for this context, but also potential inadequate quality control and maintenance practices, and other factors can shape human performance in unwanted ways, which is why their identification is an important step in understanding human behaviour in this context. The so-called *performance shaping factors (PSF)* can be internal or external to the person. The internal PSFs include all those characteristics of the person that influence this performance, e.g. skills, motivation, and the expectations. The external include the work environment, including the equipment design, and the procedures. The work environment, which places high demands on the operator and does not match the operator's capabilities and limitations, can cause psychological and physiological stressors, considered as one of the most influential PSFs. A good match between the internal and the external PSFs will lead to a more reliable and optimal performance. On the contrary, a mismatch will lead to disruptive stress and suboptimal performance (Swain & Guttmann, 1983).

2.2.2. Models of human performance in NDT

Since the first studies in this field emerged, the scientists have attempted to understand the abundance of factors affecting the NDT inspectors in their work by setting up theoretical conceptual models based on observations and the understanding of factors affecting human performance.

Figure 5 illustrates the model of mechanised UT/ISI man-machine system developed by Spanner Sr. (1986).

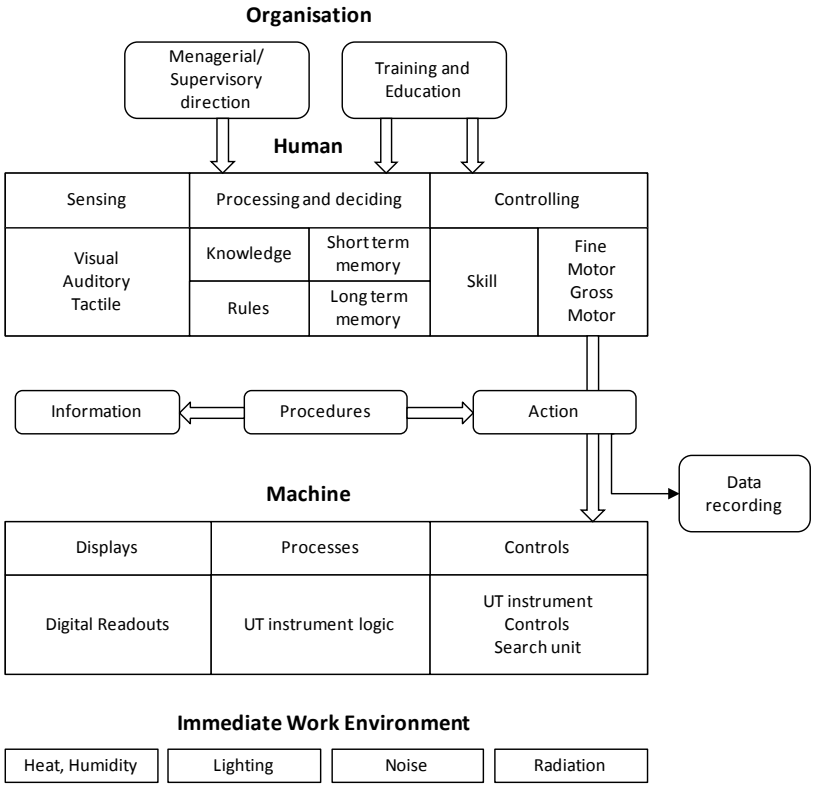


Figure 5: Model of UT/ISI man-machine system (From "Human Reliability Impact on In-Service Inspection", by J. C. Spanner Sr., 1986, in D. Stahl, ed. *Proceedings of the 8th International Conference on NDE in the Nuclear Industry*. Orlando, FL: American Society for Metals, vol. 21(19), p. 91. Reprinted with permission of the American Society for Metals.)

In this model, the mechanised system is performing the sensing, information processing, decision-making, and action functions. Training, education, and management are seen as inputs into the model. The inspector's task consists of signal interpretation, which is treated as a signal processing, pattern recognition, and a decision making process. Important outcome of this model is that the performance is seen as a result of interaction between the inspector and the technology in the context influenced by the management, the procedures, and the immediate working environment.

Two years later, Harris (1988) described human performance as an information-action-feedback loop influenced by a set of performance shaping factors (Figure 6). Based on this model, effective performance can be expected only when sufficient feedback is provided between an input of information through some sensory channel (visual, auditory, tactile, etc.) and an execution of the action in a form of some motor activity (manual, speech, etc.). This feedback must be complete, relevant, and timely appropriate. The nature of the task, the equipment, and the procedures affect this relationship internally, and the performance shaping factors externally.

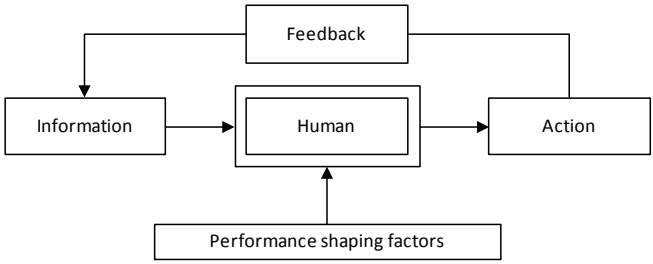


Figure 6: Model of human performance (From "Human performance in NDE inspections and functional tests (EPRI report NP-6052)", by D. Harris, 1988, Santa Barbara, CA: Electric Power Research Institute (EPRI), p. 2-2. Reprinted with permission of the Electric Power Research Institute)

In the conceptual framework of Behravesh, Karimi, and Ford (1989) the inspectors' capability of finding defects is influenced by a set of internal (personal) and external (environmental) factors. In line with the systems approach, poor performance is not equated with personal deficiency, neither with the situation alone, but is seen rather as a result of "*behaviour in a person-environmental system characterized by constant change in both internal and external (situational) requirements"* (p. 2236). Human factors can be identified by a theoretical "top-down" approach and by an empirical "bottom up" approach. In the first, effective performance is seen as a product of skilful, motivated person interacting with a responsive environment. Inspector's poor performance can be primarily understood as a function of contextual (physical and social environment) and motivational factors. Skill, though important, can be acquired through training. Performance can assume meaning only when it is evaluated within the context in which it occurs. In a complex environment (such as that of a nuclear power plant) motivation, decision-making and problem-solving skills, attention, and emotional resources, physical and mental stress, and feedback may determine the success of the performance. According to the latter, i.e. the empirical "bottom-up" approach, effective performance is defined by people's conception of a competent worker and/or a productive work episode.

The variations of internal and external factors affecting NDT reliability have been suggested by other researchers as well. According to Dickens & Bray (1994) inspection reliability is affected by the operator, engineering, and stochastic factors (see Figure 7).

Engineering decisions—related to the duration of the inspection, the working environment, or the procedures—create conditions extrinsic to the operator that can have a direct effect on the inspection reliability, as well as on the inspectors themselves. On the other hand, factors intrinsic to the operator, such as training, experience, motivation, and expectations, affect the operator and indirectly the inspection reliability.

Bertovic, Gaal, Müller, & Fahlbruch (2011) conceptualised possible influences on the manual ultrasonic in-service inspection (UT/ISI) performance in nuclear power plants (Figure 8).

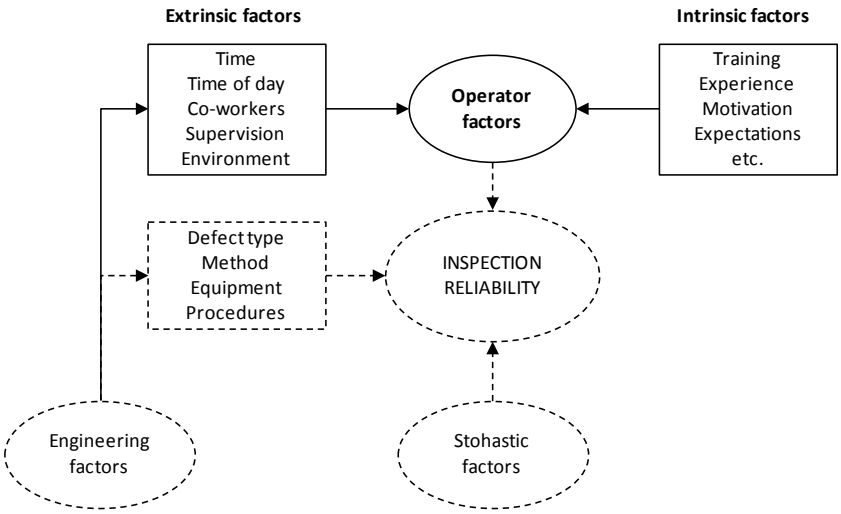


Figure 7: Extrinsic and intrinsic influences on human performance within the depiction of primary influences on inspection reliability (From "Human performance considerations in nondestructive testing" by J. Dickens, and D. Bray, 1994, *Materials Evaluation*, 52(9), p. 1040. Reprinted with permission of the American Society for Nondestructive Testing, Inc.)

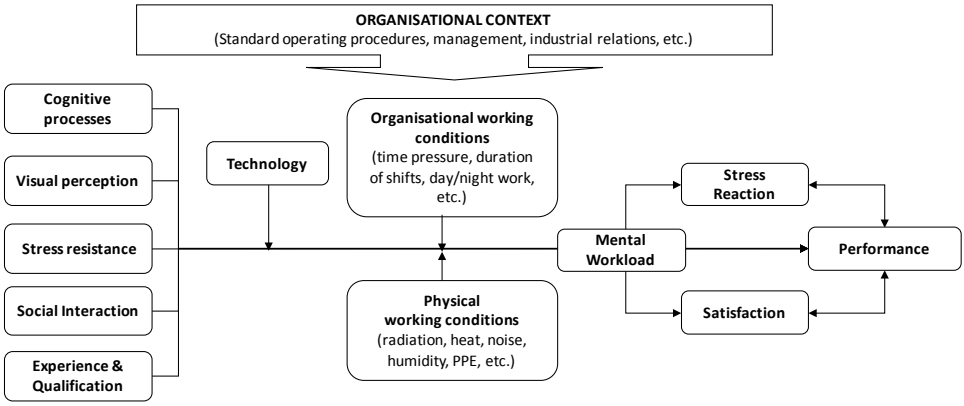


Figure 8: Combination of influences on the manual ultrasonic in-service inspection performance in nuclear power plants (From "Investigating human factors in manual ultrasonic testing: testing the human factors model" by M. Bertovic, M. Gaal, C. Müller and B. Fahlbruch, 2011, *Insight*, 53(12), p. 673. Reprinted with permission of the British Institute of Non-Destructive Testing.)

According to this conceptual framework, the quality of manual UT/ISI inspection performance is influenced by a set of internal personal predispositions (cognitive, perceptual, social, personality, knowledge, and skills), by a set of external influences (organisational and physical working environment), by technology, and by the organisation. The working environment, coupled with the equipment one uses has a moderating effect on the performance. For example, difficult working conditions, e.g. high radiation, heat, and high time pressure, could give rise to mental workload and the arousal, resulting in a decreased

inspection quality. Under optimal conditions, on the other hand, mental workload would remain constant and could give rise to work satisfaction, and through that positively affect the inspection performance. The organisational context, including both the intra- and inter-organisational factors (management practices, standard operating procedures, industrial relations, etc.), is given special importance, as the entire model of influences is embedded into it.

Unlike other models, which remain only theory-based, parts of this framework were empirically tested. The results revealed time pressure (an example of organisational working conditions) and mental workload as significant moderators of the NDT inspection performance. Moreover, the study highlighted a number of organisational factors of noteworthy effect on the inspection performance (Bertovic et al., 2011; Gaal et al., 2009).

In summary, the presented models jointly explain human performance in NDT as a function of internal inspectors' characteristics and predispositions, working environment, technology, and the organisation – a view that is in line with the trends in general human factors research.

Even though not designed exclusively to empirically verify the aforementioned models, a number of studies has been carried out to a) identify sources of variability in the inspection results and b) to find ways to mitigate the effects of that variability on NDT reliability. In the following section, an overview of those studies will be given.

2.2.3. Human factors in NDT reliability: literature review

In comparison to other fields, human factors in NDT have been a rather poorly investigated field. In the past 30 years, only a handful of institutes engaged into the research in this field.

The focus of study can be divided into six areas: the individual, group/team, the working conditions, the organisation, inspection procedure, and technology. Thus, the review will be presented accordingly.

2.2.3.1. Individual

A frequent discussion in NDT is the distinction between a "good" and a "bad" inspector. Which qualities make out a "good" inspector is a question researchers in this field have dealt with theoretically and empirically. Typically, that description refers to abilities, attitudes, skills, personality traits, cognitive strategies, and the experience an inspector should have that is expected to lead to an optimal NDT performance and, consequently, to higher NDT reliability. Table 1 summarises the main findings and conclusions of the studies concerned with the individual.

Table 1: Theoretical (T), experimental (E), and survey (S) considerations of the effect of individual traits and differences on NDT performance

Performance shaping factors	Conclusions	Reference	Type
Knowledge and understanding of NDT theory	Associated with better performance	Behravesh et al., 1989; Wheeler et al., 1986	T
Abilities			
Mechanical comprehension		Bell et al., 2012; McGrath, 2008	E
Personality			
Conscientiousness		Behravesh et al., 1989	T
Cautiousness		Bell et al., 2012; McGrath, 2008	E
Original thinking			E
Attitudes			
Interpretative habit of action (Preference for the routine)	Associated with high theoretical knowledge and long practical experience	Norros & Kettunen, 1998; Norros, 1998	S
Procedural habit of action (Preference for adherence to the procedure)	Associated with long experience with less theoretical background		S
Adherence to the inspection procedure	Associated with better performance	McGrath, 2008	T
Experience	Negative correlation with the error score	McGrath, 2008	E
	Higher inspection precision	Bertovic, et al., 2011	E
	Has not shown to increase performance	Enkvist et al., 2001c; Enkvist, 2003; Spanner & Harris, 1999; Wheeler et al., 1986	E
Cognitive strategies Development and testing of explicit hypotheses Avoidance of reaching a conclusion early in the inspection process, before all available information has been obtained and considered Application of knowledge during the inspection by putting it in the if-then logic Avoidance of the arbitrary elimination of information from consideration during the process of reaching an inspection conclusion	Support effectiveness (can be applied in training, or transformed into a checklist to be followed during the inspection)	Harris & McCloskey, 1990; Harris, 1990, 1992	E
Capacity			
Mental workload	Increases variability and negatively affects precision	Bertovic, et al., 2011	E
Attention Motivation Self-efficacy Self-confidence, task orientation, consistency	Associated with better performance	Behravesh et al., 1989; McGrath, 2008; Wheeler et al., 1986	T
Stress resistance/tolerance of environmental conditions	Fosters inspection precision/ Associated with better performance	Bertovic, et al., 2011; Wheeler et al., 1986	T

The above-mentioned suggests that individual differences have puzzled the NDT community and efforts have been made to identify those individual traits that can foster optimal inspection performance. E.g. suggestions have been made to use the acquired knowledge in the selection of the inspecting personnel as well as to develop those traits predictive of good performance. However, there have been only a few reports showing how their implementation shapes performance. They include implementation of the cognitive strategies into a checklist to be used during the inspection (Harris & McCloskey, 1990; Harris, 1990, 1992) or using aptitude and personality tests or in a) tailoring a better training course adapted to the needs and the capabilities of the personnel (Bell et al., 2012) and in b) the selection of the inspecting personnel (Spanner & Harris, 1999; Spanner, 1999). Nonetheless, they are still scarcely, if at all, implemented in the practice.

Contradictory results were obtained about experience, often thought of as a good predictor of NDT performance. Typically, the more experience an inspector has, the better in NDT he is assumed to be. For example, McGrath (2008) established a negative correlation between years of manual ultrasonic experience and the overall ultrasonic *error* score indicating that good performance is related to field experience. Some studies contradict this assumption, by providing no evidence that experience correlates with the performance (Enkvist, Edland, & Svenson, 2001b; Enkvist, 2003; Spanner, 1999; Wheeler et al., 1986). According to Enkvist et al. (1999), the danger lurking behind a belief in the experience is an exaggerated feeling of self-competence that may lead the inspectors to disagree with the guidelines of the procedure. This is not to say that experience is not a significant predictor of good NDT performance, but rather that relying *mainly* on experience might not be the best approach to assuring high standards of NDT reliability.

2.2.3.2. Group/team

Working in a team is generally believed to improve performance. For example, Dickens & Bray (1994, p. 1040) suggested that it is "*less likely for two people to miss a potential discontinuity*". One third of inspectors in the study of Wheeler et al. (1986) reported they are likely to find indications previously not found by another inspector. The social influence is seen as beneficial in aiding in the decision-making process by "*confirming or disputing signals that have been interpreted*" (Taylor et al., 1989, p. 342), which is why human redundancy is often suggested as a method giving rise to NDT reliability (McGrath, 2008).

The only found study examining the benefits of adding more inspectors to improve inspection accuracy was reported by Harris & Chaney (1969). They had ten experienced inspectors independently carry out the inspection task resulting in an observable improvement in the accuracy for the first six inspectors, whereas the next four did little to improve performance of detecting critical defects. In the case of less-critical defects, the accuracy continued to improve as additional independent inspections were added, however at lower overall detection rate than the critical ones. The authors suggest that under certain conditions the team approach to increasing inspection accuracy may be useful.

Plenty has been achieved in the qualification and training of the inspection personnel, in the technology, as well as in the way inspections are carried out since Harris & Chaney's study to futher foster these findings. The topic of team and the group influences possibly affecting the reliability of the NDT inspections is insufficiently investigated and remains not well understood.

2.2.3.3. Working conditions

The influence of the environmental conditions on the inspection performance has been a topic of a few empirical studies concerned primarily with manual UT during in service inspections in Table 2.

Table 2: Experimental studies (E) on the influences of working conditions on the manual UT performance

Performance shaping factors	Conclusions	Reference	Type
Environmental conditions			
Temperature, humidity, noise level	No significant difference in performance between optimal and suboptimal conditions	Murgatroyd et al., 1994	E
Noise	A certain amount of stress is associated with better performance	Enkvist et al., 2001a, 2001b	E
Organisational working conditions			
Shift length, shift times, breaks, number of working days, number of rest days	No significant difference in performance between optimal and suboptimal conditions	Murgatroyd et al., 1994	E
Time pressure	Increases variability and decreases inspection precision	Bertovic, et al., 2011	E
	A certain amount of stress could increase the performance in a familiar task	Enkvist et al., 2001a, 2001c	E

The experimental results of Murgatroyd et al. (1994) seem to show that suboptimal working conditions do not significantly affect performance. However, as elaborated by Pond, Donohoo, and Harris, Jr (1998) this happened because the simulated subotpimal conditions were not difficult enough and not representative of the practice. To support that, they referred to a number of studies and reports showing that a) the conditions in NPPs are in fact more difficult than simulated in the study and that b) decrements in performance could be expected only under conditions more difficult than the ones measured. Pond et al. suggested that more attention should be given to developing strategies of mitigating the performance decrements already established to be influenced by the environment, rather than to further investigating them. Further research should be invested into those factors the industry has the ability to control (as opposed to heat and noise), such as the job design or the managerial practices.

A shift of attention from the working conditions towards organisation has been a topic of several, mainly theoretical, considerations, as will be elaborated in the following section.

2.2.3.4. Organisation

Organisation, in line with the modern safety research, can be singled out as one of the most salient influences on reliability and safety. Already in 1986, J. C. Spanner Sr. pointed out the importance of organisation, by considering management as an important input in the UT/ISI man-machine system (see Figure 5). Continuing efforts have been invested into raising the awareness of the superior influence of the organisation on the reliability of NDT. Table 3 summarises those efforts.

Table 3: Theoretical (T) and survey (S) considerations of the effect of organisation on the reliability of NDT

Performance shaping factors	Conclusions	Reference	Type
Organisational climate	Some of the major contributors to NDT reliability	Spanner Sr., 1986; Taylor et al., 1989; Wheeler et al., 1986	S
The support or frustration attributed to the organisation			
Faith and trust placed in individuals and teams by the employer or contracting organisations			
Extent and quality of supervision		Behravesh et al., 1989; Spanner Sr., 1986	S
Management		Behravesh et al., 1989; Spanner Sr., 1986	S
Organisational context			
Regulators, utilities, plants and contractors		Bertovic, et al., 2011; Harris, 1988	T
Internal—the business process (financial agreement between the customer and the service provider), the information process (exchange of information between both parties), and the delivery process (the delivery of the service, i.e. the NDT inspection)	Competitive market (demands to keep production and maintenance costs as low as possible) and technical rules (in terms of use of different standards, often required by the utility), are seen as having the largest influences on NDT reliability	Holstein, Bertovic, Kanzler, & Müller, 2014	T/S
External—safety culture, social/ ethical culture, market/financial frame, regulatory requirements and technical rules			
Feedback	Situations in which informational and supportive feedback are not available may lead to inspectors developing negative beliefs about their capabilities and opportunities of exercising them	Behravesh et al., 1989	S
	Lack of feedback can lead to an impression that the utilities do not take the effort of the inspectors seriously and do nothing to adjust their expectations	Wheeler et al., 1986	S
Preparation for the inspection	Some of the major contributors to NDT reliability		
In terms of access, safety and plant surface condition		McGrath, 2008	T
In terms of instructing the personnel about the inspection and planning of the inspection		Bertovic, et al., 2011	T
Suitable documentation (risk assessments, inspection procedures, standards, acceptance standards, access and cleaning requirements, drawings and photos and equipment inventory)		Bertovic, et al., 2011; McGrath, 2008	T
Adequate time for the inspection		Bertovic, et al., 2011; McGrath, 2008	E/T

All of these instances reflect predominantly theoretical, rather than empirical, considerations of the organisational influence on reliability of NDT. Supervisory and managerial practices within an organisation, appropriate planning, execution of the inspections, and the inter-organisational context are hypothesised to have a significant impact on NDT reliability. This was acknowledged by the community by updating the widely-accepted NDT reliability model (Müller et al., 2013; see section 2.1.2 and Figure 3) so that it includes organisational context, under which all other influencing factors (intrinsic capability, application factors, and human factors) are embedded.

What remains lacking is establishing direct causal relationships between the influences and the performance, and above all, strategies for improvement and their practical implementation.

2.2.3.5. Inspection procedure

Inspection procedures are frequently mentioned as having a substantial influence on the performance (e.g. Enkvist et al., 1999; McGrath, Wheeler, & Bainbridge, 2009; McGrath, 2008). This is not surprising, since—apart from the testing equipment—the inspection procedure is the most valuable tool in the hands of an inspector. If important information in the procedure is missing or if the information is misunderstood, this can lead to errors in carrying out the task. Moreover, inappropriate procedures may lead to violations.

Several attempts have been made to understand the extent of the influence of the inspection procedure on the results as well as to improve the existing procedures and NDT instructions. Table 4 summarises the main topics and their findings.

Table 4: Theoretical (T), experimental (E), and survey (S) considerations of the quality of the existing inspection procedures and their improvement possibilities

Performance shaping factors	Conclusions	Reference	Type
Improved inspection procedures and protocols	Lead to higher satisfaction of the inspectors	Bertovic, et al., 2011	T
Procedural deficiencies	Associated with higher event and incident rate	Bento, 2002	S
Inspection procedure Usability	Leads to higher effectiveness, efficiency and satisfaction	Bertovic & Ronneteg, 2014	E
Reading and applying	Not always done as it should	McGrath, 2008	T
Improvement	To be achieved using task analysis	Harris, 1988	T
	To be achieved using a user-centred design and the application of human factors principles	McGrath, 2008	T

The improvement of the inspection procedures has most frequently been dealt with by continuously improving its content. When approved, the procedure is expected to be used by the inspector, but neither verification nor its adequacy is questioned, let alone its usability. The human factors analysis of the inspection procedure presented by McGrath (2008), including guidelines on how to write better procedures, e.g. with reference to length, structure, and consistency, provided the first step in addressing the importance of not only accurate procedures, but also understandable and usable ones.

Overall, the research on this topic suggests that inspection procedures require careful attention with respect to comprehensibility and usability, and a human-centred approach to the design.

2.2.3.6. Technology

There is no doubt that the technology can affect performance. Interaction between humans and technical systems is in the core of the socio-technical systems approach and of the human factors field in general. On the one hand, the equipment, with various controls and displays, might be difficult to use or be poorly designed. On the other hand, the interaction between the inspectors and the equipment might lead to errors.

Frequently associated with errors are manual forms of NDT, i.e. most commonly, the manual UT. This is where the variations between the inspectors have been often observed, and this is why those methods received the majority of attention from the human factors perspective. The attempts to understand the underlying causes of that variability and finding ways to decrease it have been described in the previous sections.

A frequently suggested method to decrease, if not eliminate, the chance for human error has been to replace manual with mechanised methods (e.g. Forsyth, Komorowski, Gould, & Marincak, 1999; Herr & Marsh, 1978; Liao & Li, 1998; Lingvall & Stepinski, 2000; Shafeek, Gadelmawla, Abdel-Shafy, & Elewa, 2004). In the discussion of mechanised over manual NDT, mechanised is almost exclusively seen as the front-runner in achieving higher reliability. Other advantages include an increase in data consistency, inspection speed, and the probability of detection; the ability to store data and re-evaluate at a later point in time; the possibility of computerised or completely automated evaluation of data; the ability to augment scans from different directions; and so on (Carvalho, Rebello, Souza, Sagrilo, & Soares, 2008).

Even though these arguments are not open to debate, the overwhelming focus on technology in this matter has led to neglecting some human performance issues. Even when fully-automated systems are used, human inspectors will still most probably be sent to control (Dickens & Bray, 1994). Selecting appropriate standards, preparing the equipment and the procedures, calibrating the equipment, analysing data, applying acceptance criteria and reporting findings are still carried out by the inspector. In mechanised testing, the role of inspectors is even more prevalent, especially as the result of the inspection relies on the evaluation of data conducted solely by people. Enkvist, Edland, & Svenson (1999) report that, even when the inspection is to a certain extent automated, the largest source of performance variation can be found in the inspector. According to them, the most significant task is still the evaluation of collected indications, not only their detection.

That variability is not something exclusively assigned to manual NDT, was to some extent shown by Gaal et al. (2009). Even though they admit their conclusions are limited since they are based on a sample that is too small, they observed variability in results between three mechanised teams, indicating that mechanised NDT might not be as reliable as thought of.

The major problem with respect to technology seems to be the transition from manual into the application of automated and semi-automated (mechanised) testing methods. In spite of the automaton's superiority over manual methods in terms of reliability and efficiency, the aforementioned considerations suggest that mechanised methods are not without problems and variability between inspection teams can still be observed. However, empirical evidence of the influence of the automation use on the inspection performance is still missing.

2.2.4. Main conclusions of the literature review

The literature review reveals a number of potential causes of the variability in the inspection results arising from the individual, the group, the working conditions, the organisation, the inspection procedure, and the technology. In the following, the major results will be summarised, and the missing research identified.

In spite of the variety of the presented investigated influencing factors, an overwhelming majority has been dedicated to the individual differences and manual inspection methods (mainly manual ultrasonic testing during in-service inspections, which appears to be one of the most demanding tasks). This conclusion is based not only on the quantity of presented studies, but also on the fact that the majority of considerations of *other* influencing factors were primarily theoretical, whereas empirical investigations are scarce. Mechanised testing—due to its higher perceived reliability—is considered a good method to decrease or eliminate human error as it replaces the "faulty" manual inspector with more reliable automated equipment.

Common ways to tackle with human factors-related problems and the varying performance in NDT include advances in the personnel qualification (e.g. JRC-IE, 2010), improvement of the equipment by automating parts of the task (Wall, Burch, & Lilley, 2009), changing of the inspection procedures (Enkvist, 2003) and reliance on a knowledgeable NDE practitioner and his experience (e.g. Annis & Gandossi, 2011; Fücsök, Müller, & Scharmach, 2002). The human factors' specialists, in contrast, see potential for improvement if human factors principles are applied and the knowledge gained from the studies is appropriately implemented.

There have been almost no studies concerning group influences and the interaction with automation technology. Considering that the typical measures to increase reliability and safety include automation and, often, human redundancy; the human-automation interaction and group effects have become some of the most salient characteristics of modern complex systems (Manzey, Boehme, & Schöbel, 2013; Manzey, 2012). Even though automation is designed with the aim of increased efficiency, improved safety, lower operator workload, etc.; improper use, poor design, or inadequate training for automation can be counter-effective to its aims (Parasuraman & Manzey, 2010). Similar counter-effects can be expected when the principles of technical redundancy are transferred to social systems (Clarke, 2005). In the human factors field, automation and group processes elicited plenty of research. It is clear that more attention to these topics is required in the NDT field.

This suggests that the attempts at identifying the causes of variability in NDT performance may have been rather narrow and that in looking for failure causes one should look beyond the inspector and into interactions of the inspector with other systems. Hence, a coherent picture of the human and organizational factors influencing NDT inspections is still missing.

The largest threat to NDT reliability in the field is the assumption of the majority of the researchers (e.g. Behravesh et al., 1989; Murgatroyd, 1992; Spanner Sr., 1986; Wall et al., 2009; Wall, 2013; Wheeler et al., 1986), that the variability observed in the experimental settings is even larger in the field. This is a worrisome fact. For that reason, human factors research needs to strive towards field research in conditions as close to reality as possible. This often presents with difficulty with respect to methodology.

The typical methods employed in studies include interviews, questionnaires, think aloud method, and, to a somewhat lesser extent, experiments. One of the major problems these studies jointly encounter is a rather small participant sample, which makes the conclusions of

the studies weaker. A further problem is the rare use of statistics (the studies are often analysed descriptively, and conclusions based on observations). There is a need for engineers working together with psychologists in designing experiments and interpreting the data and a need for a proper experimental design to obtain meaningful results (Enkvist et al., 1999; Harris, 1992).

2.3. Challenges and the aims of the study

In general terms, this chapter thus far provided the information necessary for the basic understanding of NDT (section 2.1) and presented an overview of existing literature of the study of human factors in this field (section 0). The latter concluded with identifying gaps in the existing research that lays the foundation for the work that will be presented in this dissertation.

However, in spite of a number of studies carried out to explain the variability in the inspection results, the NDT community remains unaware or unaffected by its findings. In the following, some of the practical and context-related challenges will be outlined, leading to the aims and objectives of the study.

2.3.1. Practical challenges of the study of human factors in NDT

During the European-American workshop on reliability of NDT—fifth in a row of international gatherings of experts from the field—human factors were brought again in focus. Some of the main discussed issues refer to the following: A communicational gap "*between what is known about human behaviour under difficult working conditions in psychology and what is known by the engineers*" and the gap "*between the utilities and the service providers, causing problems in the transfer of knowledge and, hence, posing a difficulty to implement the findings in the field*" (Bertovic et al., 2014, p. 604). Hence, the currently recognised problems in this field are related to three gaps: the gap in knowledge, the gap in communication, and the gap in implementation. These gaps will be discussed in the context of the conclusions from the literature review and observations made in the field.

2.3.1.1. Gap in knowledge

The first gap, i.e. gap in knowledge, can be derived from the literature review. The predominant orientation towards the human inspectors as the largest sources of variability in NDT indicates that research into the effects of other influencing factors is still missing. Especially with regard to group influences and interaction with automation technology.

The application of mechanised testing in NDT is still in its beginnings. Up to this date there have been no studies reporting problems that can arise from its application. There is a need for the identification of risks and their prevention. This is of special importance in industries with high reliability demand, such as the nuclear industry, in which the use of mechanised testing is increasing.

2.3.1.2. Gap in communication

It has become clear that NDT and psychology do not always communicate well, as the practitioners and scientists are not acquainted with a vast amount of research on human performance. Hence, one of the still active discussion points in NDT is the topic of vigilance and the effects of working conditions on the performance (Bertovic et al., 2014), topics for

which a body of research already exists (e.g. Echeverria, Barnes, & Bittner, 1991; Pond et al., 1998).

When looking into the studies on human factors in NDT, the influences of organisation and a mismatch in human-machine interaction have been recognised early. However, NDT technical literature and the definition of human factors conceptualised by the NDT community suggest that the NDT field is unaware of the extent of the existing literature and its findings, which indicates a large gap between practitioners and the human factors field. The inspection personnel is still frequently referred to in the NDT literature as "*the weakest link in the quality chain of NDT*" (e.g. Trampus, 2013, p. 9) and human factors are frequently considered in reference to "*the operator and* [his] *skills, attentiveness, mental attitude,* [and] *health* (Gandossi & Annis, 2010, p. 59).

In the presence of contradicting evidence from the field of human factors, the belief that the negative effects of human factors can be decreased if parts of the task are replaced by automation—without considering new potential risks—suggests a need for a bridge between NDT and psychology. Whereas the risks associated with automation, for example, are continuously discussed in the field of human factors, this topic has not yet been addressed by the NDT community.

2.3.1.3. Gap in implementation

The conducted studies presented with numerous suggestions how to improve NDT reliability. However, not much of the acquired knowledge had been implemented in the practice.

This is best outlined in a following example: The group of scientists behind the Programme for the Assessment of NDT in Industry, PANI 3 (McGrath, 2008) summarised their suggestions in a small booklet and handed it out to the inspectors on site, and promoted their findings in a myriad of scientific and practitioners' circles. Five years after the study, Carter & McGrath (2013) presented a paper titled "We Know How To Improve Inspection Reliability - Why Don't We Do It?" concluding that in spite of their efforts, and of all the other reliability studies worldwide, the NDT community still remains unaware of the findings, which are seldom, if at all, implemented.

2.3.2. Context-related challenges

The studies, which will be outlined in this dissertation, were conducted on an example of mechanised NDT methods that are currently under development to be used for the final disposal of the spent nuclear fuel (See "Digression" for a short description of the Swedish and Finnish spent nuclear fuel management programme).

The specific challenge of this field application is that the NDT inspections of the canister will be carried out only once—the components after being manufactured and the weld after the spent nuclear fuel has been sealed into the canister—with no possibility of inspection while in service and potential repair. This presents a challenge for NDT and high demands are placed on the quality and reliability of NDT inspections. To ensure that the canisters are fit for this purpose, an extensive NDT inspection programme is being developed. Processes such as choosing the right NDT methods and techniques, appropriate and reliable equipment, customising training, qualifying the personnel, organising the inspection process, and developing inspection procedures and instructions are under ongoing development. Thus, they represent a challenge for the developers, as well as give room for human factors design.

Another challenge related to this application is that a prospective approach to risks—in a process that is yet not fully known—needs to be undertaken.

DIGRESSION: A description of the Swedish and Finnish approach to the final spent nuclear fuel disposal

Responsible use of nuclear energy includes maintaining of the existing nuclear power plants, but also developing solutions for the storage of its production by-products, i.e. the spent nuclear fuel and the produced waste. Long-term management of spent nuclear fuel is one of the most critical issues affecting the acceptance of nuclear power. Geological disposal in a deep repository is up to date the only available approach that can guarantee the required level of safety, given that the repository is properly implemented at a well chosen site (IAEA, 2011).

The safety principle of the most successful national geological disposal programmes, i.e. in Finland and Sweden, is based on a three-barrier concept (known as the KBS-3). This concept entails encapsulating spent nuclear fuel in copper canisters and depositing them in the bedrock at a depth of about 400-500 meters, and additionally protecting them with a buffer of bentonite clay intended to protect against corrosion and movements in the rock (see Figure 9 for the illustration of the three-barrier concept). Stored like this, the canisters should withstand untouched for the next 100,000 years (the extent of the safety assessment) up to one million years, leaving the radioactivity in the spent fuel to decline naturally through the decay of the radioisotopes in it (Posiva Oy, 2010; SKB, 2008). The operation start of the repository is scheduled for the year 2022 in Finland (Posiva Oy, 2015) and 2030 in Sweden (SKB, 2013).

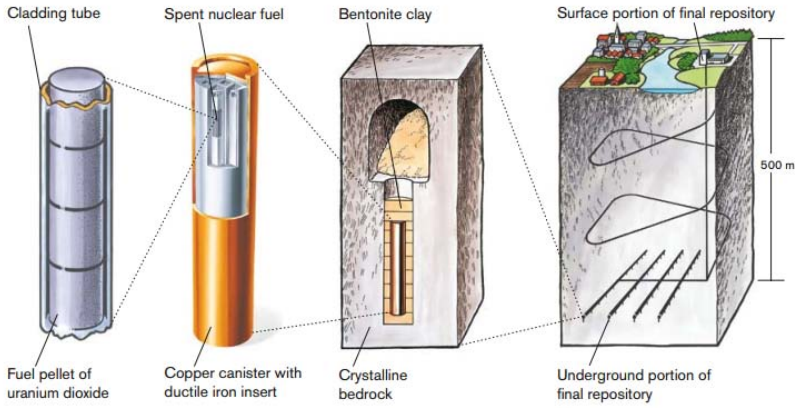


Figure 9: Illustration of the three-barrier concept for the disposal of the spent nuclear fuel (Source: SKB; with permission)

The most important safety barrier for the spent nuclear fuel is the canister. Thus, substantial effort is being invested into the development of the canister and into its mechanical properties to ensure that it can withstand anticipated loads caused by, for example, potential earthquakes, and even upcoming ice ages.

The canister, consisting of a copper shell (tube, lid, and bottom), and a cast iron insert is in the centre of interest for the NDT. Figure 10 depicts the canister's parts.

NDT is used to ensure that no critical defects are present in the materials and welds because such defects could lead to the leakage of radionuclides from the spent nuclear fuel into the environment. Considering the high reliability requirement and the fact that—after being filled with the spent fuel—the canister will be highly radioactive, the use of mechanised NDT methods is foreseen. Based on the current developments, this inspection is planned to be achieved by means of up to four NDT complementary methods: Ultrasonic (UT) and radiographic testing

(RT), to complement each other in search for defects in the material's volume, and eddy-current (ET) and visual testing (VT) for the defects near or at the surface[3] (Pitkänen, 2013; SKB, 2013).

Figure 10: Canister parts: the copper tube and the cast iron insert with their respective lids and bottoms (Source: Posiva Oy; with permission)

2.3.3. Aims and objectives of the study

The ultimate goal of consideration of human factors in NDT is to decrease variation in the inspection results, reduce the risk of failure, and, therewith, increase the reliability of NDT.

The overall aim of this study was to face some of the current challenges of the NDT field and, for the first time, explore risks associated with mechanised NDT and find ways of mitigating their effects on the inspection performance.

The objectives, by means of which this aim plans to be achieved, are as follows:

- Identify and analyse potential risks in mechanised NDT.
- Devise measures against the identified risks.
- Critically address the preventive measures with respect to new potential risks.
- Suggest ways for the implementation of the preventive measures.

By accomplishing these objectives, the gap in knowledge will be addressed. The gap in communication shall be approached by using approaches and well-known theories from psychology and applying them in the field of NDT. The communication can be strengthened by involving the NDT organisations in the accomplishment of the work, which—in combination with the suggested preventive measures—can foster implementation of the findings into the field. The context-related challenge will be overcome by using a prospective approach to the risk assessment, by providing with missing knowledge about potential performance-degrading influences, and by offering suggestions for the improvement of the reliability in the field.

[3] Eddy current testing is used to look for defects at or near the surface, and visual testing for defects at the surface. Both methods complement each other in search for defects not in the component's volume and are partly redundant.

3. Empirical Study 1: Assessing and treating risks in mechanised NDT

"Clearly, 100% safety (zero risk) is not achievable, but safety has definitely improved […] due in part to the detection of vulnerabilities by formal risk analysis"

Sheridan, 2008, p. 418

NDT has to provide reliable results. Only that way it can achieve its goals and serve as a contributor to assuring safe operation of complex organisations with high reliability and safety demand. The potential variability in the inspection results observed in manual NDT presents a risk that NDT may *not* provide reliable results. In turn, unreliable NDT may not be successful in achieving its purpose and may result in unwanted consequences. As suggested by the Modular Reliability Model (Müller et al., 2013), human and organisational factors play an important role in reliability, however still to a large extent an unknown one.

This is especially true for mechanised testing, considered by many to be more reliable than manual testing and far less prone to the possibility of human error. This assumption is largely based on experiences from the field and generally higher technical reliability (expressed in terms of probability of detection). As elaborated in the previous chapter, the understanding of potential influencing factors, their interactions, and the potential risks that can arise during mechanised testing is still missing.

The issues surrounding human error and risk management are as old as the human factors science (Woods et al., 2013). The concern for human error arose after observing how vulnerable technological systems are to the actions of human operators (Hollnagel, 1993). According to different statistics, human error is responsible for up to 90% of organisational accidents. Reason (1990, 1997) talks about the "80 - 20 problem": Some 50 years ago, 20% of accidents were attributed to human error, and 80% to technology; today 80% is attributed to human error and 20% to technology. Giesa and Timpe (2002) summarised different reports on accidents that state that 52% of accidents in the nuclear industry, 70% in aviation, and up to 90% in general can be assigned to some kind of human failure. FAA (2009) reports that as many as three out of four aviation accidents result from some kind of human error. Reason and Hobbs (2003) analysed maintenance, calibration, and testing activities in four nuclear power plants and identified that 42 - 65% problems arise from human performance associated with those activities (only 1 - 8% of human performance problems are present during

abnormal and emergency operations). It is no wonder human error is feared of. These statistics make a strong point that human factors are no longer seen as the problem solver but rather as the problem itself (Badke-Schaub et al., 2012).

Every complex socio-technical system with high safety and reliability demand invests effort into the avoidance of adverse effects that could affect people and the environment. Since erroneous actions are unavoidable—they have happened and will continue to happen—means have to be found to identify them, to determine the consequences and the seriousness of those actions, to assess the likelihood of their occurrence, and to find ways to reduce either the actions themselves, or their consequences. To do that, Hollnagel (1993) suggests a combination of theory of human action, appropriate methods for risk and reliability analysis, and a set of strong principles for the man-machine system design. In other words, in order to tackle the risks in NDT, first we need to a) understand the mechanisms underlying potential failure associated with human operators, i.e. human error, b) utilise methods to identify and analyse potential risks, and finally, c) invest efforts into changing faulty practices by appropriate design, therein applying the knowledge of human factors and man-machine design.

The aim of the first study conducted within the scope of this dissertation is to identify risks associated with mechanised NDT and generate methods for preventing them. Furthermore, this work is meant to serve as a foundation for further empirical work.

Considering these aims, this chapter will provide with understanding of the theory of human error, explain modern approaches of its contribution to failure (section 3.1), and present the fundamental principles of risk management (section 3.2). The theoretical part will conclude with the existing risk management practices in NDT and their shortcomings (section 3.3). The identification of risks will be addressed in the empirical part of this chapter, consisting of objectives of the study (section 0), selection of the appropriate risk assessment technique (section 3.5), method, results, and discussion (sections 3.6 - 3.8). The chapter will conclude with the selection of topics for empirical study.

3.1. Human error and its contribution to failure

When thinking about risks of potential failure of NDT to detect all critical defects, it is impossible not to mention the notion of human error. As stated by Hollnagel (1993): "*To err is human; to understand the reasons why humans err is science*" (p. 1). With that respect, science has dealt with human error and thanks to undertakings in cognitive psychology and accident analyses we are today closer to understanding human error and its underlying mechanisms.

3.1.1. Traditional and modern approaches to human error

The commonly accepted and widely used definition of human error is that of James Reason (1990), who defined it as "*all those occasions in which a planned sequence of mental or physical activities fails to achieve its intended outcome, and when these failures cannot be attributed to the intervention of some chance agency"* (p. 9). In simple terms, if an action fails to achieve its intended outcome, we talk about human error. Human error typically refers to mistakes, slips, and lapses.

Whereas cognitive psychologists are concerned with the internal psychological or cognitive mechanisms of the mind that are assumed to explain the action, practitioners look at human error mainly as an exacerbating feature. Hence, the term "human error" is widely used to explain human *action* or an event that happened (observable failure), the *cause* of a mishap or

an accident, or is seen as a *symptom* of deeper trouble (Dekker, 2002; Hollnagel & Amalberti, 2001; Hollnagel, 1993; Woods et al., 2013). Considering human error as a cause and as a symptom do not only constitute only two different views, but also two different eras in the approach towards human error—a difference that is still sometimes blurry to the managers of complex systems.

Underlying the first approach is the tendency to assign "blame" to the operators and inspectors at the sharp end of the line for mishaps, events, and accidents. After all, the errors do become obvious at the hands of the person handling the equipment and making the decisions. People are *available* to be blamed: Since they are working with the equipment, it is probable that the accident would not have happened had the operator not been present. People also have a temporal and a physical relationship with the outcome (Woods et al., 2013). However, this approach is nowadays considered as a *traditional* approach. By concentrating on the individual origins of error, according to Reason (2000), the act is wrongfully isolated from its context and, therefore, important features can be overlooked. First, it is often the best people that make the mistakes, and second, the same combination of circumstances can provoke the same errors, regardless of the people involved. In addition, people in high-reliability organizations are generally motivated to do a good job - what they do generally makes sense to them at the time (Dekker, 2002). Therefore, this view is being replaced by the modern *systems* approach focusing on the underlying conditions that create possibilities for failure, and view human error as a symptom of problems hidden deeper in the system. Efforts are thus invested into the conditions under which people work and ways to prevent the failures (Dekker, 2002; Hollnagel, 1993; Leveson, 2011; Rasmussen, 1997; Reason, 1997). This is achieved by implementing defences. Hence, when adverse events do occur, the question should not be *who* failed, but rather *how* and *why* the defences failed.

To illustrate the difference between the observable failure at the hands of an operator at the sharp end and the underlying causes in the system that may lead to an organisational accident, Reason (1997) introduced the terms *active failure* (human errors and violations that have immediate adverse effects and, through that, a direct impact on the safety of the system) and *latent conditions* (e.g. poor design, gaps in supervision, undetected manufacturing defects or maintenance failures, unworkable procedures, clumsy automation, shortcomings in training, or less than adequate tools and equipment). Latent conditions arise from strategic and top-level decisions made by governments, regulators, manufacturers, designers and organizational managers and are present in all systems, being an inevitable part of organizational life. They can be present for years, before being combined with local circumstances and active failures to penetrate the system's many layers of defences. The impact of these decisions spreads throughout the organization forming a distinctive organizational culture, which then results in the creation of error-producing factors within individual workplaces.

3.1.2. Classifications of human error

The manner by which an error or a failure is observed is called *error mode* or *failure mode*. This term describes the way a failure occurs and its impact on the equipment or operation (MIL-STD 1629A, 1980). There exist several classifications of failure modes, depending on whether the failure is observed from the cognitive perspective or from an empirical one.

Two classifications will be presented here. One of the most common classifications includes that into errors of omission and errors of commission, developed by Swain & Guttmann (1983) for the purposes of the Human Reliability Analysis. They refer to those events that constitute incorrect human inputs to the system. They are regarded as errors *only* if they can

result in a consequence that might be undesirable for the system, thereby affecting the system reliability and safety.

Looking for a way to describe situational and organisational factors that can contribute to failure, i.e. the latent conditions, Reason, Shotton, Wagenaar, Hudson, & Groeneweg (1989), identified 11 general failure types (GFTs) for the purposes of their error management tool Tripod-delta (Hudson, Reason, Bentley, & Primrose, 2013). Both classifications are presented in Table 5.

Table 5: Selected classifications of human error

Active failures	Latent conditions
Error of omission, i.e. omitting a task or a part of a task (e.g. a step in the task)	**Hardware (H)**, i.e. quality and availability of tools and equipment
Error of commission, i.e. adding something that should not be there	**Design (D)**, i.e. no external guidance by the designer, designed objects are opaque, the designed object does not provide feedback
Selection error, i.e. incorrect choice among a range of options (e.g. selects the wrong control, issues wrong command or information)	**Maintenance management (MM)**, i.e. safe planning of operations
Error of sequence, i.e. incorrect sequencing of actions or events	**Procedures (P)**, i.e. quality, accuracy, relevance, availability, and workability
Time error, i.e. action carried out too early or too late	**Error-enforcing conditions (EEC)**, i.e. error-producing and violation-promoting conditions related to the individual or to the workplace
Qualitative error, incorrectly carrying out an action (e.g. too much, too little)	**Housekeeping (HK)**, i.e. the problem has been there for some time, the organisation was aware of it, but did not deal with it, e.g. insufficient personnel, poor definition of responsibility, bad hardware
	Incompatible goals (IG), i.e. individual (preoccupation with private issues), group (norms incompatible with safety goals) and organisational goal conflicts (incompatibility between safety and productivity)
	Communications (C), i.e. communication channels do not exist; necessary information is not transmitted; information is sent, but misinterpreted by the receiver
	Organisation (O), i.e. organisational structure, organisational responsibilities, and the management of contractor safety
	Training (T), i.e. failure to understand training requirements; incompatibility between training and the operation, poor mixes of experienced and inexperienced personnel, poor task analysis, inadequate competence, etc.
	Defences (DF), i.e. failure in detection, warning, personnel protection, recovery, containment, escape, and rescue

3.1.3. Error prevention

Typical methods for the prevention of errors include designing the system so that it is simple and easy to use, training, effective warnings that can anticipate a system state that will likely lead to error, and restricting the exposure of the operator to opportunities for error (Sheridan, 2008).

The attempts to minimise the occurrence of errors are either proactive or reactive in nature. The *proactive* approach is based on improving the human-system interface. This is most commonly achieved by creating decision aids, improving the training or the procedures, automating features of the system interface, etc. The *reactive* approach focuses on eliminating the reoccurrence of already occurred errors. The common term used for these error prevention or minimisation techniques is defences or barriers.

Installing defences can sometimes even harm the system, because in spite of their original purpose, they can backfire (Dekker, 2002; Reason, 1997; Woods et al., 2013). The basic

premise is that any change could give rise to new risks. Reason (1997) refers to them as "defence-related ironies and paradoxes". The most frequently cited examples include automation (e.g. Bainbridge, 1987) and the procedures (e.g. Reason, 1995).

In summary, it is human to err and even the best organisations with the best and highly motivated people face a risk of accidents. This is because the state of no-risk is not achievable. Nonetheless, it is something all organisations strive to. In the attempt to prevent adverse effects, one must take measures. To start with, organisations need to stop looking for the one to blame (the person at the hands of which an event happened) and look deeper for the underlying mechanisms that may lead to errors. Problems with inattention, forgetfulness, or distraction can be only partly tackled with, but not exterminated. The conditions, under which people work, on the other hand, can be subject to change and, hence, should be optimised.

3.2. Fundamental principles of risk management

All organizations face risks that can have an effect on the achievement of their objectives. The ability to identify in advance the events that may lead to adverse outcomes as well as the outcomes themselves is a critical prerequisite for safety (Hollnagel, 2008b). Risk is defined as an "*effect of uncertainty on objectives"* (ISO 31000:2009, p. 1), where *effect* refers to a deviation from an expected objective, and *uncertainty* to a state of deficiency of information related to an event, its potential consequence, or its likelihood. In simpler words, risk can be seen as "*the notion of an adverse outcome or a potential negative impact that arises from some present process or future event*" (Hollnagel, 2008b, p. 33).

Accurately assessing (risk assessment) and successfully containing those risks (risk treatment) are in the core of an effective risk management. Management of risk, among other things, allows the organisations to achieve their goals with higher likelihood, encourages proactive measures against risks, raises awareness of possible threats, improves organisational learning and resilience, and so on (ISO 31000:2009).

Risk management refers to "*coordinated activities to direct and control an organisation with regard to risk*" (ISO Guide 73:2009, p. 2). The major steps include understanding of the problem (and if there is a problem at all), understanding of the underlying mechanisms related to the potential adverse outcomes of the risk with respect to their causes, consequences and their magnitude, and, finally, providing with ways of reducing or eliminating the risks, or of protecting from their consequences (Hollnagel, 2008b).

The ISO standard on risk management (ISO 31000:2009) proposed the main principles of risk management, the framework, as well as a description of the risk management process. Figure 11 depicts the relationships between different steps in the risk management process.

According to this model, the *communication and consultation* (1) of the organisation with internal and external stakeholders about risk should take place during all the stages of the risk management process, and, therefore, be developed early in the process. This includes a dialogue about the existence, nature, form, likelihood, significance, evaluation, acceptability, and treatment of risk.

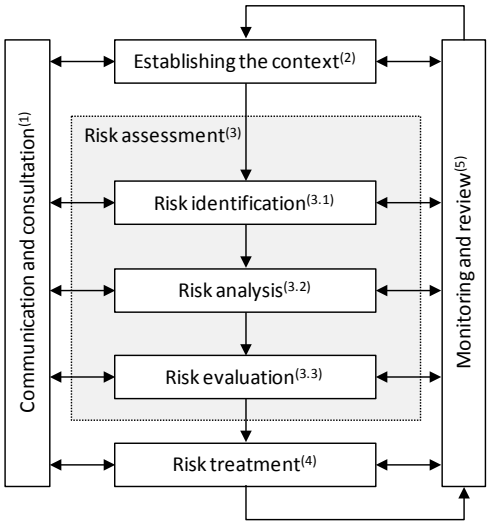


Figure 11: Risk management process (ISO 31000:2009; reproduced with permission of DIN Deutsches Institut für Normung e.V.)[4]

Establishing the external and internal environment in which the organisation seeks to achieve its objectives and determining the risk criteria (term of reference against which the significance of a risk is evaluated) are relevant contributors to *establishing the context* (2) for the assessment. The external context can refer to the social, regulatory, or economic context—for example, and internal to the organisational structure, policies, information flow, decision making processes, and so on.

To be able to eliminate the risks successfully, they first need to be identified, which is the purpose of *risk assessment* (3). The following fundamental questions are in the core of risk assessment (IEC/ISO 31010:2009, p. 6):

- What can happen and why (by risk identification)?
- What are the consequences?
- What is the probability of their future occurrence?
- Are there any factors that mitigate the consequence of the risk or that reduce the probability of the risk?

As illustrated by the questions, a holistic approach to risk assessment includes not only the identification of risks, but also aims at developing a higher understanding of the risks. This is achieved by means of risk identification, analysis, and evaluation (IEC/ISO 31010:2009; ISO 31000:2009; ISO Guide 73:2009).

Risk identification (3.1) refers to finding, recognising and describing risks. It includes identifying the risk sources[5], events, their causes[6], and their potential consequences. It is of utmost importance that all risks are included at this stage because a failure to do so will result in the

[4] The definitive version for the implementation of this standard is the edition bearing the most recent date of issue, obtainable from Beuth Verlag GmbH, Burggrafenstraße 6, 10787 Berlin, Germany.
[5] Risk source is defined as an element that can give rise to risk, either alone or in combination with other elements (ISO Guide 73:2009).
[6] Cause is typically related to the event. An event can have several causes (ISO Guide 73:2009).

risk not being further analysed. In IEC/ISO 31010:2009 it is stated that "*(...) it is important that due recognition is given to human and organizational factors when identifying risk. Hence, deviations of human and organizational factors from the expected should be included in the risk identification process as well as "hardware" or "software" events"* (p. 12).

Risk analysis (3.2) aims to develop an understanding of the risk, i.e. it is a process of comprehending the nature and determining the level of risk. (*This process is a basis for risk evaluation and risk treatment.*). In risk analysis, the causes, and sources of risk, their consequences and the probability that those consequences can occur are considered in depth. The consequences, probability, and level of risk can be analysed using qualitative (using significance levels such as "high", "medium", or "low"), semi-quantitative (using numerical rating scales), or quantitative methods (estimation of the risk in specific units defined when developing the context). It is important to note that failures do not always happen due to cause-effect chains, but may also result from unordinary combinations of conditions resulting from poorly understood characteristics of socio-technical systems (Hollnagel, 2008b), which is why the process of understanding of risks is of utmost importance.

During *risk evaluation* (3.3) the results of the risk analysis are compared with the risk criteria (defined during the establishment of the context) to determine the risk's significance and type. The purpose of risk evaluation is to use the knowledge gathered during the previous analyses to make decisions about future actions. Those decisions include considering whether a risk needs treatment, deciding whether they should, and if yes, which activities should be taken, and assigning priorities for treatment.

Addressing the risks in terms of *risk treatment* (4) takes place after the risk assessment. A risk can be treated by avoiding the risk (deciding to exclude the activity that bears risk), increasing the risk (to pursue an opportunity), removing the risk source, changing the likelihood, changing the consequences, sharing the risk with another party and by retaining the risk by decision. The risk treatment is cyclical in its nature and it involves assessing a chosen risk treatment, deciding whether the remaining risk is tolerable; and if not, a new risk treatment is generated and that treatment's effectiveness is again assessed. It is important to note that risk treatment can introduce new or secondary risks, which is why the process needs to be carefully *monitored and reviewed* (5).

Typically, risks can be managed during the design of processes by thinking ahead and imagining all those situations, in which a system can fail (prospectively), or—more often—after a system's failure by learning from accidents and incidents (retrospectively).

3.3. Risk management in NDT

NDT has long been a part of a risk management practice in organisations with high reliability and safety demand. It is typically taken into account in assessing the risk of failure of a component or of the entire system. This is achieved by, first, assessing the risk of the component failing (typically a field of fracture mechanics) and by determining which NDT method needs to be applied and where. After the NDT had been applied, the statement about the integrity of the component is taken as an input into the risk assessment. Based on this result and on the general likelihood that the component will contain defects and eventually fail, future inspections are planned. In this process, the necessary methods and the frequency of their application are taken into account. This process is known as the *risk-based management*. However, it does not include management of risks arising from the application of NDT in a

specific environment and neither does it include the consideration of a potential failure caused by the interaction of human operators with technical systems and the organisation.

Personal and technical risks associated with the programme of final disposal of spent nuclear fuel and waste have already been included in the long-term safety assessment undertaken by the Swedish and Finnish authorities. The aim is to satisfy nuclear safety and environmental objectives, as well as corporate and occupational safety (Posiva Oy, 2015; SKB, 2013). The risks that could influence the safety of the spent nuclear fuel disposal are identified already at the planning stage of the operation. Hence, the conditions of the manufacturing and welding processes of the canister components to be used for the disposal of the spent nuclear fuel are under continuous scrutiny. Determining the risk of premature canister leak caused by defects in the sealing weld and the enveloping components is an important part of the final risk assessment, for which purpose the quality of the production process and the reliability of the NDT system must be assessed. Attempts to minimise the risk of missed defects, i.e. to increase the NDT reliability, is, hence, a risk minimising measure (Müller et al., 2007; Pitkänen, Salonen, Bertovic, Müller, & Pavlovic, 2011).

Even though risks associated with the expertise level of human resources and with the production and installation of critical components are already a part of the long-term risk assessment in the disposal program (e.g. Posiva Oy, 2015), there have been no such attempts in the NDT inspection process up to this date. Hence, the risks associated with the NDT inspecting personnel and with the process of carrying out an inspection, using mechanised NDT systems, are still unknown.

There are two major reasons essential for studying the human factors in this application of NDT methods. First, the effects of human factors on the reliability of mechanised testing remain unknown. Second, an attempt to apply mechanised NDT methods in the management of spent nuclear fuel is the first of its kind. Some of the differences in the application of NDT in this field compared with, for example, nuclear power plants, include:

- the inspected materials (e.g. copper and cast iron) and the material thickness,
- the need for the components to withstand greater loads (i.e. deep geological disposal),
- difficult environment (e.g. upcoming ice ages and rock shear movement)
- longer periods of time during which the condition of the structure has to remain unchanged (a minimum of 100,000 years),
- no possibility of a repeated inspection (once the canister has been sealed and placed into the repository, it will no longer be possible to repeat the inspection), etc.

Inability of a repeated inspection, as well as unimaginably long periods of time during which the components need to serve their purpose (protect the highly radioactive spent nuclear fuel) require the highest possible reliability standards.

These differences constitute a challenge for the development of NDT as well as for the development of the inspection procedures and the design of the workplace. Consequently, this presents a challenge for the consideration of human factors.

3.4. Objectives of the study and assumptions

Observations in the field and the communication with experts revealed the following problems with risk in NDT:

- The risks associated with human and organisational factors in mechanised NDT are unknown,
- The variability in the inspection results is frequently assigned to the inspectors and their working environment, thereby neglecting other potential influences, and
- Organisations with high safety and reliability demand, such as the management of spent nuclear fuel, rely on reliable NDT methods.

Thus, the objectives of this study were to:

- Identify the potential failures that increase the risk that mechanised NDT will not fulfil its objective, i.e., detect all critical defects,
- Analyse the potential failures, with respect to their origin and effects on the execution of the NDT task, and
- Provide countermeasures to minimise the future risk of failure.

In line with the current state of the art in human error research, the following was assumed:

- There is a risk of failure in mechanised testing.
- The sources of failure can be seen not only in the technology but also in the individual and the organisation.
- The currently installed preventive measures are insufficient to prevent failures in the execution of the mechanised NDT inspection task.

3.5. Selection of a risk assessment technique

There are four NDT methods planned to be used for the inspection of the canister components and welds for the purposes of the spent nuclear fuel disposal, and all four were investigated within the scope of this study, i.e. ultrasonic testing (UT), radiographic testing (RT), eddy-current testing (ET), and visual testing with a remote camera (rVT).

Before the risks can be assessed and managed, it is important that the NDT task under scrutiny and the different roles played by the technology, inspectors, and the organisation are fully understood.

Because some NDT subtasks are allocated to the equipment and others to the inspectors, and because it was never thoroughly done before, the first approach was to conduct a detailed task analysis of the single NDT methods. Task analysis is one of the most commonly used human factors methods. It is used to help the analyst to understand and represent human and system performance in a particular task or a scenario. Task analyses are used for understanding the required human-machine and human-human interactions by decomposing tasks or scenarios into component task steps or physical operations.

Four different NDT methods of interest in the study were observed and consequently described by means of the Hierarchical Task Analysis (HTA; Annett, 2004). The initial plan to use the HTA descriptions to identify positions in the task sequence at which errors can occur was soon abandoned, after it was realised that the methods are under development and, hence, being continuously changed. Consequently, there was a need for a new approach to

identifying risks. Because NDT methods remain under development and the operation is scheduled to start after the year 2020—instead of a deterministic approach, such as was the HTA—a prospective approach was needed (i.e. one that would enable the identification of future inspector-related risks).

3.5.1. Prospective risk assessment techniques

The IEC/ISO 31010:2009 standard on risk assessment techniques suggests that the following factors can influence the selection of the appropriate risk assessment technique:

- The applicability of the method to the desired steps in the risk assessment process (some methods are applicable to identify, analyse and evaluate the risks, as well as support risk treatment, whereas some methods are only applicable to some steps).
- The availability of resources (e.g. skills, experience, capacity, and capability of the team; time and budget restraints).
- The nature and the degree of uncertainty associated with the risk (the availability of sufficient amount of information needed to assess the risks).
- The complexity of the problem and the methods required to analyse it (consideration of single risks versus consideration of dependencies between risks).

The same standard suggests a number of techniques, describing them with respect to applicability for different stages of the risk assessment process (identification, analysis, evaluation), and with respect to their attributes (necessary resources, the degree of uncertainty, complexity, and the availability of a quantitative output).

The consideration of the applicability of different techniques for the purposes of this investigation resulted in a handful of suitable techniques. They include: Hazard and operability studies (HAZOP), Structured "What-if" Technique (SWIFT), Fault-tree analysis, Failure Modes and Effects Analysis (FMEA), as well as different types of Human Reliability Analyses (HRA). All of these techniques are applicable for the entire risk management process, i.e. they enable risk assessment and generate methods for risk treatment. However, they all contain strengths and weaknesses with respect to the following conditions of the investigation (see Table 6):

- The aim of the study is to identify potential failures, analyse them with respect to their origin and consequences, and generate risk reduction measures. Thereby the focus is on the individual, technology, and the organisation, and on the observable failure, rather than on the cognitive aspects of human error.
- The NDT methods and the procedures to be analysed are not completely developed, and hence, not fully defined.
- The understanding of the system, as a part of which NDT will operate, is still missing.
- The participants are not adequately experienced (they are not experienced *inspectors*, but rather experienced *developers*).
- The aim is a qualitative empirical description and understanding of the failures, rather than a cognitive or a quantitative one.

The consideration of strengths and weaknesses resulted in the choice of the FMEA. Apart from fulfilling the requirements (applicable for risk identification, analysis, and evaluation), FMEA also overcomes most of the weaknesses of other methods (e.g. FMEA does not require high level of documentation or good experience of the team, and it is not too detailed

or concerning only the cognitive aspects of human error). The method will be described in the following section.

Table 6: Strengths and weaknesses of the selected risk assessment techniques

Technique	*Strengths*	*Weaknesses*
HAZOP	Identifies failures, causes, and their consequences, and generates risk treatment measures.	Is carried out at a detailed design stage, when a full process diagram is available and the changes are still possible. Requires a high level of documentation and procedure specification.
SWIFT	A simpler alternative to HAZOP (see above).	Requires good experience from the team.
Fault-tree analysis	Identifies and analyses factors that can contribute to an event using a top-down approach.	Requires understanding of the system and of the failure causes. Evaluates the failures in binary terms (failed/not failed). Characteristics of human error not covered by the analysis.
FMEA	Identifies human failures, causes, and their consequences, and generates risk treatment measures. Improves design of the procedures and processes. Applicable to human, equipment and system failures; hardware, software, and the procedures.	Identifies only single failure modes, not the combinations of failure modes.
HRA (THERP, CREAM, ATHEANA)	Investigates the impact of humans on system performance. Evaluates human error influences on the system. Can be qualitative and quantitative. Identifies human error probabilities. Identifies performance shaping factors. Evaluates degradation of the man-machine system (MMS) likely to have been caused either by humans or by the man-machine interaction (MMI) It is standard practice in the nuclear industry.	Requires strict definition of tasks, practical experience of the error that can occur. Has difficulty with partial failures and poor decision making. THERP - requires detailed decomposition of the activities into task elements. CREAM – cognitive method (focuses on the phenotype and genotype of human error). ATHEANA – identifies vulnerabilities of the operator's knowledge base.

3.5.2. Failure Modes and Effects Analysis (FMEA)

Failure Modes and Effects Analysis (FMEA) is a standard risk assessment tool. Originally developed by the US Armed Forces in 1949 and revised in 1980 (MIL-STD 1629A), FMEA is defined as "*a procedure by which each potential failure mode in a system is analyzed to determine the results or effects thereof on the system and to classify each potential failure mode according to its severity*" (p. 4). FMEA is used to identify potential failure modes (the manner by which a failure is observed), to determine their effects (or consequences) on the operation of the system, to identify the mechanisms of failure, and to identify actions to avoid and/or mitigate the effects of the failure on the system. It is applicable for the entire scope of the risk assessment process, that is for the identification, analysis and the evaluation of risks (IEC/ISO 31010:2009).

A crucial step in this process is anticipating what might fail or go wrong. Its use is advantageous in examining potential reliability problems early in their development cycle when taking action to overcome these issues is easier, thereby enhancing reliability through design (Cassanelli, Mura, Fantini, Vanzi, & Plano, 2006).

Traditionally used to identify failures and failure modes of technical systems, the FMEA extended its application over time and is nowadays used to (IEC/ISO 31010:2009, p. 46):

- assist in selecting design alternatives with high dependability;
- ensure that all failure modes of systems and processes, and their effects on operational success have been considered;
- identify human error modes and effects;
- provide a basis for planning, testing, and maintenance of physical systems; and
- improve the design of procedures and processes.

FMEA has found its use in various industries, such as the nuclear industry, transportation, wind turbines, health care, etc. (Arabian-Hoseynabadi, Oraee, & Tavner, 2010; Dhillon, 2003, 2007; FAA, 2000; Haapanen & Helminen, 2002; Wetterneck, Skibinski, Schroeder, Roberts, & Carayon, 2004). It is applicable to human, equipment, and system failure modes; as well as to hardware, software, and procedures. Instead of listing components and their failures—if interested in the human failure modes—an investigator can list the human errors and violations that can occur (omissions, commission, etc.) and their possible effects on the system (Wickens, Lee, Liu, & Gordon Becker, 2004). Algedri and Frieling (2001) suggest that a human-oriented FMEA can lead to improvements on personal, ergonomic (optimised working conditions), and organisational level (optimised interaction of the man, technology, materials, and method), which can result in a significant decrease of system failures.

FMEA may be followed by a criticality analysis, which assigns significance to each failure mode, thereby extending the FMEA to a FMECA, i.e. Failure Modes and Effects and Criticality Analysis. Criticality can be assigned qualitatively, semi-qualitatively, or quantitatively, usually by assessing the probability that a failure mode will result in system failure, by assessing the level of risk associated with a failure mode (typically used for equipment failures, systems or processes), or by assigning a risk priority number—RPN (IEC/ISO 31010:2009).

RPN is a semi-quantitative method of criticality obtained by assigning a numerical value to each failure mode with respect to its severity (or relevance), occurrence, and the probability of it being detected. Since RPN is obtained by multiplying these three values, its outcome is an assessment of the criticality of the system, where a higher RPN is associated with the most risky elements in the process. RPN is typically evaluated on a scale from 1 to 10, but other variations are also possible, i.e. 1-3, 1-5, as reported by van Leeuwen et al. (2009).

3.6. Method

3.6.1. Participants

The workshops were carried out at the two nuclear waste management companies: at SKB in Oskarshamn (Sweden) and at Posiva Oy in Helsinki (Finland) in duration of 1-1.5 days per method.

Four to five experts took part in the evaluation of each method. All participants were considered experts in their respective NDT methods (even though not certified). They were all involved in the development of the methods to be used for the inspection of the canister

components for the storage of spent nuclear fuel either in Finland or in Sweden, and were, therefore, qualified for the participation in the analyses.

3.6.2. Procedure

Altogether six FMEA analyses were carried out: five were carried out to assess risks during the evaluation of data collected with UT, RT, ET and rVT, and one during the acquisition of data with phased array UT. (*Since two companies wanted to assess the risks of the methods they use, the FMEA for the data evaluation with UT—a method used by both companies—was evaluated two times.*). More attention was given to data evaluation than to data acquisition, for the following reasons: first, due to its higher perceived criticality for the future of the component (whereas errors in data acquisition can be detected during evaluation, it is harder, sometimes even impossible, to detect errors during evaluation), and second, due to a larger involvement of human inspectors in the task.

The FMEA/FMECA carried out within the scope of this study was adapted to the needs of identifying potential risks in mechanised NDT and conducted using the following steps:

- Decomposition of the task into sub-tasks.
- Definition of aims for the sub-tasks.
- Identification of possible failures/errors.
- Consideration of potential causes and effects of failures.
- Identification of existing preventive measures/barriers.
- Identification of potential preventive measures/barriers.
- Assessment of error probability (EP), relevance of effects (R), and detection probability (DP) according to a key shown in Table 7.
- Calculation of the risk priority (RPN = EP x R x DP).

Table 7: Risk priority assessment key

Category	Assessment		Description
Error probability (EP)	*	Low	The occurrence of errors and deviations is improbable
	**	Medium	Errors or deviations occur seldom
	***	High	The probability that errors or deviations will occur is very high
Relevance (R)	*	Low	No observable effects of an error / deviation
	**	Medium	Effects lead to dissatisfaction, e.g. delays, increasing efforts
	***	Serious	Safety related effects or violations of rules and regulations
Detection probability (DP)	*	High	Error will be detected in successive steps
	**	Medium	Error will be detected by 100% testing / quality checks
	***	Low	There is no testing / possibility of independent tests

3.7. Results

Even though the potential errors were analysed separately for each NDT method, and are, therewith, method specific, some similarities in the way each method is applied can be found. It is for this reason that it was possible to combine the results for evaluation of data with different methods, with the aim of reaching general conclusions about the process. The detailed results of the analyses have been presented in unpublished internal reports and can be

obtained upon request (e.g. Bertovic, 2014). In the following, a summary of the collected results will be presented.

The results section starts with the description of the evaluated tasks (section 3.7.1), and is followed by the identified failure modes, their causes, consequences, preventive measures, and finally, the risk priority assessment (sections 3.7.2 - 0).

3.7.1. The evaluated tasks

Before the risks can be successfully assessed, it is important to understand the task that is to be analysed. For that purpose, this section will offer a description of the main characteristics of the task. This description is a result of the conducted HTA and the FMEA.

The process of NDT using mechanised systems can be divided into two major processes: the data acquisition (or *collection*) and the data evaluation (or *analysis*).

The data acquisition refers to a process of collecting data by mounting the NDT system onto the component and then rotating the component. The so-called manipulator is used to move the equipment along the component in the axial direction (Figure 12).

Figure 12: The ultrasonic equipment mounted onto a rotating component
(Source: SKB; with permission)

The data acquisition with UT is typically carried out following these steps:

- **Preparing the component**, i.e. mounting the component onto the rotator[7].
- **Preparing the equipment**, i.e. choosing the correct hardware and software, and ensuring they are functioning correctly.
- **Setting up the sensitivity** (also referred to as *calibration*), i.e. the process of scanning of a component specimen equivalent to the actual component (i.e. reference specimen) containing built-in defects (i.e. reference defects) of desired properties that must be detected during the inspection. This process establishes the sensitivity

[7] A rotator is a device that rotates the component while the UT equipment is at a fixed position being manipulated only in the axial direction along the component.

level needed for the inspection. It is carried out by mounting of the equipment onto the reference specimen, scanning of the component and evaluating the collected data.

- **Scanning the component**, i.e. the process of inspection of the component carried out by positioning the equipment, setting up the software, collecting data during the scanning process, and by saving the collected data.
- **Checking the sensitivity** (*calibration check*), i.e. repeated scanning of the reference specimen to assure that the sensitivity had not been changed over the course of the scanning, and corresponding evaluation of collected data.

Apart from the scanning process itself, during which the ultrasonic phased array probe is automatically moved over the component, the inspector takes part throughout the entire process.

The data evaluation process, as planned for the purposes of spent nuclear fuel disposal, can be carried out following these steps (UT, RT, ET, rVT):

- **Preparation for the evaluation**, i.e. it includes the preparation of the software and selection of the data.
- **Identification of indications**, i.e. visual search for indications on a computer screen.
- **Characterisation of indications**, i.e. determining whether an indication stems from a defect or from the geometry (e.g. edges of the component) and determining the defect type (e.g. crack vs. pore, single indication vs. cluster of indications).
- **Sizing and localisation**, i.e. measuring the size of an indication (length, width) and its position in the component (e.g. depth).
- **Decision making**, i.e. a recommendation[8] about whether the component should be *accepted*, i.e. the component does not contain critical defects, or *rejected*, i.e. the component contains critical defects and is not fit for purpose. Note that incorrect rejection will lead to a financial cost, whereas incorrect acceptance will present a threat to safety.
- **Reporting**, i.e. documenting the results of the inspection.

Data evaluation is a complex signal detection, evaluation and decision making task, requiring knowledge, skill and experience from the evaluator. The entire process is carried out following an inspection procedure.

3.7.2. Failure modes

Thirty-eight tasks in data evaluation and 30 in data acquisition were analysed resulting in the identification of altogether 90 failure modes in evaluation and 68 in acquisition.

Table 8 and Table 9 contain examples of evaluated potential failure modes during data acquisition and data evaluation, organised according to the sub-tasks.

[8] The final decision about the component's (or the weld's) acceptance or rejection is not made by the NDT personnel, which is why they can only *recommend* further action based on the collected data.

Table 8: Failure modes associated with the subtasks in data acquisition with UT

Task	Sub-task	Failure modes
Data acquisition	Preparation of the component	Incorrect component or incorrect component orientation
	Preparation of the equipment	Inappropriate choice of the equipment, equipment malfunctioning
	Sensitivity settings (calibration)	Incorrect physical setup of the equipment, incorrect scanning parameters, misinterpretation of the collected data, incorrect evaluation of the data and decision about the quality of the completed process
	Scanning of the component	Altering of the physical conditions between the calibration and the scanning, incorrect scanning parameters, incorrect scanning process, incorrect verification of the data
	Sensitivity check (calibration check)	Change of the physical conditions from the initial calibration, incorrect scanning parameters, misjudging data inconsistency

Table 9: Failure modes associated with the subtasks in data evaluation (UT, rVT, ET, RT)

Task	Sub-task	Failure modes			
		UT	rVT	ET	RT
Data evaluation	Preparation of the software for the evaluation	Incorrect selection of the data file, inspection technique or evaluation area, incorrect settings	Missing image quality check, incorrect starting point, missing initial scanning run	Missing or incomplete data validity check	Incorrect image adjustment, inappropriate image quality and scale calibration
	Identification of indications	Missing indications, false alarms			
	Characterisation of indications	Incorrect defect type, misjudgement of the defect's origin (geometrical indication vs. actual defect)			
	Sizing & localisation	Incorrect size measurement, incorrect location of the indication			
	Decision making	False recommendation (acceptance/rejection of the component)			

3.7.3. Causes

The potential causes of failures are associated with the individual, the technology, and the organisation.

- **Individual**: The individual can be a source of error both unintentionally and through rule violations. Some examples include:
  - Unintentional, e.g. subjective assessment criteria, cognitive biases (confirmation bias, representativeness bias, and availability bias), sensitivity to colours, reduced attention, lapses, over trust in automation, inexperience, and so on.
  - Rule violations, e.g. not following the inspection procedure.

- **Technology**, e.g. image quality, display characteristics, defects' characteristics (e.g. indication "hidden" behind a geometrical indication, too many indications, indications close to one another), equipment malfunction, and so on.
- **Organisation**, e.g. the working environment, organisation of the inspection process (e.g. long working hours), flawed inspection procedures, commercial pressure (i.e. time pressure), and so on.

The information about potential failures and their assumed causes was taken a step further by classifying the active failures and possible latent conditions that might have lead to those failures, according to the known classifications of *incorrect human outputs* (Swain & Guttmann, 1983) and the *11 General Failure Types* (Hudson et al., 2013; Reason et al., 1989). The assignment of the active error types and of the identified latent conditions is presented in Table 10.

Table 10: Assignment of incorrect human outputs and general failure types to the failure modes at different steps of the execution of the NDT task

Task	Method	Subtask	Incorrect human outputs					General failure types, GFTs										
				Commission														
			Omission	Selection	Sequence	Time	Qualitative	Hardware	Design	Maintenance management	Procedure	Error-enforcing conditions	House-keeping	Incompatible goals	Communication	Organisation	Training	Defences
Data acquisition	UT	Preparation of the component		✓					✓		✓	✓			✓	✓		
		Preparation of the equipment	✓	✓			✓	✓	✓	✓	✓	✓					✓	
		Sensitivity settings (calibration)	✓	✓			✓	✓	✓		✓	✓	✓					✓
		Scanning of the component	✓	✓			✓	✓	✓		✓	✓						✓
		Sensitivity check (calibration check)	✓	✓			✓	✓	✓		✓	✓	✓	✓	✓		✓	
Data evaluation	UT, RT, ET, rVT	Preparation	✓	✓			✓	✓	✓		✓	✓	✓				✓	✓
		Identification	✓	✓			✓	✓	✓			✓		✓			✓	✓
		Characterisation		✓			✓	✓			✓	✓	✓				✓	✓
		Sizing and localisation	✓	✓			✓	✓	✓			✓					✓	✓
		Decision making		✓			✓					✓		✓		✓	✓	

The Table 10 shows that the typical errors in NDT include both omissions and commissions (selection and qualitative errors). With regard to the possible latent conditions that might pave the way to failure, the most frequently suggested include error-enforcing conditions (they include both individual and the environment), design, hardware, training, and the procedure.

3.7.4. Consequences

The consequences of the listed failures were categorised into two levels: *direct,* which can be directly observed and *indirect consequences*, which can occur if the failures leading to direct consequences are not recovered. For examples, see Table 11.

Table 11: Direct and indirect consequences of potential failures in data acquisition and data evaluation

Task	Method	Direct consequences	Indirect consequences
Data acquisition	UT	Invalid calibration, e.g. incorrect settings during the calibration because of incorrect, damaged or incorrectly adjusted equipment	The whole inspection needs to be repeated; incorrect sensitivity in the inspection
		Incorrect sensitivity in the inspection	Incorrect results for the following data evaluation, i.e. defects can be missed or incorrectly sized, including a risk of false calls
		Poor data quality, e.g. missing data, incomplete coverage of the component	Defects can be missed or incorrectly sized in the following data evaluation
Data evaluation	UT RT ET rVT	Incorrect evaluation of the image quality	False recommendation (acceptance/rejection)
		Non-inspected areas	False acceptance of the component
		Missing defects	False acceptance of the component
		False alarms	False rejection of the component
		Incorrect positioning of the indication	False recommendation (acceptance/rejection)
		Incorrect sizing of the indication	False recommendation (acceptance/rejection)
		Time delay	Financial cost

3.7.5. Error detection

Errors in data acquisition could be detected through consecutive steps, equipment malfunctioning or through data check. Errors in the evaluation of the calibration data are not always easily detected.

In data evaluation, errors could be detected through crosschecking of the results or through complimentary methods[9], if available. However, there is a high probability that some errors—e.g. missing inspection areas—may not be detected, thereby increasing the risk of missing defects.

Considering the possibility of undetected errors, one has to evaluate the existing preventive measures and generate new ones, if needed.

3.7.6. Existing preventive measures/barriers

At the time of the analysis, the installed barriers to prevent the errors in data acquisition included relying on operator skill, training, inspection procedures, and, if available, a checklist

[9] At the moment it is planned that some parts of the component, e.g. the weld, may be inspected by more than one method. E.g. ET and VT could complement each other in search for surface or near-surface defects, whereas UT and RT could be used to search for defects deeper in the volume of the component.

with steps that needed to be signed off during the task. In addition, having complimentary methods served as a barrier in the evaluation of data.

3.7.7. Potential preventive measures/barriers

The consideration of potential barriers suggested a potential for building *new* ones.

In data acquisition (UT), these include:

- **Automation** of the component identification and of the choice of proper tools (e.g. probes and cables) using a bar code reader; automatic refill of the coupling when it reaches a certain level, synchronisation of the movement between the UT system and the manipulator, and so on.
- **Hardware and software solutions,** e.g. redesign of the probe fixture; alarms for inconsistencies, insufficient amount of couplant, or to indicate the incorrect orientation of the component; automatic data archiving; and so on.
- Improvement of the inspection procedures and instructions in terms of quality and information they contain.
- **Organisation,** e.g. maintenance of the equipment, instructing and training the personnel; providing a disturbance-free environment; ensuring that the process is performed according to the appropriate up-to-date procedure and that all of the correct tools are used; motivation; clear responsibilities of the personnel; and so on.
- **Human redundancy** in the evaluation of the sensitivity settings (i.e. calibration) and in deciding whether or not to accept the inspection.

In data evaluation (UT, RT, ET, rVT), the following potential preventive measures were suggested:

- **Automation,** i.e. automated detection and sizing of indications (with confirmation by an inspector).
- **Software solutions,** e.g. software alarms for areas not being inspected, changing the colour scale (e.g. red denotes a high magnitude of the signal (alarm), and green a low one (safe)), defining screen view parameters (resolution, size, distance from the screen, and so on), plausibility checks in reporting.
- Improvement of the inspection procedures and instructions in terms of their content and usability.
- **Organisation,** e.g. disturbance-free environment; better time management; organisational learning (learning from previous events through event analyses).
- **Human redundancy,** i.e. evaluation performed by two independent inspectors, e.g. in cases of uncertainty, after a critical defect had been found, or randomly by the supervisors (the frequency of which would depend on the frequency of error occurrence).
- **Training,** i.e. in terms of introducing human factors training, by e.g. increasing awareness of possible cognitive biases, group effects, mistakes, etc.
- **Detection and decision aids,** i.e. visual representations of possible known defects (defect catalogue) and further development of detection and sizing aids.

3.7.8. Risk priority number (RPN)

In the following, the results will be discussed with respect to their risk priority rating. This assessment was used as an indicator for subtasks that require primary attention.

3.7.8.1. Data acquisition (UT)

Among all analysed subtasks in the data acquisition, six that were assigned the highest RPNs are presented in Figure 13.

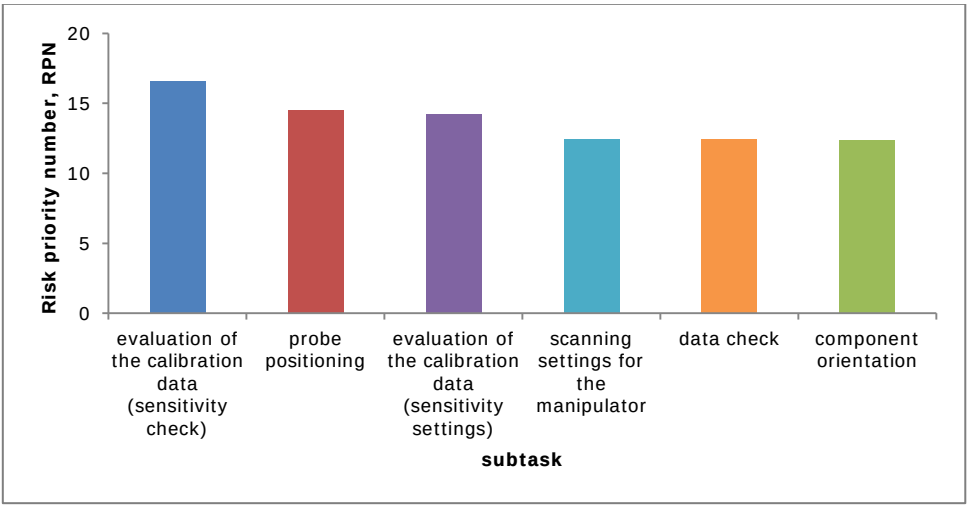


Figure 13: The subtasks in UT data acquisition with the highest assigned risk priority

The evaluation of the sensitivity settings, both before and after the scanning, was rated highly on the risk scale suggesting that the evaluation of data by the human inspector—even during the data acquisition—plays a highly important role. Here, errors may be highly critical for the safe disposal of the component. High risk was also assigned to the positioning of the probe. Errors in this assignment can affect the quality of the collected data and consequent misplacement of critical defects, which may affect the assessment of their criticality. Scanning settings, control of the quality of the collected data, and the component orientation, can, if carried out incorrectly, furthermore present with a high risk to NDT reliability.

3.7.8.2. Data evaluation (UT, RT, ET, rVT)

In data evaluation, the risk priority ratings were assigned to subtasks of four NDT methods (UT was analysed twice and the results have been presented separately for each participating company). Their results are shown in Figure 14. Note that the *characterisation* (in RT) and *decision making* (in VT) subtasks were not developed yet, at the time of the analyses, to take part in the evaluations, which is why no RPN ratings were assigned for them.

The results illustrate high potential risk on tasks associated with identification, characterisation, and sizing and localisation of indications in all four methods. This is not surprising, considering that these subtasks—if carried out incorrectly—may lead to an incorrect assessment of the criticality of potential structure-breaking defects. Whereas *identification* is typically singled out as the most critical task in RT, UT, and rVT, in ET the situation is different: Data evaluation task in ET is assisted by an automated aid, i.e. software

that is pre-programmed to identify indications and record their size and location, whereas the inspector's task consists of controlling of the results. This was why less potential risk was associated to identification, sizing, and localisation, as stated by the participants. Deciding whether an indication is a reflection of an actual material defect, or of the material's geometry (e.g. a corner); or deciding about the defect type (e.g. volumetric vs. planar-type defect, porosity, crack etc.) are still allocated to the human inspector and assigned the highest risk priority in ET.

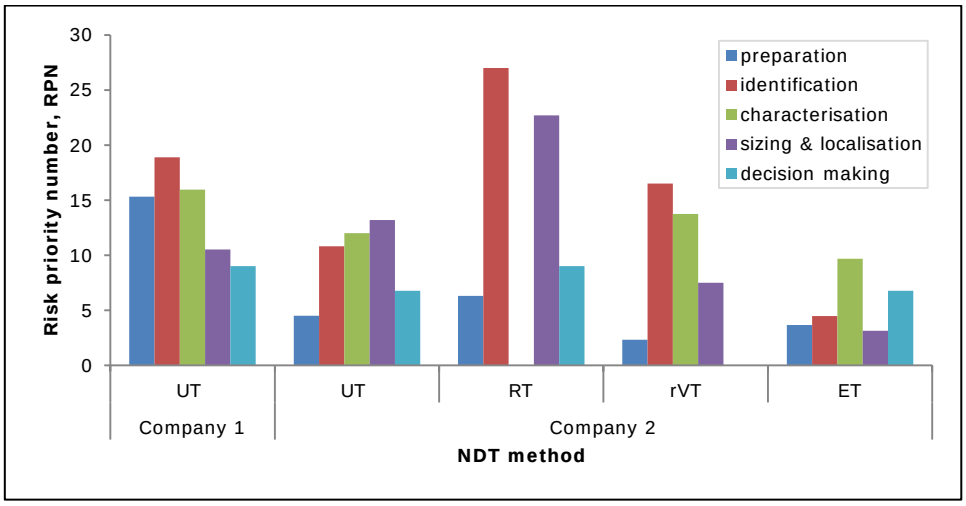


Figure 14: Risk priority rating for subtasks in the evaluation of data collected with ultrasonic (UT), radiographic (RT), remote visual (rVT), and eddy current testing (ET)

The decision-making tasks were assigned a rather low risk priority across tasks. This is because this task refers only to a *recommendation* about the structural integrity of the component and not the final decision itself.

3.8. Discussion

The aim of this study was to identify potential risks during mechanised NDT for the spent nuclear fuel management and find methods that can counteract their effects.

This chapter will start with a summary and interpretation of the results of the conducted analyses (section 3.8.1) with focus on the preventive measures and their possible practical downfalls (section 3.8.2). It will continue with limitations of the study (section 3.8.3) and conclude with the selection of topics for further empirical study (section 3.8.4).

3.8.1. Summary and interpretation of the results

The presented results confirmed the assumptions made before the study. First, they showed that there is a chance for failure in mechanised NDT during both the acquisition and the evaluation of data. This was illustrated by a number of identified potential failure modes and their effects. Second, the consideration of the potential error causes showed that next to the technical factors, human and organisational factors play an important role throughout the

entire process and could give rise to failure. It also indicated that for sources of error one should look beyond the person carrying out the inspection task—in line with the current safety practices. Third, the list of preventive measures, relying up to the point of the analysis on experience, training, qualification, and inspection procedures, has been expanded following the results of the analysis.

The analysis has shown that NDT is most frequently associated with errors of omission (e.g. missing defects) and commission (errors in selection and errors during the preparation for the inspection and the inspection process itself). Even though the errors of commission might be more prevalent during the single subtasks, a failure to detect a defect (omission) is of major concern. Omissions are frequently associated with failures in maintenance, which NDT is a part of. According to Rasmussen's (1980) analysis of 200 event reports, 33% of all omissions in maintenance happen during testing and calibration, i.e. NDT. Positive effects in reducing omissions were shown by providing with cognitive aids such as check-lists and by improvements in the training (Reason & Hobbs, 2003; Reason, 1997).

The most prevalent identified latent conditions present almost at all stages of the NDT task include error-enforcing conditions, hardware, design, procedures, and training. The focus on technology (hardware, design) and on training is consistent with the stage of development of the NDT methods. The participants, i.e. the developers of the NDT methods and techniques, are expectedly interested in the improvement of the technology and are aware of its current shortcomings. This is, also, where the most improvements are expected, since continuous effort is being invested in developing the procedures and the equipment, as well as considering necessary requirements for the future training. Since failures are difficult to be detected and understood in isolation from their context, some *now* not so salient latent conditions, i.e. those related to the organisation, may have a greater role when the operation of the fuel management starts. Considering that most errors have their origins in managerial and organisational actions or inactions (Reason, 1993), it is reasonable to assume that organisation, incompatible goals, communication, or housekeeping will merit more attention once the operation starts. Inadequate tools, unworkable procedures, design deficiencies, poor communication, and housekeeping are cited as some of the most influencing local-error provoking factors (Reason & Hobbs, 2003). Holstein et al. (2014) proposed that the reliability of NDT can be affected by not only the internal organisational context (i.e. the business, information, and the delivery processes), but also by the external organisational context (i.e. regulatory practices, technical rules, social/technical rules, safety culture, and the market itself).

The risk priority rating indicated a number of tasks that bear substantial risk that NDT could fail in its assignment and lead to wrong conclusions about the structural integrity of the material. The highest risk priority number was assigned to tasks associated with the evaluation of data, i.e. the evaluation of the calibration data, identification, characterisation, and sizing and localisation of indications. This was in line with the expectations, considering that the detection and characterization of defects are frequently cited as the most critical tasks of inspection (Norros & Kettunen, 1998; Norros, 1998). Whereas failures in the preparation phase could still be detected with some consecutive steps, the failures in data evaluation could remain undetected, if no complimentary methods are provided.

The discussion on possible preventive measures yielded a number of ways some of the identified risks could be prevented. Some of the most salient preventive measures include the improvement of the inspection procedure and instructions, implementation of human redundancy, and a number of hardware and software improvements, including further automation of the process.

3.8.2. Critical reflection on the preventive measures and outlook

Failures are contained and prevented not only by installing defences, but also by identifying gaps in the defences (Reason & Hobbs, 2003). As discussed during the FMEA, the existing barriers might be insufficient to prevent the potential failures: For example, the task is to a high extent aided by automation, but errors do occur; inspection procedures and instructions do exist, but are flawed and need optimising, and so on. This indicates that the existing barriers might need improving and that having barriers alone does not necessarily prevent from failure. Considering the potential paradoxes associated with protective barriers (e.g. Bainbridge, 1987; Dekker, 2002; Reason, 1997), three suggested measures will be further discussed: improvement of the inspection procedures and instructions, human redundancy, and automation.

3.8.2.1. Inspection procedure

Inspection procedures and instructions are some of the most important tools in the everyday working life of an NDT inspector. They are typically written by certified personnel in accordance with standards, codes, or specifications. During the FMEA, the procedures and instructions were identified as a potential error cause, and their optimisation as a potential barrier. Failure was generally assigned to insufficient content or to the inspectors not following the procedure, and the suggestions made for its improvement were focused mainly on its content.

Operating experience and research over the years have shown that procedures and instructions are not always used properly and, thus, might need to be optimised. In his analysis of scrams (emergency shutdowns of the nuclear reactor) and LERs (Licensee Event Reports) in Swedish nuclear power plants in the period 1995–1999, Bento (2002) reported that 15% of all scrams and 31% of MTO-related scrams as well as 10% of all LERs and 25% of MTO-related LERs occurred due to procedural deficiencies. Of all LERs, 23% were related to testing activities. Deficient procedure content was assigned to 70% of the procedure-related LERs and 85% of the procedure-related scrams, followed by missing procedure and missing updates. Lack of adherence to the procedure was the most important contributing cause of LERs. Procedure-related events were more related to maintenance, testing, and modification tasks (74%) than to operational tasks (20%). In Gaal et al.'s (2009) study on human factors influences on manual UT inspection performance, a procedure that was written by a highly experienced and qualified writer, was not entirely understood by the users. After improvements have been made together with the participants, they reported higher satisfaction. The research initiative *Programme for the Assessment of NDT in Industry - PANI*, revealed that each inspector applies the procedure differently and that the inspectors do not necessarily read the full procedure or apply the procedure as intended by the procedure writers (McGrath et al., 2004; McGrath, 1999). In the PANI 3 study (McGrath, 2008), a review of the procedure from a human factors perspective was completed to identify improvements that may encourage the full use of procedures during inspections. Issues such as length and structure, content and presentation of information, procedural steps, procedure format, and record keeping were addressed in detail. The author suggested that the inspection procedure is central to a reliable inspection, and as such needs to be written in a way that not only contains all the relevant information but also supports their systematic application. For that purpose, the procedures need to be developed together with the user.

With this in mind, it becomes clear that attention should not be given *only* to the procedure content, but also towards its usability. Hence, the suggested approach to further development of the inspection procedure is to direct focus on understandability of the procedures and on

the format in which they are presented to the inspectors, in hope that procedures will be used and will be used appropriately.

On another note, adding more procedures has been a frequent engineering approach to deal with human variability. However, most organisations have a problem between accepting a bit of human variability and highly regulating the activities of its members. The risk with the procedure is that of over-specification that tends to lead to routine violations. Violating the procedure does not necessarily lead to an accident, but the tendency to violate increases the possibility of an adverse event, especially in the case of breaching safe operating procedures (Reason, 1995). Hence, NDT community should weigh the advantages of the extent of the procedures and of investing effort in a better training of the personnel. Furthermore, as the procedure is written by *any* inspector with sufficient qualification, it may be useful to develop a unified approach to its writing and guidelines that would improve its usability.

3.8.2.2. Human redundancy

One of the suggested methods to detect possible errors during data evaluation was to introduce human redundancy. The suggestions include random checks by the supervisors and a repeated inspection/evaluation by another inspector once a critical defect had been found.

Although human redundancy is generally used to increase reliability, its implementation can also carry risks, especially when the principles of technical redundancy (i.e. two independent systems that perform the same function) are applied to social systems. One of these risks is *social loafing*, i.e. the phenomenon of investing less effort when working on tasks collectively than when working alone (Karau & Williams, 1993). Whereas technical systems are assumed to function independently of each other, this scenario is often not the case with regard to social systems (Sagan 2004). Swain and Guttmann (1983) indicate the checker's familiarity with the inspector, who had already conducted the task, and his or her knowledge of the other inspector's technical level as some of the factors influencing human redundancy. Clarke (2005) added that a checker might fail to perceive an error because of a belief in the colleague's competence.

In the case of the management of spent nuclear fuel, in which the demand for the inspecting personnel will be low (at least in the first years of operation), independence might be difficult to achieve. In small companies, such as the two investigated in the scope of this study, but also in other inspection companies active in other domains, inspectors are highly likely to know each other and be aware of each other during the inspection. This raises concern with respect to independence. The latest studies have indicated that due to social loafing effects human redundancy might not necessarily be an effective safety measure in working with automated systems (Manzey et al., 2013; Marold, 2011).

Taking this into consideration, the implementation of human redundancy in NDT requires further consideration with respect to potential negative effects that can outweigh the benefits expected from redundancy.

3.8.2.3. Automation

Further automation of parts of the data acquisition and evaluation tasks was frequently suggested during the FMEA. The benefits of automation in form of a bar-code reader were especially seen in data acquisition, as a result of which mistyping errors or opening of the wrong setup file could be avoided. In data evaluation, automation was identified as a potential aid in identification and characterisation of indications. An example of existing automated aid is the software used for the evaluation of data with ET. This software aids in the evaluation by

automatically detecting and sizing the indications, whereas the role of the inspector remains to control the results.

The major goal of introducing automation into the working environment is to reduce human error (Skitka, Mosier, & Burdick, 1999). The advantages and disadvantages of automation with regard to the reduction of human error have been widely investigated in various industrial applications, including, among others, aviation, healthcare and the military (e.g. Alberdi, Povyakalo, Strigini, & Ayton, 2004; Bahner, Hüper, & Manzey, 2008; Dzindolet, Dawe, Beck, & Pierce, 2001; Lee & Moray, 1994; Madhavan, Wiegmann, & Lacson, 2006; Manzey, Reichenbach, & Onnasch, 2012; Mosier & Skitka, 1996; Parasuraman & Manzey, 2010). Despite its many benefits regarding processing speed and accuracy, and reduction of human error to some extent, automation has shown to lead to new error sources and new risks in ways that are unintended and unanticipated by the designers (Bainbridge, 1987; Parasuraman & Manzey, 2010; Parasuraman & Riley, 1997). This is because automation does not necessarily replace human operators; rather, it changes what they do. Note that automation mentioned here is not meant to replace the human operator completely, but rather aid the operator in carrying out selected tasks. For example, instead of manual control, inspectors now need to cope with the complexity of constantly developing technology and multitasking, by relying on what the equipment tells them. This can occasionally result in overlooking errors with regard to the functioning of the automated system, thereby leading to errors that may compromise safety. An uncritical reliance on the proper functioning of an automated system without recognising its limitations and the possibilities for failures often occurs when the task demands are too high and when the automated system is perceived to be reliable and is trusted (Lee & Moray, 1992; Manzey, 2012; Mosier & Skitka, 1996; Parasuraman & Riley, 1997). Considering the perceived superiority of automated systems in NDT, i.e. higher perceived reliability of mechanised over manual NDT, uncritical reliance could be one of the automation ironies associated with NDT. Nevertheless, automation offers many advantages, with regard to reducing human error. When functioning properly, automation saves time, decreases workload, and generally reduces human error. The key for NDT is to be aware of the potential errors that arise from this interaction and to find means of avoiding them.

In conclusion, implementing preventive measures is a process that requires detailed consideration. Risk assessment and risk treatment are cyclical in nature. To ensure the highest profit, the FMEA should be repeatedly applied to identify new risks that can arise over time resulting from, e.g. implementing barriers or from the changes in the way NDT inspections are carried out.

As a result of this study, several changes have already been made to the way NDT inspections are carried out, as well as with regard to the used equipment and the inspection procedures. These are shortly summarised below.

Digression: Implemented preventive measures/barriers as a result of the FMEA

The major changes include further development of own evaluation software and specification of the requirements for the new equipment and the evaluation software. Some of these specifications include equipment that is more reliable, improved presentation of information and access to information on the screen, and the evaluation of data in static, as opposed to dynamic, mode. Some parts of the inspection task have been automated, i.e. the identification of indications in eddy-current, and the reporting of indications in ultrasonic testing. In remote visual testing, the viewing parameters such as the magnification or speed of the component's

rotation have been further optimised and in radiographic testing, a defect catalogue to help with the characterisation of indications is under development.

Convincing evidence from the literature, coupled with the results of the FMEA, inspired a study into the further development of a selected NDT instruction, as well as its optimisation with regard to its content and format using a user-centred approach (by means of eye tracking, individual and group discussion with the users, among others). In a follow-up study, instruction content was subjected to questioning its understanding by experienced readers, and the format to the questioning of its usability. The study resulted in identifying a number of factors that contribute to a high-quality instruction, regarding both its content and usable interface. The results showed that the understanding of the information could be affected by the information order, information organisation, logics and clarity in the writing, and by cognitive demands it poses on the reader. Higher efficiency in use was achieved by clearly distinguishing relevant information (e.g. warnings, deviations, reminders) from the remainder of the text and by presenting the tasks in a stepwise manner, with one action per step (Bertovic & Ronneteg, 2014). The changes also yielded higher user satisfaction and higher effectiveness in locating the information in the instruction.

3.8.3. Limitations of the study

The primary limitations of the study can be found in the fact that the analysed NDT methods are under ongoing development, thus lacking field application and experienced participants. On the other hand, this approach was valuable in aiding in further development of the methods.

Further limitations refer to the method. For example, the participants expressed using different criteria in assigning risk priority when evaluating consecutive methods (in some cases, the participants reported in hindsight that they may have underestimated or overestimated the risk of the previous tasks). Hence, it was decided not to compare the ratings for *different methods* quantitatively, i.e. with respect to the magnitude of their differences, but rather to assess them qualitatively and use the highest ratings as indicators for subtasks that require primary attention.

The FMEA's RPN is criticised to be overly subjective and not comparable. If a team were to be composed of different members, the rankings might be different (van Leeuwen et al., 2009). Rhee and Ishii (2003) go as far to say that RPN is meaningless, being that the three indices used for RPN are ordinal scale variables, which preserve rank but the distance between the values cannot be measured since a distance function does not exist. In addition, it can be influenced by the social group processes between members, organisational policies, and organisational norms.

The FMEA in combination with the criticality assessment, i.e. the FMECA, has shown to be a valuable tool for identifying and evaluating potential failures in NDT. Still, there are restrictions to this method that need to be mentioned. For example, the FMEA is suitable for identifying single failure modes, but lacks the combinations of different failure modes and could be difficult to conduct for multi-layered systems (IEC/ISO 31010:2009). Hollnagel (2008b) points out that explanation for risks cannot always be found in single components of the socio-technical system, such as the operator or the technology, but can also stem from their interaction or normal variability in human performance combined in unexpected ways. Thus, future attempts to assess risk in NDT should also include interactions between different systems.

Next, unless adequately controlled and focused, the analysis can be time consuming and costly (IEC/ISO 31010:2009). Difficulties can be encountered when there is insufficient data to

evaluate, if the leader is biased and not well trained, or if the participants' experience with the problem at hand is variable (Wetterneck et al., 2004).

Benefits of the FMEA are seen in the multidisciplinary approach and in its ability to identify failures early in the design. In addition, it is easy to understand, highlights safety critical tasks that require attention, provides with countermeasures to prevent failure in the future, etc. (Dhillon, 1999). All these benefits had made FMEA a suitable choice for assessing and treating risks in NDT in an application still under development.

Prospective approaches, such as that of the conducted FMEA, have been criticised for focusing mainly on the technology and the individual, in comparison to event analyses that have the ability of taking into account team, organisational and environmental factors (Fahlbruch, 2009). In this analysis, teamwork and extra-organisational environment were not considered due to their current non-existence. However, according to Fahlbruch (2009), customising the FMEA for the identification of human and organisational contributions to failure can be close to that of the event analyses.

Numerous alternatives to FMEA exist. It remains to be seen which of those methods is the most suitable for the purposes of identifying risks in mechanised NDT in the management of spent nuclear fuel. The process might differ greatly from the process today and new risks can and probably will arise. Other or new prospective and retrospective approaches can be used, and their suitability should be decided based on the relevant criteria at the time of the analysis. Risk and error management work best if both proactive and retroactive methods are combined (Hollnagel, 2008a; Latorella & Prabhu, 2000).

In conclusion, risk management is dynamic, iterative, and responsive to change (ISO 31000, 2009). For it to be successful, the need for risk management has to be recognised, the risks need to be identified, the underlying mechanisms of their effects understood and, finally, measures have to be taken for the risks to be successfully treated. This is most frequently achieved by preventing something unwanted from happening or by protecting the organisation from its consequences (Hollnagel, 2008a). This study raised questions regarding the suggested protective measures, which is why a deeper look into the potential implications of their implementation is needed.

3.8.4. Selection of the topics for the empirical study and research questions

As discussed in this chapter, the FMEA has shown a potential for failure in mechanised NDT, but also raised motivation for installing preventive measures.

As discussed in this chapter, the FMEA has shown a potential for failure in mechanised NDT, which raises motivation for installing preventive measures. As elaborated in the discussion, the preventive measures have to be carefully implemented, keeping in mind potential new error sources that come along with them. With this in mind, two empirical studies were initiated: one concerned with the human redundancy and the other with the use of automated aids in the evaluation of data.

Ultrasonic testing (UT) is one of the most frequently used NDT methods to inspect the volume of thick components. As such, there is a high demand that this method provides highly reliable results and continuous efforts are invested into its optimisation. The risk priority rating of the UT (Figure 14) declared *identification, characterisation,* and *preparation* as the most critical tasks in data evaluation. In discussion with the experts about prevention possibilities, the most frequently mentioned method was to implement human redundancy

and rely on the premise that if one person missed a defect or misinterpreted it, the second one will most likely not. As elaborated in the discussion section, the benefits of human redundancy are closely related to the degree of independence between the two inspectors. When independence cannot be guaranteed, it is questionable whether it can be profited from human redundancy. The notion of social loafing in redundant teams is no novelty to social psychology and to the field of human factors (e.g. Clarke, 2005; Latané, Williams, & Harkins, 1979; Skitka, Mosier, Burdick, & Rosenblatt, 2000; Williams & Karau, 1991), but a completely unexplored topic in the NDT field. To present this issue to the NDT community, the following research question was raised:

> Q1: **What happens when you *know* someone else had already conducted the task, or will do it after you?**

The aim of this question was to explore the effects of social loafing in sequential redundant teams and suggest the appropriate means for the reduction of the social loafing effects and for the appropriate implementation of human redundancy in NDT. This question will be explored in the first empirical study.

The ratings of the single subtasks of different methods have shown slight differences between the methods in the risk priority. To be more specific, eddy current method can be singled out as a method to which, in comparison to other methods, rather low risk priority was assigned. In addition, the identification of defects—a task usually assigned the highest priority in other methods—was assigned a rather low risk priority. One of the explanations for this, according to the experts, was found in the fact that ET data evaluation is aided by automated software and, hence, identification has been made more reliable. Considering that high perceived reliability of the aid has shown to lead to the tendency that the aid is more trusted, and hence, its directives uncritically followed (e.g. Parasuraman & Riley, 1997), the following question was raised:

> Q2: **What happens if an automated aid is highly trusted and it fails?**

With this question, the aim was to raise awareness of potential downfalls of inappropriate automation use and suggest methods to support appropriate use of automated decision aids in NDT. This issue will be explored in the second empirical study.

4. Empirical studies 2 & 3: Recruitment of the participants and the design of the experimental task

The two empirical studies that are going to be presented in the following chapters (Study 2: *Application of human redundancy in the evaluation of NDT data*; Study 3: *Use of automated aids in the evaluation of NDT data*) are the first studies conducted to explore difficulties encountered during the evaluation of data within the scope of a mechanised NDT inspection. Furthermore, the results of the studies are supposed to expand the understanding of the behaviour of NDT personnel in their actual working environment and, hence, be applicable to the NDT practice.

For that purpose, it would be optimal to conduct the studies with experienced NDT personnel carrying out their daily task. However, NDT methods that will be employed to inspect components used for the final disposal of spent nuclear fuel are under development, with only a handful of people familiarised with the methods and qualified to take part in our studies (2-3 experts per NDT method). Still, instead of observation, survey, or any other form of qualitative study, it was opted to carry out an experimental study to be able to profit from the advantages of the experimental control and establishing of the causal relations among variables. For that purpose, a larger sample population would be beneficial. Since recruiting NDT inspectors with actual field experience was limited for financial reasons, the acceptable alternative was found in recruiting NDT trainees, researchers, and NDT instructors (trainers) familiar with the methods of concern for the studies.

The evaluation of NDT data is a complex task, usually carried out by certified, experienced, and trained personnel with the aid of application-specific software. There is a large variety of available software for each NDT method, which each NDT service provider is free to choose from, or develop on its own. To use the actual software used in the spent nuclear fuel management would require extensive training and sufficient practice, before the participants would be able to carry out the task properly. Hence, the task needed to be simplified. For this purpose, the experimental task was designed using open–source software that does not require extensive practice, but still simulates the actual task well.

Due to the similarities in the applied method in the two empirical studies, the recruitment of the participants (section 4.1) and the design of the NDT data evaluation task (section 4.2) will be jointly presented in this chapter.

4.1. Recruitment of the participants

Altogether 154 participants took part in the two studies. They were sampled at research institutes, schools, and facilities for vocational training (Table 12).

Table 12: Institutes, schools, and training facilities from which the participants were sampled and the assignment of the participants to both empirical studies

Institute/school/training facility	Qualification	Participant count (N)		Total
		Study 2	Study 3	
Federal Institute for Materials Research and Testing, BAM, Berlin	Researcher	12	NA	7.8%
German Society for Non-Destructive Testing, DGZfP, Berlin	Trainer	9	NA	5.8%
Lise-Meitner School of Science, Berlin	Trainee	6	NA	
W. S. Werkstoff Service GmbH, Essen	Trainee	10	22	86.4%
Siemens AG – Siemens Processional Education, SPE, Berlin	Trainee	47	48	
TOTAL		84	70	100

The participants were predominantly NDT trainees (86.4%). All of the participants possessed knowledge and experience in NDT, even though to a different extent. For example, the majority of the researchers possessed sufficient theoretical knowledge and laboratory experience but lacked practical experience. Most of the NDT trainees, on the other hand, had practical experience, but somewhat less theoretical knowledge. Basic understanding of NDT coupled with practical experience and training provided by the experimenting team was sufficient for the understanding and the completion of the task.

Even though the experimental task is concerned with the evaluation of data acquired with mechanised NDT methods, the participants were not experienced in this task. This is because data evaluation is not taught during the initial training of the NDT personnel, as it is highly dependent on the used software that varies between inspection providers.

4.2. Design of the data evaluation task

In the scope of the empirical work, two NDT methods were investigated: ultrasonic testing (UT) and eddy current testing (ET). Even though these methods differ in their physical capabilities—one method is acoustic, the other electromagnetic; one searches for defects in the volume and the other at or near the surface; etc.—the data evaluation task bares similarities when narrowed down to detection and sizing of found indications.

The main task in data evaluation is to search for discontinuities in the material, characterise them (determine their size and location, and sometimes the type), and report the findings. Once an indication (a collection of pixels with colour intensity that differs from the background) is found, i.e. detected, it has to be further analysed by determining its size and location (sometimes the type). This is done *only* for those indications, which equal or exceed a predetermined *reporting threshold*. This value is usually determined based on careful consideration of what constitutes a potential threat to the structural integrity of the material,

influenced by a number of factors, e.g. the expected mechanical loads, the expected corrosion, inspection intervals, and so on.

Which apparatus was used to simulate the task and how the task was designed to be carried out will be presented in the following.

4.2.1. Apparatus

4.2.1.1. Software

The solution to overcome the obstacle of the variety of available software and the fact that working with it requires extensive training and practical experience was found in the use of a public domain Java-based image processing program, i.e. ImageJ, version 1.43u (Rasband, 2010). ImageJ possesses a large number of necessary features for image processing necessary to carry out this simulation of the NDT evaluation task, such as the area and pixel value statistics, measuring distances, labelling of indications and so on. It offers a *simplified* evaluation, considering that the evaluation of data usually requires a more complex and multidimensional approach.

4.2.1.2. Images

For the purposes of creating a realistic task, images resulting from actual UT and ET inspections containing real defects in the components were used. Those images are characterised by a colour-coded representation of pixels with varied intensity, i.e. each individual pixel has a unique intensity value presented by a corresponding colour that reflects changes in the homogeneity of the material. Figure 15 shows an example of an image of the component lid, acquired by means of eddy current testing and an example of an indication distinguishable from the background by its colour.

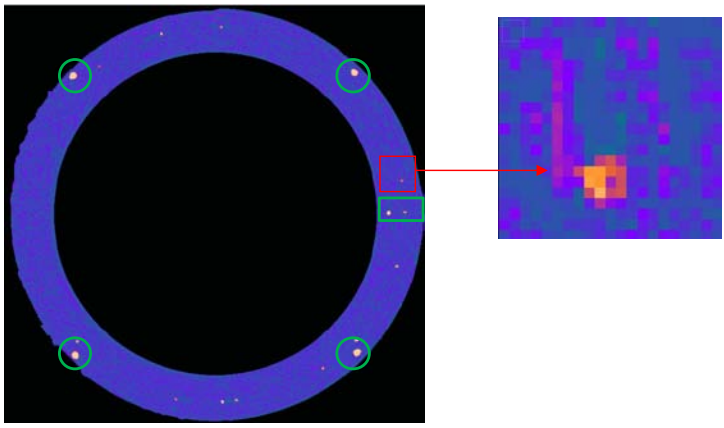


Figure 15: An example of an image of the canister lid (left) with an enlarged indication (right). Note: Marked in green are the indications resulting from screws used to lift and carry the lid (circles) and from the starting point of the measurement (rectangle), which the participants are taught to exclude from the evaluation.

Note that the signals collected by means of e.g. ultrasonic testing can be presented to the evaluator in different ways. In the scenario adopted in this study, the so-called C-scan presentations, i.e. images of the results of UT showing a cross section of the test object

parallel to the scanning surface (Schmitz & Mißmann, 2009), were used. Other presentations include A-, B- and D-scans (see Glossary).

Real images of the components also bear difficulties for the evaluator due to varying signal-to-noise ratio[10] (SNR): If low, it is hard to distinguish the signal from noise, which may lead to the signal not being detected; if high, the signal is clearly distinguishable from the background noise. Images with varying SNR were used in the study. The examples of (a) low and (b) high SNR in the case of UT are depicted in Figure 16.

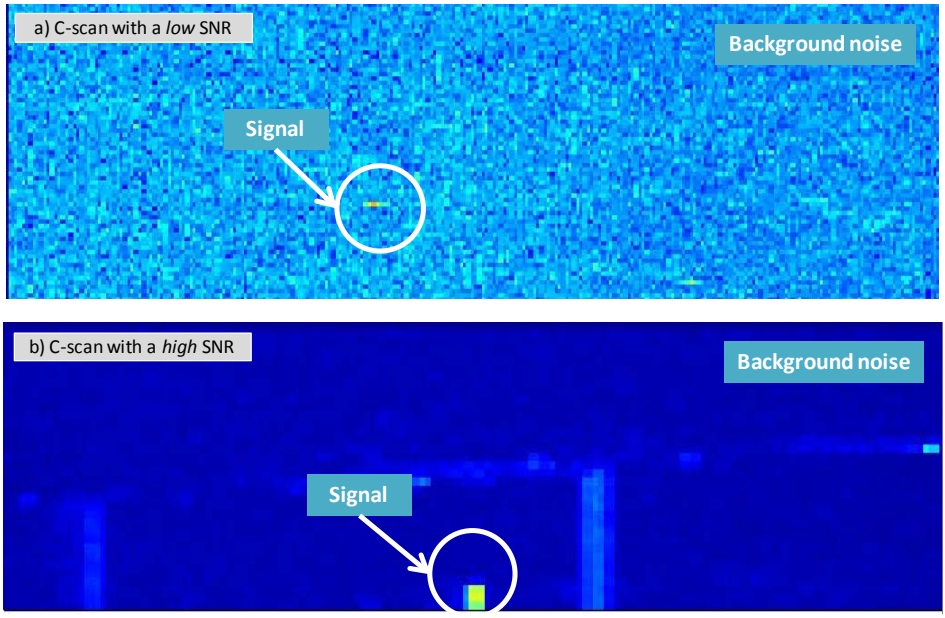


Figure 16: UT C-scans with different signal-to-noise ratio (SNR)

4.2.2. Procedure

4.2.2.1. Detection

The search for indications is carried out visually by searching for pixels that are distinguishable from the background. In order to determine whether the indication equals or exceeds the registration threshold, and, hence, needs to be further analysed and reported, the intensity of the pixels needs to be measured. To do so, the participant is expected to zoom into the area, mark the area around the indication by setting the contour around it, and measure the intensity of the pixel values, either by pointing with a mouse over the single pixels, or by using the "measure" function of the software. If the maximum pixel intensity value exceeds the registration threshold, the participant is expected to continue with the sizing of that indication, and if not, continue searching for other indications. The reporting threshold in the first study

10 The signal, arising from some kind of discontinuity in the material (e.g. defect, edge of the material), is usually distinguishable from the background noise. The closer the signal is to the background noise, the harder it gets to detect it, which is why establishing the signal-to-noise ratio is an important step in the evaluation.

was set to the pixel intensity value[11] of 100 for all images, and in the second study, it varied from image to image[12].

4.2.2.2. Sizing

Sizing of the indications is typically carried out by registering all those pixels that exceed a value of a predetermined *sizing* or *decision threshold*. Following this rationale, all the pixels with a value equal to or exceeding the decision threshold should be judged as belonging to the indication. This is achieved by setting a contour around the suspected size. An example of sizing can be seen in Figure 17.

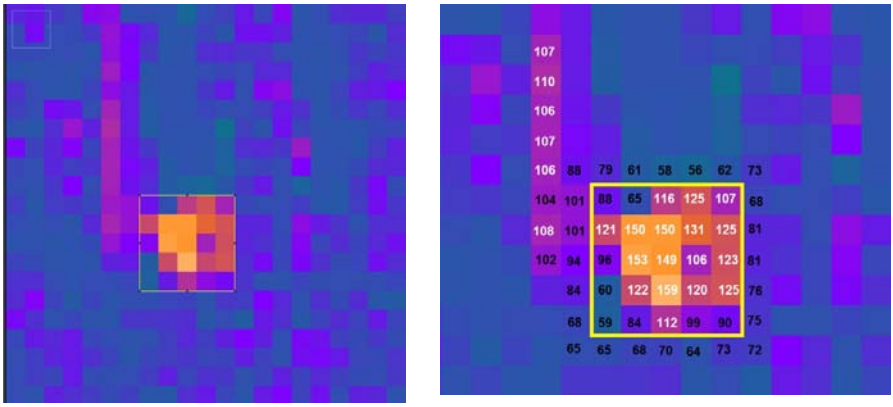


Figure 17: An example of indication sizing. Note: The figure on the left shows an indication with a marked size (yellow contour) and the figure on the right depicts detailed sizing based on the pixel intensity value. In this example, reporting threshold equals 135; the maximum intensity is 159; and the decision threshold is 106. This means that all those pixels, exceeding the value of 106 that are in the direct contact with the maximum intensity value are counted as belonging to the indication.

According to the sizing criterion conveyed to the participants, only those pixels in the *direct* contact (not diagonal) with the maximum intensity value should be judged as belonging to the indication. The employed sizing criterion is depicted in Figure 18.

4.2.2.3. Reporting

The position of the indication is automatically calculated after the participant draws a rectangle around the suspected area and measures it with the tool provided by ImageJ. Using the software function "measure" the maximum intensity, area, width, length, and the position coordinates of the marked indication are measured (Figure 19) and then copied into a reporting protocol (in this case, a spreadsheet document with exactly the same categories as the characteristics measured in ImageJ).

[11] Usually called grey value.

[12] The determination of the registration criterion can be based on a fixed value i.e. the size of the reference defect or on the signal to noise ratio. In the first study, the first criterion was used, and in the second, the second criterion. This difference is a result of different practices by the two companies, which provided the data for the studies.

The final step in the evaluation is to mark the indication (set a contour around its suspected size), label it, and save the image. This process is to be repeated until all images have been inspected, all found indications reported and appropriately saved.

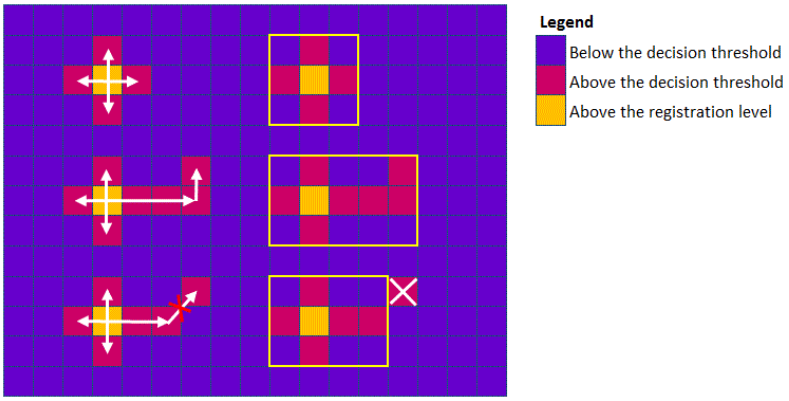
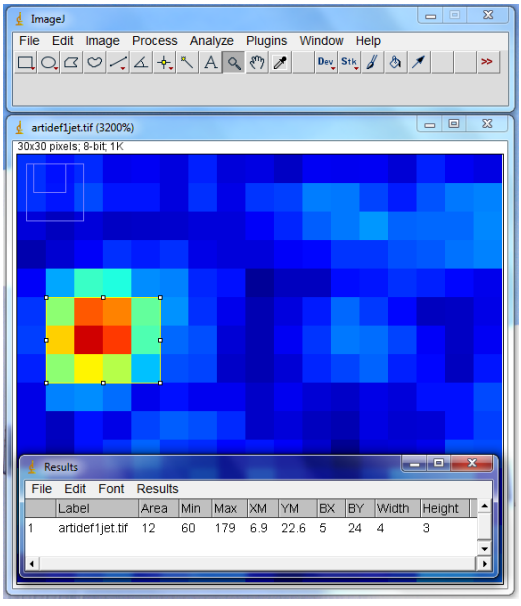


Figure 18: The sizing criterion. Left: Determining pixels that belong to the indication. Right: Correctly marking the indication. Only those pixels in direct contact with the pixels—horisontal or perpendicular—are to be judged as belonging to the indication. Those in diagonal contact are *not* belonging to the indication.



LEGEND

Label	name of the image, e.g. UT03
Area	the surface area of the contour
Min	the minimum pixel intensity value within the contour
Max	the maximum pixel intensity value within the contour
XM	the position of the maximum pixel intensity value in x-direction
YM	the position of the maximum pixel intensity value in y-direction
BX	the position of the upper left corner of the contour in x-direction
BY	the position of the upper left corner of the contour in y-direction
Width	the width of the contour (in pixel)
Height	the height of the contour (in pixel)

Figure 19: Marking and measuring the size of the indication.

5. Empirical Study 2: Application of human redundancy in the evaluation of NDT data

As elaborated in the third chapter, implementation of human redundancy to the NDT evaluation task could have benefits in increasing the likelihood that the outcome of the NDT data evaluation provides with accurate information about the true state of the component. However, with every change, new risks lurk. One of those risks—the risk of social loafing in the evaluation of NDT data—could decrease the likelihood that the error recovery mechanism of human redundancy will function as expected, thereby increasing the risk of missed defects. Hence, it is important to address this issue before human redundancy can be successfully implemented in the NDT practice. This chapter will present the first study of human redundancy in NDT and discuss the implications of the results for the NDT practice.

In order to identify potential roots of the risk, this chapter will explain the main principles of human redundancy and the reasons because of which it could fail (section 5.1). In the following, the effects of social loafing and social compensation in working in groups will be described, including the myriad of variables that can moderate the effects (section 5.2). The final theoretical section (section 5.3) will focus on the application of human redundancy in non-destructive testing, with examples of problems of sequential redundancy, finishing with its application in NDT, and the aims of this study (section 5.4).

In the empirical part of this chapter, two studies conducted to identify potential problems arising from human redundancy will be presented (sections 5.5 - 5.6), followed by a joint concluding discussion (section 5.7).

5.1. Human redundancy

Wherever there is an "*excess or superfluity of anything*" (Landau, 1969, p. 346), we talk of redundancy. The so-called redundancy principle relies on an assumption that the greater the number of redundant components is, the greater the reliability of the system will be, given that the components involved in the redundant system are *independent* of each other (Felsenthal & Fuchs, 1976). In simple words, if one component or a system fails or stops operation, having independent redundant components or systems will keep the system running. Typically applied in high-reliability organizations, the redundancy principle is used to increase the

system's overall reliability in monitoring the performance of technical processes (LaPorte & Cansolini, 1991).

Taught by experience with technical systems, the designers of complex systems (typically, high reliability organisations) decided to apply the principles of technical redundancy to organisational systems with similar aims, i.e. to increase reliability and as a human error recovery mechanism (Clarke, 2005; Swain & Guttmann, 1983). Assigning several individuals to solve the same problem or carry out the same task (fully or partially) avoiding, therewith, failures of single individuals, is the core principle of human redundancy.

Human redundancy comes in various ways and forms. The way it is implemented depends largely on the number of redundant operators and their inter-relations (e.g. operator–operator or operator–supervisor), on the extent of direct involvement in the task (active, or available on need), and on the degree of cognitive diversity among the redundant operators (Clarke, 2005).

Two redundant individuals is the simplest structure of human redundancy, followed by an operating team consisting of two or more operators and a supervisor; and an operating team and another individual, such as the shift technical advisor employed in civil nuclear power plants or an independent observer (Clarke, 2005).

Essentially, the task performed by two or more redundant individuals can be described as *duplication* (all performing the same function) or as an *overlap* (the task has some functional areas in common) (Landau, 1969; Lerner, 1986). Furthermore, Clarke (2005) distinguishes between active and standby redundancy. In *active redundancy* the individual fulfilling a redundant function is actively involved in the task (e.g. an operator fulfils a function, while another monitors the performance of that operator with respect to the required function). In *standby redundancy,* the redundant individual is called upon request to review past activities and contribute to future activities. In this case, the dependence between the team and the redundant individual is lower than in the active redundancy.

Human redundancy is also present when someone checks someone else's work. In this way, redundancy is frequently applied as a measure of error recovery, i.e. a measure of error detection, indication (bringing to attention), explanation (e.g. localisation of the error), and correction (Clarke, 2005; Swain & Guttmann, 1983). This kind of redundancy is known as the *sequential redundancy*, in which one redundant element carries out the task, followed by another employed to control (or check) the work and thereby detect the potential errors. If a person makes an error that he does not detect, the error may remain undetected until the results of that error affect the system's functioning in some way. If a second person checks the task performed by the first person, there is some probability that he will detect the error and correct it. If this happens, we can say that the recovery factor of human redundancy has happened.

The rules of technical redundancy are not always easily applied to people and its implementation is not always as straightforward as it might seem. For one, redundant individuals are not identical as redundant devices and the reliability of people is hard to estimate and difficult to improve (Felsenthal & Fuchs, 1976). More importantly, the main requirement for the redundancy to fulfil its purpose—that the individuals involved need to work parallel to and completely independent of each other (Clarke, 2005)—is not often achievable in human redundant systems. Unlike machines, individuals in redundant systems are aware of each other (Sagan, 2004) and, thus, independence between them is not necessarily a given. Hence, some kind of dependence between people must be assumed. The violation of independence opens doors to a number of social influences that can lead to a decline in

individual effort on group tasks, as opposed to working alone, and counteract the benefits expected from redundancy.

What are the reasons for that? According to Carroll (2004), "*in human systems, independence rapidly breaks down because people take action based on their beliefs about what other people are doing*" (p. 955). Beliefs, such as that other people are more competent, experienced, willing, or more apt to pick up the slack, can all influence individual motivation, and thus, behaviour in a task. Relying on adequacy of the other redundant system, may also lead to the diffusion of responsibility (Conte & Jacobs, 1997), which in turn can counteract the benefits of redundancy. These beliefs may have an even larger effect on performance in complex environments, in which the individuals are working on multiple simultaneous tasks and may be forced to choose on which tasks they should allocate their attention (Conte & Jacobs, 1997).

Schöbel & Manzey (2011) suggest that interpersonal dynamics in social redundant systems represent blind spots that can "*easily outweigh any positive effects on system reliability and safety*" (p.51). They talk of two motivational interdependencies between social systems not present between technical components that could cause individuals to conform to others, and consequently, human redundancy to fail. On the one hand, it might fail due to normative and informational social influences, i.e. conforming to the expectations of others because of the desire to obtain approval and avoid rejection (normative social influence) or because of a belief that the opinions and decisions of others can improve own decisions and judgements (informational social influence). On the other hand, it can fail due to the effects of social loafing, i.e. a reduction of effort in tasks when working collectively, as opposed to working alone (Latané et al., 1979).

5.2. Social loafing and social compensation

The problem of social loafing was first introduced by a German student Ringelmann in 1913. On a simple game of rope pulling, he showed a decrease in the group performance over individual performance depending on the group size. Assuming that individuals acted upon the rope at 100% of their ability, Ringlemann observed a performance decrease of 7% for dyads, 15% for triads, and up to 51% for groups of 8 people of their potential ability, as reported by Ingham, Levinger, Graves, & Peckham (1974). In an attempt to verify Ringlemann's findings, the first scientific study carried out by Ingham et al. obtained a similar result (measuring the strain used to pull the rope using a strain gauge). They observed a drop in the individual performance for dyads and triads (with no difference for groups larger than 3), even when they only led participants *believe* they were working in groups of varying size, when in fact they were pulling the rope alone.

Since those initial discoveries, numerous studies provided evidence of social loafing in a variety of tasks, both physical (e.g. Harkins, Latané, & Williams, 1980; Latané et al., 1979) and cognitive (e.g. Harkins & Petty, 1982; Price, 1987; Williams & Karau, 1991). Furthermore, the social loafing effect has shown to be stable across genders and cultures (even though the effect is somewhat smaller for women and Eastern cultures), and it was found in optimising and maximising tasks, in real and imaginary presence of others, and in between and within-subjects designs (Karau & Williams, 1993).

Social loafing-like effects have also been discussed in the economics literature, concerned with individuals withholding effort in organisations. Referring to "*commonly-experienced phenomenon of individuals or groups acquiring more than their fair share of the benefits of other people's efforts*" (p. 123),

Albanese & Van Fleet (1985) defined the term free-riding. According to their understanding, free riding is a phenomenon characterising rational individuals directed at achieving private goals. They will contribute to the public goals if it is in their self-interest, if they are more likely to receive something in return, and if their contribution will significantly influence the amount of public good to be shared. The free riding tendency increases with the group size. Even though frequently associated with laziness, irresponsibility, and selfishness, free riding is a natural tendency driven by a desire to receive benefits without having to sacrifice more resources than necessary.

Another similar term is that of social shirking, i.e. reduction of individual effort because of a belief that others will "take up the slack" (Sagan, 2004). Withholding effort due to social shirking can be assigned to mismatching interests of the individual and the organisation (Judge & Chandler, 1996) or to employee's perception of "fairness" of the reward systems (Bennett, 2004). Awareness of other redundant individuals can decrease system reliability if it leads an individual to shirk off unpleasant duties or responsibilities because of an assumption that someone else will take care of the problem (Sagan, 2004). Job satisfaction and general life satisfaction seem to counteract shirking (Judge & Chandler, 1996)

However, participants will not necessarily loaf, free ride, or shirk off responsibilities when working in groups. Under some conditions, individuals will invest *more* effort in order to compensate for other members in the group, an effect known as *social compensation* (Williams & Karau, 1991). This effect is likely to occur if the group outcome is relevant and meaningful to the individual and if one expects the other group members to perform poorly (either due to the general lack of trust in others or due to direct knowledge of other's insufficient abilities or efforts). The motivation for such behaviour could be a result of altruistic tendencies, i.e. a desire to protect others from poor evaluation; a need for self-validation through the validation of the group; or due to *something to gain, nothing to lose*-motivation. I.e. a poor group evaluation is expected to lead to poor external evaluation and can be blamed on the co-worker; but for good evaluation the individual can take credit for (Williams & Karau, 1991).

One of the most influential explanations for both social loafing and social compensation was provided by Karau & Williams (1993) in a unified framework dubbed Collective Effort Model (CEM). According to CEM, individual motivation in a group context depends on three factors: *expectancy* (extent of effort expected to lead to high levels of performance), *instrumentality* (the extent to which high-quality performance is instrumental to achieving the desired outcome) and *valence* (the extent to which the outcome is seen as desirable). Compared to individual motivation, where instrumentality is guided by a degree to which individual performance is instrumental for obtaining a desired outcome, the CEM suggests that working in a group setting requires additional contingencies between individual's efforts and individual outcomes. Those contingencies relate to three perceived relationships, between: a) individual performance and group performance, (b) group performance and group outcomes, and (c) group outcomes and individual outcomes.

5.2.1. Social loafing and social compensation moderators

The motivation for extensive study of social loafing has been to identify ways to reduce or eliminate the effect. Since Latané et al. (1979) introduced the effect into social psychology, a myriad of moderating variables has been identified. For example, Williams, Harkins, & Latané (1981) discovered identifiability to be one of the most important moderators to the effect of social loafing. When individuals in the group are not identifiable, they tend to loaf. However, when they become identifiable, they tend to work as hard as when working alone. Some

understanding of how lack of identifiability affects social loafing can be simply derived from common-talk expressions, such as *lost-in-the-crowd* or *neither-credit-nor-blame*.

Identifiability might not be the only way to eliminate social loafing. The knowledge one's contribution to the task could be evaluated and compared either to some subjective or objective standard, e.g. compared to the contributions of other group members, is a strong motivator not to loaf (e.g. Harkins, 1987; Szymanski & Harkins, 1987).

Expectation of co-worker's performance has a strong influence on social loafing and social compensation. If the co-worker is expected to be unreliable, unwilling, expected to perform poorly, or unable to contribute to the outcome, individuals will compensate for him, given that the task is meaningful to them (e.g. Karau & Williams, 1993).

Next to increased identifiability, evaluation potential, and the expectation of co-worker's poor performance, the following were identified as potential countermeasures to social loafing that could give rise to social compensation: strengthening of group cohesiveness, increasing the difficulty of the task, and thus, making it more challenging, having a unique contribution to the task (inducing a feeling that their unique skills and talents are required for the task to be completed successfully), increasing personal involvement in the task, thus making someone's contribution meaningful, having a task that is meaningful, or increasing the responsibility for the task (e.g. Karau & Williams, 1993). The study of George (1992) suggested task visibility and intrinsic involvement as mediators to social loafing in the real-work settings.

Based on these findings, it can be concluded that social loafing and social compensation are well-investigated phenomena, especially in simple laboratory tasks. The next step in building up the knowledge requires the researchers to step out of the laboratory and into the field.

5.2.2. Studies of social loafing in real work contexts

There have been only a few studies including actual working teams in complex environments, especially in human-automation interaction. Note that these studies have been conducted on student populations but simulating actual industrial tasks.

For example, Skitka, Mosier, Burdick, & Rosenblatt (2000) examined whether in a flight simulation task two-member crews are less likely than single individuals to rely on faulty automation's cues, or lack thereof. The results showed that the presence of another operator did not increase the chance of responding to faults that the automated aid did not indicate, nor did it decrease the likelihood of incorrectly following aid's faulty directives. This showed that crews are not better than individuals in avoiding errors arising from human-automation interaction. Hence, assigning the same task to several individuals has not shown to be a good countermeasure against overreliance on automation.

Marold (2011) extended the study of Skitka et al. (2000) by further analysing performance losses in automation monitoring potentially caused by human redundancy. She postulated that redundant individuals would monitor an automated system less carefully than the non-redundant ones, in line with the social loafing theory, but that this effect should be reduced if the participants are informed about the limited performance abilities of the team partner, in line with the social compensation theory. As expected, participants excerpted less effort when working in a redundant team as opposed to working alone. This was found for the Redundant (responsibility for the task is equally shared between the two team members), and the positively Informed-Redundant individuals (the same as *Redundant*, but the participant is told that the team partner is motivated to perform well on the task). The information about the team partner's possible loafing (negatively Informed-Redundant) lead to the participants

investing more effort in the task in order to compensate for the less adequately performing imaginary partner. Thereby, this study provided with further evidence of social loafing and compensation effects in a new domain, i.e. automation supervisory task.

Extending to Marold (2011) is the study of Manzey, Boehme, & Schöbel (2013), which furthermore explored the benefits and potential pitfalls of applying human redundancy in automation monitoring. In addition to Non-Redundant and Redundant conditions, they investigated whether logging individual monitoring performance and giving individual feedback would decrease the detrimental effects of social loafing on the automation monitoring performance in an imaginary team (Redundant-Feedback condition) and, hence, constitute a good measure to counteract social loafing effects. The results showed that, first, the redundant individuals reduced their monitoring behaviour, as opposed to non-redundant individuals. Second, this reduction raised the risk of missing a surprising automation failure. However, this effect was decreased when the participants expected their individual performance to be monitored and fed back to them indicating that assessing individual performance in a redundant team could lead to the reduction, if not elimination, of the social loafing effect. This effect could be explained through raising perceived accountability for own performance. The authors suggested that the explanations for the causes of these effects could be found in the Collective Effort Model (Karau & Williams, 1993), suggesting that the amount of effort is dependent on how instrumental individual effort is to the desired outcome, but that the result could also be a sign of social shirking of responsibility (Sagan, 2004).

These studies have shown that social loafing is not only restricted to simple laboratory tasks, but that it occurs in actual working groups. In addition, it appears that monitoring of automation will not necessarily profit from human redundancy, unless the redundant individuals are independent of each other, expect their contribution will be evaluated, and receive feedback on their performance.

The overwhelming majority of the social loafing/compensation studies and findings were carried out on an example of joint activity on a task, typically conducted in parallel, even though the presence of the other team member is in most of the cases imagined. A significantly less investigated field of study is that of the sequential redundancy, in which one individual carries out the task after another and checks his results.

Examples of sequential redundancy can frequently be found in the practice, and is also present in the field of non-destructive testing. Up to date, there have been no studies concerning human redundancy in NDT. Its application is usually proscribed in standards, codes and regulations, but the experience from the field reveals that human redundancy might not always be implemented as it should.

5.3. Human redundancy in non-destructive testing

5.3.1. Human redundancy in the NDT practice

Human redundancy is already applied in the NDT practice (primarily in the nuclear industry) as a tool for increasing reliability. Human redundancy, or the *four-eyes principle*—as it frequently referred to in NDT circles—relies on the belief that it is less likely for two (or more) inspectors to miss a potential discontinuity (Dickens & Bray, 1994).

That redundant NDT inspections need to be performed independently of each other has been in some countries proscribed by regulations. For example, German Safety Standards of the

Nuclear Safety Standards Commission (KTA), which regulate in-service inspections of the primary circuit components, e.g. the reactor pressure vessel components, state the following:

> "Manual ultrasonic tests shall be independently performed and be evaluated by the plant owner and the authorized inspector" (KTA 3201.4, 2010 [Section 10, paragraph 3], p. 30).

In the practice, this means that an inspector commissioned by the nuclear plant owner and an inspector in service of the authority will conduct two separate inspections, without insight into each other's results. The results of both are then jointly discussed until a sufficient agreement between both parties is achieved (minor deviations are allowed). However, it is frequently the case that the two inspectors know each other (NDT community is rather small); and both are likely to be aware of the previous inspection's results (from the previous in-service inspections). In the case of mechanised inspection of primary circuit components, the data acquisition process is not be conducted redundantly, whereas data evaluation may or may not be conducted redundantly, even though there are no regulations proscribing its need. The stated reason for refraining from redundancy during data acquisition is to reduce inspectors' exposure to high radiation levels. Another reason refers to the fact that the equipment had been qualified in presence of the authority. Hence, no other quality control measure is required. The data evaluation, on the other hand, may be redundant if the authority—in charge of overseeing the inspection—assesses that there is a need for it. If a critical defect had been found, the data may be partially re-evaluated. In this case, the independence may not be guaranteed, because the authority will have insight into the results of the previous inspection and evaluation [D. Schombach (TÜV Nord), personal communication, March 18, 2015]. For components outside of the primary circuit, which present lower risk for safety, redundancy is not prescribed (KTA 3221.4, 2013).

Even though stated in the standard, the application of human redundancy varies from nuclear plant to plant, and it is not always ensured that the entire inspections will be conducted by two inspectors and neither that they will be independent [A. Erhard (Reactor Safety Commission, RSK), personal communication, May 11, 2015].

According to NDT practitioners and experts from fields other than nuclear, e.g. chemical industry, human redundancy is rarely carried out by having two inspectors inspecting the same area twice, i.e. in a form of duplication. Nor is it ever an overlap. Most frequently, if a defect had been found, the supervisor, or an inspector from the authority, will go on site and check the results. However, in most of the times, the supervisor only checks the reporting sheet [A. Hecht (retired, former BASF) & N. Weidl (Butting), personal communication, June 19, 2014].

Human redundancy is rare in the NDT practice in Sweden (a country of relevance for this study). There are no official requirements for redundancy in the nuclear industry, including both manual and mechanised NDT methods. Typically, one inspector will seek assistance of another if he requires an advice or if the consequences of a failure are considered large. The way redundancy takes place can range from looking at the results to conducting the entire inspection, if the suspicion for a failure exists [U. Ronneteg (SKB) & B. van den Bos (DEKRA), personal communication, February 25, 2015].

From the mentioned examples, it can be concluded that redundancy applied in the NDT practice can take various forms. However, independence may not, be always assured and, most frequently, there will be some dependence between those carrying out the inspection. The independence, proscribed in the regulations (e.g. KTA), refers mainly to the independence between the inspectors in terms of objectivity and quality control – one

commissioned by the plant owner, and the other by the authority. The social influences and the possibility of their effects on the performance are largely neglected.

Considering the prevalent use of sequential types of redundancy in the NDT practice, further attention needs to be given to social influences in its implementation.

5.3.2. Social loafing in sequential redundancy

Conte & Jacobs (1997) examined the effect of social loafing in sequential redundant systems in a transcript-checking task. They manipulated identifiability and the redundant systems (working alone vs. working in a redundant system, i.e. with a peer, faculty, computer, or faculty and peer—who either preceded them or carried out the task after) and measured whether under those balanced conditions, the participants would identify the 12 errors implemented into the task. Thereby they measured the frequency of omission (not identifying an error), commission (identifying an error that was not present) and categorisation errors (identifying an error, but placing it in an incorrect category of the check-sheet they were asked to control). Among other things, they hypothesised that those working alone would make fewer errors than those part of a redundant team, and that redundant individuals will loaf *less* if their contributions are identifiable and loaf *more* if they perceive to be working with a highly reliable system (automation bias). All three assumptions were supported by the collected data. One of the results shows that the participants committed more errors when they perceived to be working with a reliable system, as opposed to working alone or working with a less reliable redundant system. These results suggest that individual's perception of the reliability of the redundant system affects performance. However, not all redundant systems will have poor performance. Conte & Jacobs (1997) suggest that performance can still benefit from redundancy if the individuals in the redundant system are preceded by a low reliable redundant system, in which case they will compensate for the lower reliable co-worker.

Even though this study was conducted in laboratory settings, with students and on a simple task, it has shown that principles of social loafing and social compensation apply to sequential redundant systems. Most importantly, it emphasises the importance of expectation of co-worker's performance as an important moderator of the loafing/compensation effects.

5.3.3. Expectation of co-worker's performance in sequential redundancy

In real working groups, the information about the other redundant element is not told, but observed. Often, redundant team members are familiar with each other, know each other or can at least estimate other's experience and knowledge. That kind of familiarity can lead a redundant individual to omit a check (failure of initiation) because of confidence in the individual concerned, or—following successful initiation of human redundancy—can lead to a 'perceptual set' that results in a failure to detect an error (Clarke, 2005).

According to Swain & Guttmann (1983), checker's familiarity with the operator, as well as his knowledge of the other operator's technical level are some of the most influencing factors on human redundancy. Clarke (2005) elaborates that a checker might fail to perceive an error because of a belief in the competence of a colleague. Inefficient or insufficient checking behaviour could also happen due to excessive professional courtesy between individuals of similar rank (Sasou & Reason, 1999) or due to high levels of interpersonal trust (Williams & Karau, 1991).

Swain & Guttmann (1983) and Clarke (2005) described the potential problems of dependence between the inspector and the checker in sequential redundant systems. If, for example, the checker believes that the first inspector's work is reliable, he may assume that his performance will be correct. This assumption or expectation generally reduces the checker's effectiveness, i.e. he might miss the inspector's error because he does not expect it, even when the error is clearly visible. If, on the other hand, the person being checked is relatively inexperienced, or from a different department, the checker might take extra care because he doubts the quality of his performance. In addition, inspector's knowledge that his or her work is subject to human redundancy may already influence the reliability with which an interaction is carried out (Clarke, 2005). Therefore, social factors may be important in both directions: in case of the redundant individual expecting his performance will be checked, as well as in case of the one checking.

In summary, the expectation of co-worker's performance is an important moderator of performance in human redundant systems and can influence both members of a sequential redundant team.

5.4. Aim of the study

The aim of this study was to explore potential decrements in performance caused by human redundancy in NDT and to provide with suggestions how to implement it optimally.

Human redundancy in the inspection of the components used for the spent nuclear fuel management (application in focus of this study) may be applied sequentially in two ways: either as a standby redundancy or as randomly assigned quality checks, as suggested during the FMEA. In both cases, two distinct roles and different role-associated problems could be identified. The first inspector, from now on referred to as the *redundant inspector,* may be aware of another inspector coming over to check his work, hereafter referred to as the *redundant checker*. Both roles differ not only in the order in which the task is carried out, but also in the task itself. Whereas the inspector conducts his task as if he were working alone, i.e. his task is to detect and characterise indications from the data collected with an NDT method (*identifying task*); the checker frequently receives the analysed data and is required to control whether the data is correct (*checking task*). The dependence between the inspectors is especially high in the latter case, which can only be emphasised by knowing something about the inspector, whose work is being controlled, and, hence, the expectation about his performance.

Considering that only a few canisters are planned to be sealed and disposed of per year, the demand for inspecting personnel will be low, at least at the beginning of the disposal. The inspection personnel will be either permanently hired by the utilities or contracted among the existing fluctuating NDT personnel, as is common practice in the nuclear industry. In both cases, some to strong familiarity between the inspectors is to be expected.

Taking into account that under the stated conditions independence between the inspectors is violated (the redundant inspectors might be aware of each other, know each other, work together, make decisions together and have the knowledge of the redundant inspector's results—all possibly leading to an expectation of his performance), it could be expected that the variety of factors that can moderate individual motivation could give rise to social loafing.

Both inspector roles—the inspector and the checker role—were examined in this study, conducted in two parts. First, the role of the redundant inspector was investigated by comparing non-redundant performance to the performance in an imaginary teamwork with

the redundant checker (Experiment 1). And in the second part, the role of the checker was studied by examining whether or not an information about the redundant inspector's superior experience will have a detrimental effect on the inspection performance in comparison to not having that information (Experiment 2).

5.5. Experiment 1: Role of the redundant inspector

The aim of the first experiment was to determine whether the knowledge one was a part of a redundant team could negatively affect the performance.

5.5.1. Hypothesis

In accordance with Clarke (2005), who suggested that the mere awareness that one's work could be subject to human redundancy could have an effect on the reliability of the redundant system; and the empirical evidence that and the meta-study of Karau & Williams (1993), which suggested that social loafing occurs even in the imaginary presence of others, the following was postulated:

> **Hypothesis:** Individuals—led to believe they are working in a team with another individual, who would carry out the task after him—will loaf, in comparison to those individuals, who were told they would be working alone.

5.5.2. Method

5.5.2.1. Participants

The sample consisted of 32 participants. Three participants were excluded because they did not complete the entire task. Therefore, the analysis was conducted on 29 participants (all male; average age: 26 (18-46) years). The sample consisted of 6 researchers and 23 NDT trainees.

5.5.2.2. Apparatus and tasks

The experimental task was a simulation of the NDT evaluation task carried out with a computer, i.e. the participants were instructed to look for indications above a given registration level, size them and report the findings (described in detail in the previous chapter). To do so, they were provided with twenty C-scan images of defect indications in the copper canister component acquired by means of phased array ultrasonic testing. They contained from none to a maximum of six indications per image, summing up to altogether 37 indications[13]. During the task, carried out in ImageJ image-processing and analysis software (Rasband, 2010), the participants were aided by a short written NDT instruction containing all the relevant steps to be followed and by a list of shortcuts and key combinations to ease the work with an unfamiliar software. In addition, the participants received empty spreadsheets for reporting of the detected indications.

[13] The number of the indications was determined by an inspector, who carried out the original inspection and data evaluation, and confirmed by an algorithm, programmed to detect all pixels exceeding the predetermined evaluation criterion.

5.5.2.3. Design of the experiment

To investigate whether holding the first position in a sequential redundant system would lead to social loafing, the same participants (within-subjects design) carried out the task twice: once being instructed that they are working alone (Non-Redundant Inspector, nRI) and once being instructed that they are working in a team with a partner with whom they cannot communicate and who will carry out the task after them (Redundant Inspector, RI). The experimental manipulation was achieved by means of a written instruction.

In order to exclude the effects of learning and other sources of unsystematic variation, the order in which the participants took part in the experimental conditions was balanced. In addition, the order of appearance of images for each participant and in each experimental condition was randomised.

5.5.2.4. Dependent variables

The performance measures included typical measures of quality of an NDT inspection performance, i.e. the rate of detected indications and the accuracy in sizing. The indication of social loafing was a lower detection rate (DR; frequency of detected indications divided by the number of all possible indications) and a lower correct sizing rate (CSR; frequency of accurately sized indications divided by the number of all detected indications).

5.5.2.5. Procedure

The experiment commenced with the presentation of the background of the project (context), and the study aim. The participants were told that the aim of the study was to determine the efficiency of teamwork, as opposed to working alone. Efficiency was to be defined through a) high quality of the evaluation results and b) lower amount of time needed to complete the task. The efficiency was to be calculated by combining the quality of the evaluation results (correct number of found indications and their accurate sizing) and time of the individual (non-redundant) or team (redundant) contributions. The purpose of the cover story was to simulate common demands in the practice (the demand for high reliability, productivity, and low cost), and to refrain the participants from thinking about the actual aims of the study. The purpose of presenting the context of the study, i.e. the application of NDT in the inspection of the canister components to be used for the final disposal of spent nuclear fuel, was to induce motivation for the task.

By assuring the participants that their results will be anonymous in both experimental conditions, the identifiability and evaluation potential were held constant. All participants received the same introduction into the study and the information about the context of the study, i.e. the spent nuclear fuel management project. Therewith, the task and the individual contribution to the study were aimed to be meaningful.

The introduction was followed by a half-an-hour training session, in which the participants were taught how to work with the chosen data evaluation software and how to complete the experimental task. During this time, they were given an opportunity to become familiarised with the software and to practice on up to eight UT images, as well as to ask questions. After it was ensured that the participants understood the task, they were asked to complete the questionnaire on their experience and then read the experimental instruction. After they were assured anonymity and asked to identify themselves with their own code throughout the experiment, they were asked to start the evaluation task.

The duration of the experiment was about 1.5 to 2 hours, depending on the speed of the participant. It was conducted in groups of up to 10 participants. At the end of the session the

data were collected from the computers and carefully stored, and the participants were thanked for their participation.

5.5.3. Results

The statistical analyses were conducted with IBM SPSS Statistics package, version 22.

5.5.3.1. Data preparation

From 37 indications to be found and characterised, three were especially difficult, due to a very low signal-to-noise ratio (when the noise is very high, signals become hardly distinguishable from the noise, resulting in a low signal-to-noise ratio, or SNR). Whereas the detection was possible, the sizing was difficult following the provided NDT instruction and it would require more experience and knowledge from the participants. Hence, these three indications were excluded from the analysis of the correct indication sizing (n = 34 indications).

Following the principle of *winsorising*[14] (Field, 2013), the outliers were substituted with the last value that was not an outlier. The Kolmogorov-Smirnov test was used as an initial test for establishing whether the sample distribution approximates the normal distribution ($p > .05$). As a second step, the significance of skewness and kurtosis of the distributions was analysed by dividing them with their standard error (Field, 2013). This, in combination with the exceptions with respect to the robustness of parametric tests, i.e. T test, against violations of normality[15] (c.f. Bortz & Schuster, 2010) was used to assess whether parametric statistics can be used.

An overall alpha rate of $p = .05$ was used as a criterion for the null hypothesis significance testing, i.e. the minimum of $p < .05$ is implied when differences are referred to as statistically significant (other levels include $p < .01$ and $p < .001$, as typical in psychological research). All significant results are accompanied by the Cohen's d coefficient for effect size, with d values of .20, .50, and .80 reflecting small, medium, and large effects, respectively (Cohen, 2013).

5.5.3.2. Performance measures

A Paired-samples T test[16] was used to examine the hypothesis whether the participants led to believe they are working in a redundant team would loaf, as opposed to those participants working alone. The results show no significant differences between the Non-Redundant and Redundant Inspector condition in detection and correct sizing rate.

With no difference in the performance between working alone or in a group the participants missed, in average, one critical indication and incorrectly sized five (nRI) to six (RI) indications, with no false alarms.

---

[14] An alternative to *winsorising* would be to trim the data by deleting the outliers, or to transform of data (e.g. logarithmic or square root transformations). However, the transformation did not yield satisfactory results, and it was refrained from deleting values not to decrease the sample size.

[15] T test is robust against violations of normality under the following conditions: a) if the sample sizes are approximately the same, b) if the variances are approximately the same (in case of unequal sample sizes), and c) if the paired samples are positively correlated (Bortz & Schuster, 2010).

[16] The Kolmogorov-Smirnov test indicated the difference between the scores of the paired samples for Correct Sizing Rate to be normally distributed. Detection Rate satisfied the requirements of the robustness of the *T test*.

5.6. Experiment 2: Role of the redundant checker

The aim of second experiment was to investigate the behaviour of the redundant checker in a sequential redundant team and its dependence on receiving information about a redundant inspector's superior experience.

5.6.1. Hypothesis

What distinguishes human redundancy from technical redundancy is the fact that humans are usually aware of each other. This *awareness* is the main reason why people are never completely independent of each other. One of the sources of that dependence, among others, is the expectation of co-worker's performance, which has shown to be a strong moderator of the social loafing and social compensation effects (Karau & Williams, 1993). Expectation of co-worker's performance can result from a variety of factors, i.e. among others, the knowledge of the other's technical or experience level (Swain & Guttmann, 1983). A prevalent stereotype in NDT is the belief in the inspector's experience: If he is well experienced, he must perform well, and can be relied upon. Combining this stereotype with findings on the expectation of the co-worker's performance was the motivation to carry out this part of the study. Hence, the following was postulated:

> **Hypothesis:** Individuals in charge of checking the results of another inspector will loaf more if they are led to believe that the inspector is highly experienced, than if they have no information about the inspector.

5.6.2. Method

5.6.2.1. Participants

Sixty-one participants took part in the experiment. Six were excluded from the analysis due to insufficient amount of collected data. Therefore, the final number of participants was 55 (5 female, 50 male; with average age of 26 (18-49) years). The sample consisted of 40 NDT trainees, nine trainers, and six researchers.

5.6.2.2. Apparatus and tasks

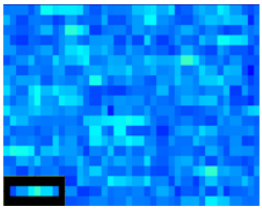
The experimental task in Experiment 2 was similar to the task in the previous experiment. The participants were given the same 20 UT images with 37 indications to be found and accurately sized using the ImageJ software, and were aided by the same tools (NDT instruction, list of shortcuts/key combinations). The major difference was that the participants were asked to *control* the results of another inspector. For that purpose, they were handed out reporting protocols allegedly filled out by their team partner. These contained the list of detected indications, with their location and determined size. In addition, the participants were provided images with the reported indications marked on them (see Figure 20). The task was to control whether all indications had been reported and their size accurately measured following the given NDT instruction.

Operator 3

Registration threshold: 100

	Nr.	Label	Area	Min	Max	XM	YM	BX	BY	Width	Height
1	1	UT10	30	39	103	5,9	19,7	1	18	10	3
2	2	UT10	3	81	95	68,4	14,5	67	14	3	1
3											
4											
5											
6											
7											
8											
9											
10											

Indication 1



Indication 2

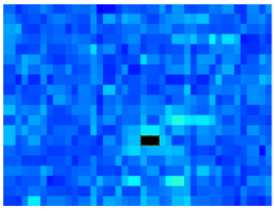


Figure 20: An example of a reporting protocol with a corresponding image with marked indications

In order to measure social loafing, 13 errors were implemented in the reporting sheets: i.e. three indications were missing, eight were inaccurately sized, and two were false alarms[17], providing opportunities for four omission and ten commission errors. Hence, the "team partner" detected about 92% of the indications, accurately sized about 78% of the indications, and committed two false alarms. These error types are representative of errors that can occur during data evaluation task.

To assess whether the manipulation—induced with a written experimental instruction—was successful, the participants were asked to assess their own and the performance of the team partner, and, in addition, their motivation for the task in a paper-pencil administered, so-called, *Performance evaluation and motivation questionnaire*. Specifically, they were asked about trust in the results of the team partner, about their ability to contribute to the overall result, the responsibility they felt for the final group outcome, whether they identified with the role, and finally, whether they thought the task was interesting and helpful for their future career (motivation questions). The altogether nine questions were rated on a 7-point Likert scale (e.g. from *bad* to *excellent*; *not at all* to *definitely yes*).

5.6.2.3. Design of the experiment

The main manipulation in this experiment was the *information about the team partner*, varied in two levels: whereas the Redundant Checkers (RC) were told they working with a team partner but given no extra information about him, the Informed Redundant Checkers (iRC) were

[17] Note that according to the original design, nine errors were implemented into the task, i.e. three omission (misses) and six commission error opportunities (four sizing errors and two false alarms). However, in the post-hoc check of the reference data (i.e. the list of indications with their exact locations and size, provided by the industrial client, who carried out the inspection) with the aid of a computer algorithm (programmed to detect and size according to the detection and sizing criteria employed in the simulated NDT task used in this experiment), it was established that some of the results deviated from those provided by the algorithm. This resulted in a larger error rate than planned.

instructed that the first inspector is a person with a yearlong experience in ultrasonic testing. The participants (between-subjects design) were randomly assigned to one of the two groups. In accordance with the previous experiment practice, the order in which the participants received the images and reporting protocols was randomised for each single participant. Therewith, since the experiment was conducted in a group setting, communication between the participants would not reveal that the information given to them is in fact the same and not belonging to a different inspector. In addition, this practice reduces unsystematic variation.

5.6.2.4. Dependent variables

The performance measures, i.e. the dependent variables, included the *agreement with the errors* committed by the previous inspector, i.e. the commitment of omission (failing to detect an indication missed by the redundant inspector) and commission errors (agreeing with a false detection, i.e. false alarm, or a sizing error committed by the redundant inspector). Social loafing was operationally defined through a higher frequency of omission and commission errors in comparison to the other group. In addition, it was expected from those participants, who loafed, to have a lower detection rate (DR) and a lower correct sizing rate (CSR), indicating that less individual effort was invested into the task and, hence, that the inspection performance was suboptimal.

5.6.2.5. Procedure

The experiment was once again conducted in a group setting of up to 10 participants, and lasted about 1.5 to 2 hours. It started with an introduction about the background of the project (to increase motivation for the participation) and followed with the training in duration of about 30 min, during which participants were taught how to work with the evaluation software and how to complete task, as well as given opportunity to practice. The cover story was again that the goal was to determine the efficiency of teamwork, as opposed to working alone (a simulation of common demands in the NDT practice – high productivity, low cost). Efficiency was to be defined through a) high quality of the evaluation results and b) lower amount of time needed to complete the task. The efficiency of teamwork was then to be calculated by combining the results and the time of both team partners. In reality, time was of no relevance for the analysis of the results. Upon completion of the task, the participants filled out the *Performance evaluation and motivation questionnaire* and they were thanked for their participation.

In line with the previous experiment, the identifiability (all participant's contributions were anonymous), the meaningfulness of the task, and of the relevance of the individual contribution to the study (by explaining the context of the study) were held constant.

5.6.3. Results

The statistical analysis was again carried out with IBM SPSS Statistics package, version 22.

5.6.3.1. Data preparation

As in the previous experiment, three indications were very difficult to size following the proscribed inspection procedure and were excluded *post-hoc* from the analysis. The same approach to outliers, normality of the distribution, significance levels, and effect sizes described in section 5.5.3.1 was applied to this experiment. The effect size for non-parametric tests was calculated using the following formula: $r = Z / \text{Sqrt } N$ (Rosenthal, 1991).

Frequency of omission and commission errors, and error rates, in general, are measured on a ratio scale. Since they represent error *counts* and, hence, take a form of finite numbers (e.g. 0, 1, 2, 3), they are defined as *discrete*. Even though not continuous in the true meaning of the term, discrete variables are frequently treated as continuous variables and analysed using parametric statistical methods (Bortz & Schuster, 2010; Field, 2013), which is the approach adopted in this analysis.

5.6.3.2. Performance evaluation

The first step in the analysis was to check whether the experimental manipulation was successful. That check was based on the *performance evaluation and motivation questionnaire*, consisting of nine questions related to the participant's participation in the study. Figure 21 shows the answers on the particular items with respect to the experimental condition (Redundant Checker, RC: Informed Redundant Checker, iRC).

Just by observing the figure, it seems that the participants evaluated the team partner (Q1) worse in the iRC condition, even though the team partner is supposed to be highly experienced. In addition, the iRC participants rated to have been able to contribute to the overall result (Q7) slightly more than the RC participants. Of special interest, next to the evaluation of the team partner's performance was the evaluation of own performance (Q2) and the evaluation of the team partner's experience level (Q3), which do not differ between the experimental groups.

No statistically significant differences in the ratings between the redundant conditions (RC vs. iRC) on any of the items were found, as established by the Independent Samples Mann-Whitney *U* Test. However, further results show that, regardless of the experimental condition, the participants (within-subjects) evaluated their own performance ($Mdn = 5$) as significantly better than that of the team partner ($Mdn = 4$), as obtained with the Related Samples Wilcoxon Signed Ranks Test ($Z = 3.69$, $p < .001$, $r = .54$).

Taking the first two questions into consideration, i.e. the evaluation of own and the team partner's performance, it was possible to create one variable dubbed *Performance Evaluation* on three levels, depending which performance was assessed to be superior: Superior Self, Superior Team Partner, or No Superiority. Figure 22 shows the difference in the frequency of participants who evaluated their own performance as superior, less superior, or the same as that of the team partner, depending on the experimental condition.

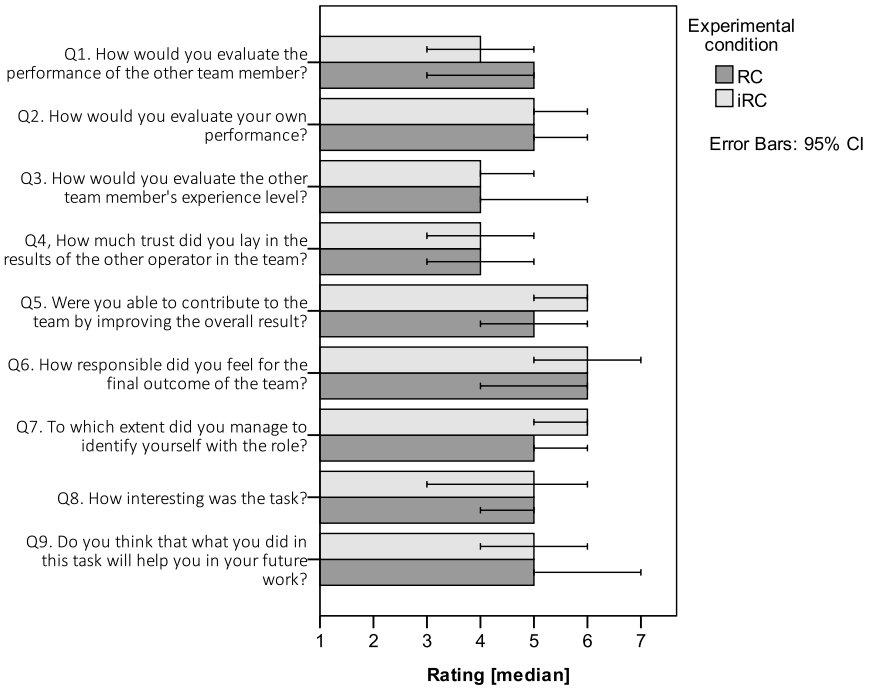


Figure 21: The difference between the Redundant Checker (RC) and the Informed Redundant Checker (iRC) in the performance evaluation and motivation

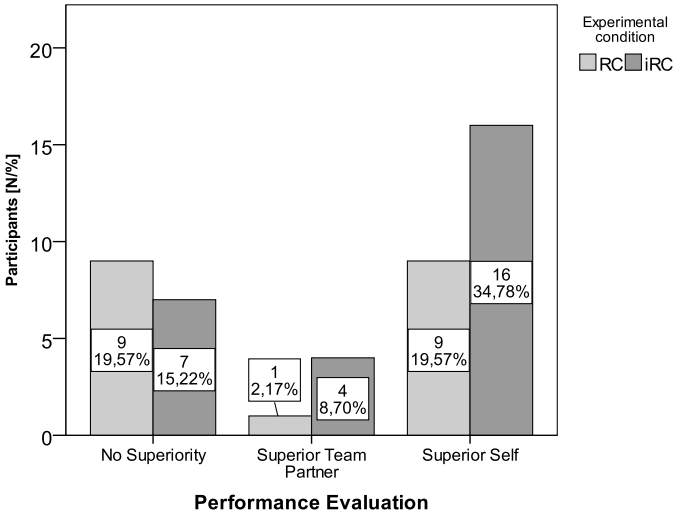


Figure 22: The difference between the Redundant Checker (RC) and the Informed Redundant Checker (iRC) conditions in the frequency of participants who evaluated their own performance (Superior Self) or that of the team partner as better (Superior Team Partner), or as equal (No Superiority)

It can be observed that only five participants declared the team partner to have been superior in the task, whereas the majority, especially in the iRC condition, evaluated their own performance as superior to that of the team partner. However, no statistical differences could be determined as the data did not fulfil the requirement for a Chi-square analysis (expected counts in several cells were lower than five [e.g. Field, 2013]).

It seems that—upon completing the task—the participants were not convinced in the team partner's superior experience, probably due to a large number of errors allegedly committed by the previous redundant inspector (about 1/3 of the task was erroneous). This suggests that the information itself (conveyed by means of experimental instruction) was overridden by the actual performance of the team member, which was the same in both conditions. Hence, it can be concluded that the information did not have the expected effect on the belief in the redundant inspector and that the hypothesis, i.e. that the information about the team partner's superior experience would lead to social loafing, was not confirmed.

The results presented in Figure 21 show that the task was judged equally interesting and relevant for the future of the participants, i.e. meaningful, by both experimental groups. Taking into account that the literature (e.g. Karau & Williams, 1993) suggests that—when the task is meaningful and the team partner's performance is poor—it is more likely for individuals to excerpt *more* effort in order to compensate for the team partner, rather than to loaf; the potential effects of social compensation were explored.

5.6.3.3. Additional exploratory analysis: Social compensation effects

To examine whether the participants compensated for the poor performing predecessor, a series of one-sample *T* tests[18] was conducted. Hence, the mean values of the sample were compared to the test values, i.e. performance of the alleged inspector. Note that the participants may have committed errors that extend beyond the implemented errors. The results of the One-Sample *T* tests are presented in Table 13, accompanied by a graphical representation of the mean differences in Figure 23.

Table 13: The results of the One-Sample *T* Test for the differences between the redundant inspector (RI) and the Redundant Checker (RC) in the sizing and detection performance

Dependent variable	*t*	*df*	*p*	*d*
Detection rate, DR	14.41	54	.000	3.92
Correct sizing rate, CSR	3.52	54	.001	.96
Frequency of misses	–5.06	54	.000	–1.38
Frequency of sizing errors	.00	54	1.000	.00
Frequency of false alarms	-24.08	54	.000	–6.55

Apart from the frequency of sizing errors, the participants significantly improved the results by displaying higher detection and correct sizing rate and by exhibiting a lower frequency of misses and false alarms.

[18] Not all variables were normally distributed, i.e. Detection Rate was associated with a minor skew and misses, sizing errors, and false alarms were associated with a significant skew (p (skew/SE skew) < .001). A non-parametric One-Sample Wilcoxon Sign Test, used to control the effects, yielded the same significant effects, with the following effect sizes: r = .79 (DR), –.57 (miss), –.92 (false alarm). For comparison, the One-Sample *T* Test effect sizes were: r = .89 (DR), –.57 (miss), –.96 (false alarm).

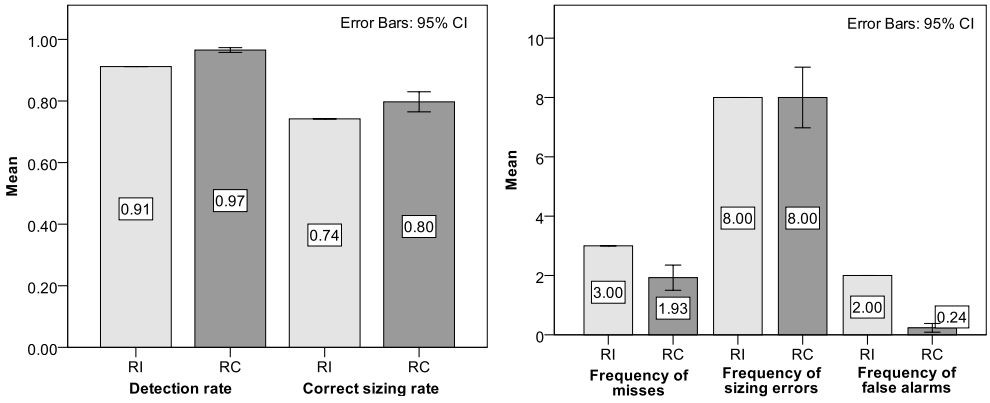


Figure 23: The difference between the Redundant Inspector (RI) and the Redundant Checker (RC) in the sizing and detection performance

5.7. Discussion

Examining the possible problems of sequential human redundancy and potential social loafing effects in the field of NDT was the first of its kind. Inspired by the literature stating that human redundancy might not be the best error recovery mechanism due to dependence between the team members and its potential resulting social loafing effects; this study aimed to demonstrate those effects in NDT and to highlight potential problems of human redundancy in the practice.

In spite of the expectations, this study was not entirely successful in demonstrating those effects. The results of the two carried out studies will be summarised and interpreted, followed by a critical reflection on the applied methodology, concluding with suggestions for the implementation of redundancy in NDT and suggestions for future research.

5.7.1. Summary and interpretation of the results

The two conducted studies were designed to explore social loafing effects in a sequential redundant evaluation of data collected with NDT. In such a redundant system, two distinct roles and role-associated tasks were examined: the role of the first inspector, whose task is to detect and characterise indications, and the role of the second inspector, the so-called checker, whose task is to control the results of the first inspector.

5.7.1.1. Experiment 1: Role of the redundant inspector

In the first experiment, it was postulated that the mere knowledge that one is working in a team and that one's results could be checked by another would lead to social loafing, defined through lower detection rate and a lower sizing precision, in comparison to working alone. This hypothesis was not confirmed, as detection and correct sizing rate did not differ between the experimental conditions. This result is not in line with previous similar studies, in which participants led to believe they are working *in parallel* with another (imagined) team partner did exhibit signs of social loafing (Manzey et al., 2013; Marold, 2011).

In the context of NDT and the experimental setting of this study, the interaction of other factors, not investigated in the scope of this study, may have affected the obtained results. For example, the participants, i.e. NDT trainees, trainers, and researchers—with some experience in NDT, but no experience in data evaluation—may have found the task as an opportunity to learn, and therefore intrinsically meaningful. Meaningfulness of the task has shown to decrease social loafing and foster social compensation behaviour (Williams & Karau, 1991). This may have led to an equal excerption of effort in both conditions (non-redundant and redundant).

Furthermore, the participants may have expected their colleagues to be their team partners. Members of cohesive groups tend to invest an equal amount of effort across working conditions, collective or co-active, due to the so-called "high-effort" heuristic. Karau & Williams (1997) elaborate this by stating that motivation and effort are primarily driven by individualistic needs when working with strangers, whereas in cohesive groups, people are far less concerned with individual needs, but rather work hard because the group members are valued. Furthermore, the participants may have felt that their contributions will be evaluated by other members of the group. Evaluation potential and high group cohesiveness may further explain why participants did not decrease their effort in the task.

The Collective Effort Model by Karau & Williams (1993) suggests that individuals will invest effort on a task depending on the degree by which they expect their efforts will be instrumental in obtaining valued outcomes. Taking into account that the participants were presented with a highly safety-relevant context of the study (spent nuclear fuel management) it may be that the outcome was highly valued by the participants and that they felt their contribution was instrumental to the successful completion of the task.

In conclusion, supported by the interpretation of the obtained results, the mere awareness one is a part of redundancy will affect reliability, as suggested by Clarke (2005), may not necessarily lead to negative performance outcomes. A task that is meaningful, a valued outcome, potential by the evaluation of the group or strong group cohesiveness may have had an effect on the motivation in the task.

5.7.1.2. Experiment 2: Role of the redundant checker

In the second experiment, the role of the redundant checker was explored. It was postulated that knowing that the previous inspector is very experienced would lead to higher social loafing (measured in terms of agreement with the previous inspector's errors) than when not knowing anything about his experience. This hypothesis was not confirmed. The evaluation of own and team partner's performance revealed that the participants were not affected by the information about the team partner's superior experience, as it seems to be overridden by his actual performance that was the same across conditions and relatively poor.

Since the participants experienced that their partner is not very reliable and were taking part in—to them—an interesting and relevant task, it was, thus, assumed that the participants might have compensated for the less reliable inspector, regardless of the experimental condition they were part of, in line with the social compensation theory. This *post-hoc* assumption was confirmed by comparing the results of the redundant inspector (in reality, data provided by the experimenters containing errors) to the average performance of all redundant checkers. Except in the frequency of sizing errors, the participants compensated for the less reliable team partner by exhibiting better performance in terms of detection, sizing, frequency of misses and false alarms.

The reasons for the occurrence of social compensation are not certain, as the potential moderating variables were not manipulated in the experiment. One potential reason for the compensation effect, as stated above, may have been that participants expressed relatively high interest in the task and its usefulness for their future work as inspectors. This, in combination with the context of the study, may have been a strong motivator to perform well and to compensate for the poorly performing team partner.

The participants were also a part of a cohesive group, participating in the experiment together with classmates or work colleagues. Whereas one group of participants was led to believe they are working with an NDT inspector with a yearlong experience, i.e. clearly not a member of the group, the other group may have assumed their colleagues would be the ones checking their results. In both cases, participants may have felt being evaluated as a group. In the social loafing studies, increased effort in individual tasks or in collective tasks is assigned to the potential that the individuals will be positively evaluated, i.e. people are motivated to work hard because of this potential evaluation. In addition, in case of low effort of the partner, the social compensation effect tends to be higher in cohesive groups (Williams & Karau, 1991).

Not only is the possibility of evaluation by an external source useful to eliminate social loafing, but also the potential for self-evaluation alone may be sufficient to eliminate it (Harkins & Szymanski, 1988; Szymanski & Harkins, 1987). The self-evaluation could have been made possible by comparing individual performance to the performance of the inspector, whose work one is checking. Considering the poor reliability of the inspector, invested effort would almost certainly result in a positive self-evaluation. This explanation is supported by the result that the majority of the participants did in fact evaluate their own performance as superior to that of the team partner.

5.7.2. Limitations of the studies

Potential problems encountered in the first experiment include primarily the fact that the same participants took part in both experimental conditions. Even though given different instructions, carrying out the task alone or as first in line of two redundant inspectors might not have made much difference in the motivation and the resulting effort; especially in a one-hour-long task. The participants may have approached the assignment in the same way. On the other hand, participants may have realised that the task they are carrying out is the same, in spite of the experimental instruction and the randomisation of the order of the images they were asked to evaluate.

In the second experiment, the participants were unaffected by the information about the previous inspector's yearlong experience and instead of loafing, exhibited social compensation behaviour. The major reason for the experimental manipulation not succeeding in its purpose was probably the reliability of the redundant inspector, whose results were being checked. Had the reliability of the inspector been high, coupled with knowing of his superior experience, the participants may have shirked off some of the responsibility by not inspecting that diligently and, hence, may have failed in recovery of his errors (the initial, theory-based assumption). However, this was not revealed in this experiment and instead, the participants exhibited compensation effects. The primary criticism here relates to the design of the experiment. The initial amount of implemented errors, i.e. nine errors—, which, in hindsight, was high, to begin with—was actually higher due to a shortcoming in the design. In a post-hoc analysis of the given data, it was established that the amount of errors was higher than initially thought of (13 errors). Problems of this kind may be encountered in future studies involving data evaluation in NDT. The only way to know the exact number of defects in the

material and their actual size is to cut slices in the material, look at the defects and measure them. Considering that in reality this is not possible—since this would mean that the component would be destroyed—it has to be relied upon the inspector doing the evaluation to his best ability. Taking into account that the original data provided by the inspector who conducted the evaluation with the original software had to be reviewed for the evaluation with different software and different criteria employed in this experiment, some differences were not identified in time. This shortcoming highlights the importance of a full-scale pilot study, even at the expense of the sample size.

In the second experiment, no control group to which the performance results could be compared was provided. There is no possibility to check someone's result and *not* be a part of some kind of a redundant team. Moreover, those working alone and those being the redundant checker differed in the type of the task they are doing (identifying vs. checking). Hence, the best measure for loafing or compensation effects was to take the given data with implemented errors as a baseline to which all other performances are compared to, or simply compare experimental groups, assuming that the difference between them would be sufficient to grant an effect.

Another noteworthy shortcoming of this study, in general, includes identifiability. In real work settings, all inspectors would be identifiable, by signing their name in the reporting sheet. However, they were, in this study, for practical reasons made unidentifiable. NDT practitioners taking part in performance evaluating studies tend to fear evaluation from their superiors and work differently than they would in the practice. In the study of Gaal et al. (2009), highly experienced NDT inspectors took hours to inspect an area they would otherwise inspect in a time significantly shorter than that. When asked about their behaviour, the inspectors stated fear from their performance being evaluated by their superiors and suffering potential consequences for that. After they were assured that their contributions are completely anonymous and will only be used for scientific purposes, the inspectors changed their behaviour and performed similarly as in the practice. Out of concern that identifiability would make the participants in this study act differently as they would in their daily job, it was opted to conduct the experiment anonymously.

Another factor that distinguishes this study from the NDT practice is the potential over-simplification of the NDT data evaluation task. Since it was not possible to train all the participants to work with a highly complex software and analyse data from—to the participants—unfamiliar components of unfamiliar geometries and properties, it was opted to design a simpler task with as many similarities to the actual task as possible. Still, this process might have over-simplified the task, which might not require too much effort to compensate.

In conclusion, based on the conducted studies, it was not possible to show decrements in the performance as a result of human redundancy, which contradicts existing literature on social loafing effects in human redundant systems. Whether in this specific task social loafing effects would occur when working with more reliable team partners demands further attention. Still, considering the body of literature showing dangers of insufficiently considered human redundancy, in the next section, their potential implications for the NDT practice will be discussed.

5.7.3. Implications of the studies for the NDT practice

Even though the conducted studies were not successful in demonstrating social loafing effects, other studies have provided evidence that if the differences between human and technical redundant systems, especially with respect to independency, are not taken into

account redundancy can fail. In this section, suggestions will be made with respect how to implement human redundancy in NDT optimally, supported by the findings from social loafing and compensation literature.

The first experiment showed that expecting someone else to conduct the same task afterwards led to the same amount of effort in the task as when working alone. Studies suggest that high levels of performance may be expected if redundant individuals are clearly identifiable (e.g. Williams et al., 1981) and expect to be evaluated by their succeeding counterpart (e.g. Harkins & Jackson, 1985), making these good strategies for tackling the potential loafing effects between dependent individuals.

The checkers are even more likely to be influenced by the dependence between them and the previous inspector, as they are frequently aware of their predecessor's findings. This influence can be strengthened by the ability to assess the predecessor's reliability and the familiarity between the inspectors, among others. Elaborate findings from social psychology suggest that if human redundancy is expected to succeed in error recovery, dependence between the redundant individuals must be decreased, if not eliminated. Since complete independence may be difficult to achieve in small inspection companies, other strategies could be used to decrease the dependence. For example, by ensuring that the inspectors neither posses any information about the other redundant inspector, nor have any knowledge of his inspection results, nor the decisions he had made. The inspector called upon after a critical defect had been found should also not be aware of that finding, as it may not only reduce dependence, but also affect his expectations about the criticality of the defect.

Furthermore, involving the inspectors in the decision making process after the inspection will counteract the tendency to loaf, since giving the inspectors more responsibility for the task and increasing the personal involvement have also shown to reduce or even eliminate social loafing (Brickner, Harkins, & Ostrom, 1986; George, 1992; Price, 1987). Providing with some kind of evaluation of the individual performance and feeding it back to the individuals has shown to raise personal accountability and eliminate loafing effects (Manzey et al., 2013).

The frequent NDT practice that involves supervisor overseeing the inspector conducting the task or controlling the reporting sheets, and calling it the four-eyes principle seems to be flawed, as effective human redundancy includes active involvement of both redundant elements (Swain & Guttmann, 1983).

Of note is the survey of Wheeler, Rankin, Spanner, Budalmente, & Taylor (1986), in which 66 percent of the inspectors stated that it was *not* likely that they would find indications not previously found by another inspector. This suggests that inspectors themselves are not convinced in the error recovery mechanism of human redundancy. However, this survey was conducted almost 30 years ago, which begs a question whether the improved NDT methods and training, reliable modern technology, and the increasing use of automated aids in the evaluation had led to a rise in certainty in their error recovery skills. However, dangers can lurk behind the feeling of certainty and safety and the theoretical background of this study provided with an abundance of evidence in favour of social loafing detrimental effects on individual motivation and effort on group tasks. Designers and planners of the NDT inspections should recognise them and make sure that the mentioned influenced factors are considered, if they decide to implement human redundancy.

Sagan (2004) concludes that organisational efforts to increase reliability and safety through redundancy can backfire in numerous and complex ways. However, the implication of this argument is not that redundancy never helps to improve reliability and safety, but rather that an organisation has to consider redundancy carefully.

5.7.4. Outlook

If the NDT field continues to use human redundancy, future studies may be needed to understand the possible pitfalls of its different forms of implementation. One of these forms includes redundancy applied to data acquisition, where redundant inspections may be carried out in a form of an overlap and with cognitive diversity.

Even though in-service inspections (routine inspections carried out at predetermined time intervals, e.g. every year) do not constitute as sequential human redundancy—as redundancy includes a time restraint (it should take place soon after the initial inspection; Clarke, 2005)—similar social factors may play a role in assuring effectiveness of the process. These inspections can include re-inspecting areas known to contain a yet-uncritical defect or areas expected to be free of defects. If the inspector receives information, or the *signed* report, indicating the existence or non-existence of the defect, this may in turn affect his behaviour in the task in a manner akin to human redundancy, i.e. through dependence between the two inspectors. This and other implications of in-service inspections merit further attention.

Considering the prevalence of redundant inspections conducted by the supervisors or inspectors in service of the authority, sometimes of higher qualification than the regular inspection personnel, it may be interesting to explore whether expectations of the redundant checker's qualification and experience may affect performance of the first inspector positively or negatively.

Going even further into the field, than this study has done, and investigating the behaviour of highly experienced personnel in their natural environment would reveal even more about the extent of individual effort in human redundant teams and at the same time, win more attention from the planners and the designers of the NDT inspections.

Problems associated with sequential redundancy refer not only to the inspection tasks, but also to decision-making tasks. That the decisions of previous decision makers affect people in a way that they ignore their own opinion and conform to others has been illustrated by Solomon Asch's studies on conformity (Asch, 1955). Deutsch & Gerard (1955) offered an explanation for that conformity by distinguishing social influence into normative and informational. Normative social influence can lead to conforming to the expectations of others because of the desire to obtain approval and avoid rejection; and informational because of a belief that the opinions and decisions of others can improve own decisions and judgements. Considering the prevalence of joint and sequential decision making in NDT, the effects of normative and informational social influences would be a logical further topic to explore.

6. Empirical Study 3: Use of automated aids in the evaluation of NDT data

Trust in automated decision aids, i.e. specifically, in the defect detection and sizing software used in the evaluation of data in NDT, was the second issue raised during the FMEA. For example, the eddy current testing method—for which such a software exists and is being further developed—was given a lower risk rating in comparison to other methods, based on the belief in its capability and reliability. Considering this prevalent belief in the superior reliability of automated systems in NDT (in general) and its contribution to a more reliable NDT, the following question was raised: what may happen if an automated system is held reliable—and as a consequence, highly trusted—and it fails?

The human-automation interaction literature suggests that automated aids are not always used as they should be and that an uncritical reliance on the aid's cues may actually degrade performance. In NDT, an uncritical reliance on an aid that failed to identify a critical defect could lead to that defect being overlooked – an event that can endanger safety. Therefore, the focus of this study is on the automation-aided data evaluation and the potential downfalls associated with the inappropriate use of automated aids.

In the theoretical part of this chapter, the use of automated aids in NDT (section 6.1) and different kinds of automated aid's inappropriate use and its influencing factors (section 6.2) will be addressed; concluding with the aims of the study (section 6.3). In the empirical part, the hypotheses (section 6.4) will be followed by the description of employed method, the design of the study (section 6.5), the results (section 6.6), and their discussion (section 6.7).

6.1. Automated aids in NDT

Non-destructive testing (NDT) has a reputation for being one of the slowest innovating sectors (Wassink, 2012). Still, over the last decades an increase in the use of automated and semi-automated systems for acquisition (e.g. Pitkänen et al., 2014; Rosado, Santos, Piedade, Ramos, & Vilaça, 2010) and—to a smaller extent—evaluation of data (e.g. Pitkänen, Lipponen, Lahdenpera, & Kiselmann, 2009) has been observed.

The reason for automating the defect detection and sizing processes is to overcome problems associated with manual evaluation: the time consumption, the associated costs, and the risk of human failure (Lingvall & Stepinski, 2000). Liao & Li (1998) consider manual evaluation "*subjective, inconsistent, labor intensive and sometimes biased*" (p. 183). Hence, engineers and NDT

experts consider it desirable and beneficial to develop computer-aided systems that can assist the inspector in the evaluation of data. Objectivity, consistency, and productivity are some of the expectations of such a system, provided that it is successfully developed (Liao & Li, 1998).

Automated defected detection and classification systems are designed by using algorithms that, first, search for suspicious regions and then proceed with a more precise identification and location of defects. Thereby, information on the shape, position and the intensity level of the defect pattern is used (Liao & Li, 1998). Experimental studies, carried out to determine accuracy of the proposed algorithms for methods such as radiography (RT), eddy current (ET) or ultrasonic testing (UT), showed satisfactory results. For example, for simple geometrical shapes (circular, sphere, or rectangular) of artificially manufactured defects, studies found a detection accuracy of 94% in UT, 98.5% in RT, and 99% in ET (Lingvall & Stepinski, 2000; Sambath, Nagaraj, & Selvakumar, 2010; Sun, Bai, Sun, & Zhou, 2005). The detection capability of more realistic complex defect geometries, i.e. artificially manufactured defects simulated to resemble real defects, was over 91% in RT (Liao & Li, 1998; Wang & Liao, 2002). However, real defects—formed during the manufacturing process or the component's service—have even more complex geometries, which can lead to even lower detection rates (Santos & Perdigão, 2001).

The management of spent nuclear fuel is one of the NDT application fields, in which only the use of mechanised NDT is foreseen (Posiva Oy, 2015; SKB, 2013). For their purposes, efforts have been invested into the development of an automated detection and interpretation software to aid in the evaluation of eddy current testing (ET) data. The software is designed for data recording, visualisation, and analysis (Pitkänen et al., 2009). The purpose of the aid is to provide the inspector with a list of detected indications, their locations, and sizes (the algorithms concentrate on real defects that can occur in the canister welds). The role of the inspector is to *control* the results before reporting the findings.

The role of automated decision aids—as the term itself suggests—is to *assist* human operators in decision-making. They have two distinct functions: to alert (or make the user aware of a change in the situation, which might require action) and to recommend, or offer an advice on, a course of action (Parasuraman & Manzey, 2010). In other words, an aid is designed to provide decision cues, whereas a human user retains authority in decision-making.

Considering that no such system is flawless, it is reasonable to expect that, on occasion, automated aids may fail. This is also one of the reasons why in many domains automated systems and aids did not entirely replace people. Instead, people are still involved as a measure of error recovery and error correction.

Apart from possible technical failure, or an inability of an automated system to carry out specific tasks, human factors scientists have raised another concern: human-automation or human-computer interaction. According to them, automated aids are commonly designed by engineers without in-depth consideration of factors that can affect human decision-making (Mosier & Skitka, 1996). As a result, studies and practical experience have identified a variety of factors that can lead to negative, rather than positive consequences of the aids' use. For example, they can be used inappropriately, incorrectly, or inefficiently, which is frequently the case due to the belief in the aids' expertise and reliability (Mosier & Skitka, 1996). In the following sections, different types of inappropriate automation use and the factors contributing to that effect will be addressed.

6.2. Inappropriate automation use

In an ideal world, automated systems would be designed with the consideration of human operators, who would have the ability and the time to accurately assess their capabilities and use them accordingly. In reality, however, it has been observed that operators can *disuse* automation, i.e. underutilise it even though it is functioning correctly, or *misuse* it, i.e. uncritically rely on its correct functioning (Parasuraman & Riley, 1997).

6.2.1. Automation disuse

Typically, new technologies are not always accepted. Before one gains experience with the system, it is almost natural to distrust it. Too many false alarms will additionally lower the trust. The consequence of this kind of *distrust* in automation may be to underestimate the "true" reliability due to automation errors, to reject the capabilities of automation and, consequently, to underutilise it. This decision (and intention) is appropriate if it corresponds to the actual condition of the automated system. In case of highly reliable, but underutilised, automated systems and automated aids, we speak of *disuse* of automation (Parasuraman & Riley, 1997).

Evidence of automation disuse has been found, e.g. over a course of studies exploring the use of target detection aids for military purposes (Beck, Dzindolet, & Pierce, 2007; Dzindolet, Beck, & Pierce, 2000; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Dzindolet, Pierce, Beck, & Dawe, 2002). These studies show a strong tendency of individuals—after experiencing automation failures–to rely on self and only then on the automated aid. Dzindolet et al. (2000), for example, reported that as many as 67-84% of the participants in their studies disused automation. A later study of Beck et al. (2007) corroborated that finding with disuse rates of 55-84%.

In light of obvious errors of the automated aid, especially when the risks are high, or when an automation error happens during easy tasks (Dzindolet et al., 2003; Madhavan, Wiegmann, & Lacson, 2004; Madhavan et al., 2006), people tend to disuse automation and choose to rely on their manual performance. In addition, when led to believe that their aid is near perfect, e.g. more reliable (*positive framing*), people are more likely to notice the obvious automated aid's errors, and thus be less inclined to rely on the automated aid (Dzindolet et al., 2002). Disuse may also be related to the perceived cost of making a mistake if the automated aid is wrong (Ezer, Fisk, & Rogers, 2008). This is often the case in domains, in which a failure may lead to catastrophic consequences.

The issues related to automation disuse have frequently been associated with high financial cost, but also with several accidents that occurred because of the operators refusing to comply with alerting systems (Parasuraman & Riley, 1997).

6.2.2. Automation misuse

The most automated systems nowadays are highly reliable and almost never fail. It is, hence, not difficult to form trust toward them. However, the risk of high trust is associated with those rare situations when automated systems *do* fail. High trust in automation may result in uncritical reliance on the correct functioning of the system, without recognising its limitations. The so-called automation *misuse* (Parasuraman & Riley, 1997) is a result of over-trust in automation and takes effect when people inappropriately rely on automated systems that are less reliable than manual operation.

Automation misuse can be observed in a failure to monitor what the automation is doing, i.e. *automation-induced complacency* (Parasuraman, Molloy, & Singh, 1993); and in a failure to notice problems because the automated aid failed to indicate them (errors of omission), or in inappropriate following of an automated aid's directives (errors of commission)—both due to *automation bias*, a term related to working with automated aids (Mosier & Skitka, 1996).

6.2.2.1. Complacency

Complacency has been operationally defined as "*poorer detection of system malfunctions under automation control compared with manual control*" (Parasuraman & Manzey, 2010, p. 387). A typical example is a pilot who relies on the proper functioning of the autopilot so much that he neglects to check or monitor whether it is functioning properly. Complacency is, hence, visible in the verification behaviour of the users, and is observed in information under-sampling (Moray, 2003), i.e. insufficient monitoring and controlling of the automation's functions. This happens under the conditions of high workload and high reliability of automation (Parasuraman et al., 1993), and when attending to multiple tasks in too little time, especially when the user does not understand what the automation is doing (Manzey, 2012). Consequences of complacency include loss of situation awareness and an increased risk of missing automation failures (Bahner et al., 2008).

Whereas the term complacency typically applies to monitoring of complex automated systems, i.e. supervisory task, the term automation bias is related to automated aids.

6.2.2.2. Automation bias

The term *automation bias* was coined by Mosier & Skitka (1996), who defined it as a failure to notice problems of the automated aid because of "*the tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing*" (p. 205). The underlying explanation for automation bias can be found in the tendency to use heuristics to decrease cognitive effort, i.e. because it is easier to delegate tasks to automation, people will do it when an automated aid is available (Mosier & Skitka, 1996). When accurate and used correctly, relying on automated aids is a very efficient cognitive strategy that will lead to an accurate and reliable performance of the system. Reliance on less than perfect aids, however, can lead to errors.

Errors can arise because delegating tasks to automation makes human decision makers less attentive and more likely to miss problems of the system (Mosier & Skitka, 1996).

Evidence of automation bias had been found in various domains, e.g. in aviation (Mosier, Skitka, Burdick, & Heers, 1996; Mosier, Skitka, Heers, & Burdick, 1998; Skitka et al., 1999), luggage screening (e.g. Madhavan et al., 2006), process control (e.g. Bahner, Hüper, & Manzey, 2008), military (e.g. Dzindolet, Pierce, Beck, Dawe, & Anderson, 2001), command and control (Cummings, 2004; Rovira, McGarry, & Parasuraman, 2007), and health care (e.g. Alberdi, Povyakalo, Strigini, & Ayton, 2004; Povyakalo, Alberdi, Strigini, & Ayton, 2013).

For example, Skitka, Mosier, & Burdick (1999) had their participants (non-pilots) carry out a flight simulation task with and without an automated decision aid designed to monitor system states and make decision recommendations. The obtained results revealed higher omission and commission error rates for those participants working with the aid, as opposed to those working without the aid. E.g., the participants in the automated condition had a significantly lower detection rate (59%) than the participants in the manual condition (97% accuracy). In addition, 65% of the participants committed a commission error, even in presence of disconfirming evidence. Further evidence of automation bias had also been provided by Manzey, Reichenbach, & Onnasch (2012, Exp. 2). In their study, up to a half of the

automation aid-assisted participants failed to detect an error committed by the aid, whereas only one person carrying out the task manually did so.

Potential downfalls of working with an automated aid was also observed in a field close to NDT: radiological evaluation of mammographic images (e.g. Alberdi et al., 2004; 2005; Povyakalo, Alberdi, Strigini, & Ayton, 2013). To that respect, Alberdi et al. (2004) investigated the impact of computer-aided detection (CAD) in aiding in detection of breast cancer from mammographic images. The CAD tool is designed to alert the human film reader to areas on the mammogram where abnormalities (indicators of cancer) may be present. The CAD can commit false negatives (i.e. misses) by failing to prompt about the cancer, or by placing the prompt away from the tissue that contains cancer, as well as false positives (i.e. false alarms). Experienced film evaluators were asked to evaluate a set of 60 mammograms, half of which contained a cancer. The participants worked either without CAD (unprompted condition), or were aided by the CAD (prompted condition) that correctly marked 10 (a), falsely placed 11 (b), and missed 9 cancers (c), and committed 12 false alarms (d). The comparison of the two studies (unprompted condition vs. prompted condition) showed a significant decrease in the detection of the incorrectly marked (b) and unmarked cancers (c) when working *with* CAD, as opposed to manual control. Hence, working with a faulty automated aid increased the frequency of omission errors—an effect that the authors assigned to automation bias (following the advice of an automated aid without checking) and complacent behaviour (paying less attention to the mammograms with no prompt).

Automation-induced complacency and automation bias have been widely investigated independently of each other—the former as decreased monitoring and verification behaviour in relation to automation supervisory tasks, and the latter as a decision bias in relation to automated decision aids. However, recent studies have shown that omission and some instances of commission errors can result from inadequate verification/monitoring of automation (e.g. Bahner, Hüper, & Manzey, 2008; Manzey et al., 2012). This supports the statement by Parasuraman and Manzey (2010) that both are "*different manifestations of overlapping automation-induced phenomenon, with attention at the center*" (p. 405) and, as such, may be investigated together.

The fact that working with an aid cad lead to a decrease in the monitoring of automation and an increase in errors inspired a myriad of research studies concerned with identifying factors contributing to inappropriate automation use. Those relevant to this study will be presented in the following section.

6.2.3. Factors affecting inappropriate automation use

Interaction with automated aids is highly influenced by trust in automation (e.g. Lee & Moray, 1992, 1994; Lee & See, 2004). In simple words, if an automated aid is deemed trustworthy, it is more likely it will be used even in situations in which the automated system does not merit it (Parasuraman & Riley, 1997).

Ideally, the level of trust should correspond to the true capabilities of the automated system, which Lee & See (2004) dubbed "appropriate trust". However, people are usually not good in this task, since they often lack sufficient information to perform the task adequately. Hence, they rely on their *subjective assessment* of the system's capability (e.g. reliability), which can result in overestimation or underestimation of its true capabilities (Lee & See, 2004; Rice & Keller, 2009). Over trust and distrust arise from a complex interaction between system characteristics, system users, and the situation in which a system is used (Manzey, 2012; Parasuraman & Manzey, 2010), which in turn can result in automation misuse and disuse, respectively. In the

following, some of the factors affecting trust and consequent automation reliance—i.e. the reliability of the aid, experience with the automated aid, and individual differences in reliance on automation—will be highlighted.

6.2.3.1. Reliability of the aid

There is a consensus between the researchers that the reliability of the automated system/aid is one of the most salient factors affecting trust in automation and its consequent misuse and disuse (Lee & Moray, 1992, 1994; Lee & See, 2004; Madhavan & Wiegmann, 2004; Parasuraman & Riley, 1997). For example, in their study of performance consequences of complacency, Parasuraman et al. (1993) observed that participants detected more automation failures when working with a low reliable system (even though not significantly). Oakley, Mouloua, and Hancock (2003) reported a decrease in the detection rate with increasing automation reliability (from 35% to 95% reliability), indicating insufficient monitoring behaviour. Another study confirmed that high *static* automation reliability increased automation-induced complacency (A. Singh, Tiwari, & Singh, 2009).

Similar effects have been observed in the use of automated aids. In the study by Metzger & Parasuraman (2005), working with a highly reliable automated aid reduced the likelihood of automation failures' detection, whereas low reliable aid was disused. Dixon & Wickens (2006) observed poorer performance with low reliable aids in detecting false alarms in a system failure task and in detecting misses in concurrent tasks. In a subsequent study, they suggested that systems with reliability below 70% will not be trusted by the operators (Wickens & Dixon, 2007). Dzindolet et al. (2001), in contrast, observed misuse of the aid regardless of the reliability (60%, 75%, or 90%) suggesting that the participants are insensitive to the varying reliability of the aid.

6.2.3.2. Experience with the aid

Studies have shown that after experience with an automated aid that can err, people seek self-reliance rather than to rely on the aid (Dzindolet et al., 2002). Hence, trust in automation and a decision whether or not to rely on automation are not only dependent on the reliability of the aid, but also on the experience with the aid (Alberdi, Ayton, Povyakalo, & Strigini, 2005; Dzindolet et al., 2003, 2002; Manzey et al., 2012). Whereas positive experience will lead to increased trust in the aid, decreased verification behavior and, potentially, the errors in the task; negative experience will reduce trust and decrease complacent behavior (Lee & Moray, 1992; Manzey et al., 2012).

Experience with the aid and consequent trust can be affected by whether the aid commits misses or false alarms (Cotte, Meyer, & Coughlin, 2001; Dixon, Wickens, & McCarley, 2007; Meyer, 2001, 2004), whether the aid's failures are occasional or continuous (Parasuraman & Riley, 1997), and whether the aid commits errors on easy or difficult tasks (Dzindolet et al., 2003; Madhavan et al., 2004, 2006).

Whereas positive experience may result in reliance, experience of aid's failure may have a different effect on the operators, i.e. they may chose to prefer manual performance and rely on themselves, rather than on the aid. If the trust in aid is higher than trust in self, people are likely to misuse automation. In contrast, if the trust in self is higher, the aid is likely to be disused (Lee & Moray, 1992, 1994).

The choice for reliance on self or the automation is not only situation-dependent, but is also susceptible to individual differences.

6.2.3.3. Individual differences in the aid's use

Individual differences in interaction with automated aids did not receive as much attention as other previously mentioned factors, at least empirically. Still, many indicate individual differences as an unexplored factor that has shown to affect both complacency (I. L. Singh, Molloy, & Parasuraman, 1993) and automation bias (Skitka et al., 1999). This suggests that investing effort into investigating individual differences may provide additional insights into automation-induced phenomena.

Some researchers suggested that individual's interaction with automated aids (and automated systems, in general) can be shaped by attitudes towards automation (e.g. Dzindolet et al., 2003; Manzey, 2012; Parasuraman & Riley, 1997), by trust in one's own abilities and self confidence (e.g. Lee & See, 2004; Riley, 1996), personality characteristics (I. L. Singh et al., 1993), responses to the environment, i.e. workload, fatigue, sleep loss, mood, etc. (e.g. Reichenbach et al., 2011; Röttger, Bali, & Manzey, 2009), and by risk behaviour, i.e. risk-taking vs. risk-averse (e.g. Madhavan & Wiegmann, 2004, 2005).

Some insight into the effects of individual risk-taking tendencies on interaction with aids was provided by Madhavan & Wiegmann (2005) in their study of cognitive anchoring. The results of the study showed that those with high risk taking tendencies are more likely to pre-diagnose aid's failures and consequently disagree with the aid. The authors conclude that the choice of a utilization strategy in opaque systems may be influenced by biases to rely on own intuitive guesses about the system state or take risks. Apart from this study, risk-taking tendencies remain an unexplored topic in the field of inappropriate automation use.

In addition to research missing, as it may be the case in the study of individual differences, the studies of automation-related phenomena bear shortcomings with respect to external validity.

6.2.4. Shortcomings of the study of the inappropriate use of automated aids

One of the major shortcomings of study of inappropriate automation use refers to its external validity. The majority of the studies were conducted in laboratory-like settings with students as participants. Several notable exceptions to that rule include studies involving professional pilots in the aviation domain (Mosier et al., 1996, 1998), and radiologists in the study of computer-aided detection of cancers from mammograms in health care (e.g. Alberdi et al., 2004).

Apart from the necessary external validity of the investigated constructs, studies with professionals may reveal influences that are hard to identify with students, as professionals may have qualities that are not found in student samples. For example, Mosier et al. (1998) noted differences in examining the effects of accountability on performance with the aid between student samples and professional pilots: Whereas accountability had an effect on students, the same effect was not found in professional pilots. The authors assigned this effect to accountability being a personality trait in individuals choosing this profession. In addition, with increasing experience, the likelihood of detecting omission errors decreased in professional pilots—an effect, which was unexpected and could not be identified on student samples.

Even studies with actual users bear difficulties with respect to artificialities and simplifications of the experimental task. Even though all of the automation bias considerations by Alberdi and his colleagues were conducted with less to highly qualified radiologists (e.g. Alberdi et al., 2004; Alberdi, Povyakalo, et al., 2005), they addressed the difficulties in generalising the results

to the real-world practice due to important differences between their trial and the everyday practice. I.e. unrealistically high proportions of pathological cases are present in the experimental settings, but not in reality; whereas access to other information in addition to the decision aid and the ability to discuss the results in a group are present in the everyday practice and absent in an experimental setting (Alberdi, Povyakalo, et al., 2005).

Other scientists also suggested that in order to explore the generalizability of the present findings to real-world scenarios, future research should be conducted in realistic settings (Dzindolet, Pierce, et al., 2001; Madhavan & Wiegmann, 2005). In addition, Bahner (2008) suggested it may be interesting to explore automation bias effects in other domains (and different experimental environments) apart from aviation, which received the most attention.

6.3. Aims of the study

The aim of this study was to explore potential inappropriate use of automated aids in the NDT data evaluation. Thereby, the prevalent *belief* in the high reliability of automated systems in NDT and individual differences in risk taking, and their effects on the performance with the aid, were put in focus.

The motivation for this study arose after the FMEA showed that due to the belief in the high reliability of defect detection and characterisation aid, the NDT method using this aid, i.e. eddy current testing, is considered to be at lower risk for human failure (see Chapter 3, Figure 14). Whether this belief alone would be sufficient to cause overreliance on the aid, even in light of its failures, was investigated.

Interaction with the aid was observed in terms of omission and commission errors and the verification behaviour, in line with Parasuraman & Manzey's (2010) integrated model of complacency and automation bias.

6.4. Hypotheses

The assumptions in this study are based on the findings that people are likely to rely on reliable automated aids that may result in errors of omission and commission, as revealed by the body of automation bias research elaborated in the previous sections. Of special note is the study by Alberdi et al. (2004), which noted effects of automation misuse assigned to automation bias and complacency in a field very close to NDT, i.e. the field of medical radiology using computer-assisted detection of cancer lesions from mammographs. Hence, no control group was provided.

Building up on this assumption, the effects of belief in the aid's reliability on the performance in an NDT data evaluation task were studied. Even though researchers (e.g. Riley, 1996) suggest that trust in automation is affected only by its *actual* reliability, there have been no studies examining whether a belief in the reliability of the aid may affect performance, even when it is not in accordance with the system's actual reliability. Madhavan & Wiegmann (2005) provided some hints that this may be possible. Their study revealed that even under conditions of extreme system opacity, operators tend to diagnose the state of the aid, *prior* to its actual use. Their decision whether or not to agree with the aid has shown to be anchored to that "hunch", irrespective of the actual reliability of the aid. They called this effect *cognitive anchoring*. Studies have also shown that people generally trust in automated systems and consider them more reliable than manual operation (*"perfect automation schema";* Dzindolet,

Pierce, et al., 2001) or other people (Dzindolet et al., 2002; Madhavan & Wiegmann, 2004, 2007). Those positive attitudes may result a higher tendency to rely on the aid and comply with its suggestions, i.e. to automation bias and complacency effects. Hence, the following was hypothesised:

> **Hypothesis 1:** The belief in the aid's reliability will influence the perception of the performance of the aid and, thus, lead to differences in behaviour even towards equally reliable aids. Those being told that the aid is highly reliable—and evaluating the aid's performance after the task as superior to own—will commit more omission and commission errors and verify the results less, than those led to believe that the aid is not highly reliable.

The second hypothesis relates to the relationship between complacency effects and agreement with the aid (typically a behavioural consequence of automation bias). E.g. Parasuraman & Manzey (2010) suggested that the decreased verification behaviour may be responsible for the occurrence of omission and commission errors. In line with this statement, the following hypothesis was made:

> **Hypothesis 2:** Individuals, who committed more omission and commission errors, will have verified the results less frequently, than those with a lower frequency of errors will.

Guided by suggestions that risk prone behaviour may moderate automation use in a way that high risk taking is associated with higher disagreement with the aid (Madhavan & Wiegmann, 2005), the third hypothesis was made:

> **Hypothesis 3:** Risk-seekers are less likely to agree with the automated aid, i.e. they will commit less omission and commission errors, than risk-averse participants will. In contrast, cautious participants are more likely to misuse the automated aid, than those less cautious.

Finally, NDT reliability is affected by the overall performance in the detection and sizing task. Whereas the detection directly affects reliability, inaccurate sizing can affect the criticality assigned to the indication and, correspondingly, to the defect in the material. Considering that omission and commission errors directly affect the overall performance in the task—in terms of its detection and sizing capability—it was assumed that the above-mentioned factors would also affect the overall performance. Hence, the fourth hypothesis is as follows:

> **Hypothesis 4:** Those led to believe that the aid is highly reliable and those less likely to take risks will exhibit lower overall performance in the task.

6.5. Method

6.5.1. Participants

Seventy NDT trainees (69 male, 1 female) with average age of 27 years (range 16-55) took part in the experiment. The participants were sampled at two different organisations in departments in charge of education of the material inspectors (this includes NDT and other methods of material testing). The participants were sufficiently experienced in NDT to understand the task and complete it after a provided training, with no previous experience in data evaluation.

6.5.2. Apparatus and tasks

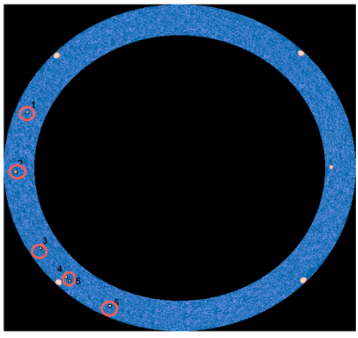
Similar to the previous study, the eddy current (ET) data evaluation task was simulated using the ImageJ version 1.43u software. In the NDT practice, the automated aid—ET data evaluation software—typically provides the inspector with a list of detected and measured indications, with the task to control its correctness in detection capability and its accuracy in sizing and positioning of the indications. The inspector assumes responsibility for the task by signing off the report sheet.

In line with this task, the participants were asked to control the results produced by the automated aid, i.e. establish whether all the indications have been correctly found and their size accurately measured. They conducted this task at a computer. For that purpose, the participants received seven C-scan images of the component weld. These contained from none to a maximum of 15 indications per image, summing up to 36 indications representing real defects in the weld. In addition, the participants were presented with filled-out reporting protocols containing all detected indications, their measurements, and snapshots of the images with marked indications (Figure 24)—all allegedly compiled by the automated aid.

In reality, the protocols were prepared by the experimenter and they contained three types of errors: three misses, two sizing errors, and two false alarms—corresponding to typical errors that can occur during data evaluation in NDT.

During the task, the participants were guided with a short inspection procedure, i.e. NDT instruction (designed especially for this task), containing all the necessary steps to be followed in order to complete the task successfully, and a list of shortcuts and key combinations to ease the work with unfamiliar software.

The behaviour during the task was recorded using TechSmith software Camtasia Studio 7 © for Windows. This software allows capturing videos of actions on the computer screen. Thereby it was possible to record all mouse movements across the screen in order to quantify the participants' verification behaviour.



	INTENSITY		POSITION				SIZE	
Indication No.	Min	Max	XM	YM	BX	BY	Width	Height
1	141	254	42,9	212	40	209	6	6
2	125	250	21,4	329,5	18	326	7	7
3	132	253	66,5	480,6	64	477	5	7
4	132	251	113,9	537	111	534	6	6
5	128	251	120,5	544,5	118	542	5	5
6	128	252	193,9	593,9	191	591	6	6

Area The surface area of the contour
Min The minimum pixel intensity value within the contour
Max The maximum pixel intensity value within the contour
XM The position of the maximum pixel intensity value in the x-direction
YM The position of the maximum pixel intensity value in the y-direction
BX The position of the upper left corner of the contour in the x-direction
BY The position of the upper left corner of the contour in the y-direction
Width The width of the contour (in pixel)
Height The height of the contour (in pixel)

Figure 24: An example of a reporting protocol with a corresponding image with marked indications[19]

To measure the perception of the aid's performance and risk propensity, two questionnaires were provided: the so-called *Performance evaluation questionnaire*, and the Risk Orientation Questionnaire, ROQ (Rohrmann, 2005). In the first, the participants were asked to rate their

[19] The white collections of pixels not marked by the aid include geometrical indications from the screws used to transport the lid and the starting point for the evaluation. The participants are trained to exclude those from the evaluation.

own performance in the task and the performance of the aid on a 7-point Likert scale, varying from *bad* to *excellent*. The second, i.e. ROQ, is designed to identify general orientations towards risk taking. It measures two risk-taking qualities: *Risk Propensity* and *Cautiousness*, each with 6 items (e.g. "I'm quite cautious when I make plans and when I act on them" or "Even when I know my chances are limited I try my luck"). The participants are instructed to read each sentence and then rate to what extent that statement is true for him/her using a 7-point Likert scale (from *no, not at all true for me* to *yes, very much so*). The scores for the two scales are obtained by multiplying the mean value with 10, resulting in a range of scores between 10 and 70.

6.5.3. Design of the experiment

The main manipulation in this experiment was the reliability of the automated aid, varied in two levels, i.e. high reliability (HR) and low reliability (LR). The manipulation was achieved through a written experimental instruction handed out to the participants at the beginning of the task. The instruction was used to describe shortly the task and to provide with information about the reliability of the aid. Thus, they comprised of the same information, with the exception of the information about the reliability of the aid. In the high reliability condition (HR), the participants were presented with the following information [originally in German]:

> "In our previous experience, it was shown that the results of this automated evaluation are almost always correct. Nevertheless, these results need to be controlled by an inspector."[20];

and in low reliability condition (LR), with the following information:

> "In our previous experience, it was shown that the results of this automated evaluation are not always correct. Your task is, hence, to control the results."

Different participants were randomly assigned to the experimental groups (N = 35 per experimental group; between-subjects design). The order of the images and the reporting protocols was randomised for each participant to prevent participants' communication during the experiment, to exclude learning effects, and to avoid all other sources of systematic variation.

Based on the ROQ questionnaire, that measures risk behaviour on two scales, i.e. risk propensity (ROQ-P) and cautiousness (ROQ-C), the participants were divided into those high and low in the propensity to take risk and high and low in cautiousness.

Post-hoc, the participants were divided into those low and high on agreement with the errors of the aid, based on the frequency of committed omission and commission errors. Those that committed none or one omission error were assigned to the *low agreement* group, and those with two or three errors into the *high agreement* group. The same procedure was applied to commission errors, i.e. those that committed none or one error were assigned to the *low agreement*, and those with two or more into the *high agreement* group.

6.5.4. Dependent variables

The perception of the reliability of the aid was measured through the performance evaluation, i.e. responses on the ratings of own and aid's performance (range 1 to 7, with seven reflecting high performance evaluation). If the aid's performance is rated higher than own performance, then the participants perceived the aid as more reliable.

[20] "Almost always correct" was used instead of "always correct", as technicians and engineers in NDT are said not to believe that an automated system can be perfectly reliable.

The performance was measured through the agreement with the errors of the aid, i.e. the commitment of omission and commission errors. In this experiment, there was an opportunity to commit three omission and four commission errors by agreeing with the aid that made three misses (omissions), two sizing errors, and two false alarms (commissions).

The verification behaviour was related to the steps in the procedure that needed to be followed in order to complete the task: opening the image, zooming, and sizing of the indications. Those tasks and the possible events, including the ranges of the values are presented in Table 14. The higher value on all procedure tasks indicates higher verification behaviour. Reversely, lower value indicates lower verification behaviour and hence, indicates signs of complacency.

Table 14: Procedure tasks and possible events in the verification behaviour

Name	Procedure task	Possible events
Opening	Opening the image	Values ranged from 0 (no image was opened) to 7 (all images were open)
Zooming	Zooming in the area around the indication	Values ranged from 0 (no area was zoomed in) to 38 (all 36 indications and 2 false alarms were zoomed in)
Sizing	Sizing of the indication (the participant controlled the pixel intensity values, and/or set the contour around the suspected size of the indication, and/or measured the obtained values within the contour)	Values ranged from 0 (no indication was sized) to 38 (all 36 indications and 2 false alarms were sized)

And finally, the complete performance in the task was observed, in terms of *detection rate*, defined as the frequency of detected indications divided by the number of all possible indications (DR), and the *correct sizing rate*, i.e. the frequency of accurately sized indications divided by the number of all detected indications (CSR). The values range from zero (no indication had been detected/correctly sized) to one (all indications had been detected/correctly sized).

6.5.5. Procedure

The experiment was carried out over the course of six days at the two training facilities where the participants were sampled. Up to ten participants worked simultaneously on ten provided computers and were all assigned to the same experimental condition. When possible, the order of the experimental conditions over the course of the day (morning vs. afternoon session) was varied to avoid the effects of time of the day.

Upon the entrance to the room, the participants were seated to their working stations and presented with a short summary of the project—the use of NDT in the final disposal of the spent nuclear fuel—aimed to raise motivation for the task and raise relevance of individual contribution. They were told that the aim of the study is to investigate the efficiency of automated software, defined through a) the quality of the results and b) the time needed to complete the task. In the following 30 minutes, the participants were trained how to use the evaluation software and how to complete the experimental task. Thereby, they were given an opportunity to practice evaluation and reporting on up to five example indications.

After assuring that the participants were comfortable with the task, they were handed out the experimental instruction (containing the description of the task, the experimental manipulation, and the assurance of anonymity) and asked to start with the task. After each

participant had completed the task, they were asked to rate their own and the performance of the aid, to complete the ROQ, and were thanked for their participation.

6.6. Results

All the statistical analyses presented in this chapter were conducted with IBM SPSS Statistics package, version 22.

6.6.1. Data preparation

The data were examined to establish whether there are any outliers, whether the used instruments are sufficiently reliable, and whether the distributions of the dependent variables satisfy the conditions for parametric statistics.

Two participants were excluded because their correct sizing rate was about 2%, i.e. the indication size of almost all indications was changed and hence, incorrect. It was thus assumed that they did not understand the task, or did not carry it out according to the trained instructions and the evaluation procedure. The final sample size counts 68 participants, with 34 per experimental condition.

The reliability analysis of the ROQ subscales showed insufficient internal consistency of the two subscales (Cronbach's alpha for both Risk Propensity and for Cautiousness, $\alpha = .38$). With removal of one item on the Risk Propensity and two items on the Cautiousness subscale, satisfactory reliability of $\alpha = .63$ (Risk Propensity) and $\alpha = .61$ (Cautiousness) with acceptable correlations (above .30 according to Field, 2013) of single items with the overall score was obtained. Similar reliability was also obtained by Hatfield & Fernandes (2009).

Even though the values on both ROQ scales can range from a minimum of 10 to a maximum of 70, the values on the Cautiousness scale reached a maximum of 28 indicating a rather low cautiousness. Dichotomising this variable into those less or more cautious would have no sound basis. Hence, it was excluded from further analysis. In contrast, the values on the Risk Propensity scale range from 20 to 68 and were divided into *low* and *high* based on the median-split method, with the cutting point at 48 (*Mdn*) assigned to the high risk propensity.

The verification behaviour was recorded for 45 participants. One participant's recording was incomplete, which is why only 44 recordings were used in the analyses (22 per experimental condition). The variable "Opening" was excluded from the analyses because all participants opened all images.

Adopting the same approach as in the previous study (see sections 5.5.3.1 and 5.6.3.1), discrete variables (i.e. omission, commission errors and error rates) were treated as continuous, as common in psychological research (Bortz & Schuster, 2010; Field, 2013). Outliers were treated using a procedure called *winsorising* (Field, 2013). If the distributions of the dependent variables were not normal, but complied with the requirements (i.e. sample size, variance) that still enable parametric tests to be robust against violations of normality (Bortz & Schuster, 2010), parametric statistical procedures were used. A minimum of $p < .05$ is implied when differences are referred to as statistically significant. All significant results are accompanied by the Cohen's d coefficient for effect size, with d values of .20, .50, and .80 reflecting small, medium, and large effects, respectively (Cohen, 2013).

6.6.2. Performance evaluation

To examine the first hypothesis and determine whether the information about the reliability of the aid would affect perception—and consequent evaluation—of the performance of the aid after working with it, the answers to the ratings of own and aid's performance were analysed. Figure 25 shows the average ratings of own and aid's performance with respect to the experimental condition (*N*=62; higher mean value indicates higher performance rating).

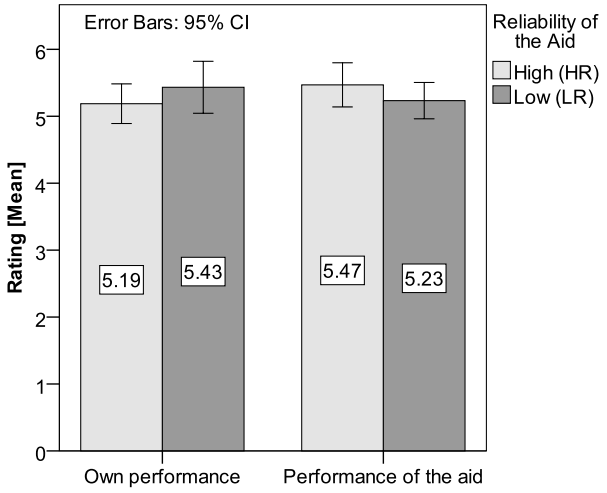


Figure 25: The evaluation of own and the aid's performance with respect to the reliability of the aid

No statistically significant difference in the ratings was found between the conditions (Independent samples *T* test), or within participants (Paired samples *T* test) in the evaluation of own and aid's performance. In other words, the evaluations of own performance and the performance of the aid were independent of the given information about the reliability of the aid.

By combining the two performance evaluation questions into one variable, i.e. a variable Performance Evaluation, the participants were divided into three groups depending on how they evaluated their own performance and that of the aid: No Superiority, Superior Self, and Superior Aid. Figure 26 shows the distribution of the participants in the three groups, depending on the experimental condition (Reliability of the Aid).

It may be observed that somewhat more participants in the HR condition (19.4%) evaluated the aid as superior, than in the LR condition (11.3%). The evaluation of own performance (Superior Self) is almost equal across the conditions, whereas somewhat more participants rated both performances as equal in the HR condition. However, the Chi-Square test (*N* = 62) revealed no association between different levels of performance evaluation and the reliability of the aid. Moreover, those that evaluated the aid as superior did not commit more omission and commission errors than those who evaluated their own performance as superior, or those that rated both performances equally, as hypothesised (One-way ANOVA).

Based on these results, it is reasonable to conclude that the information about the reliability of the aid will have no further effects on performance in the task. Hence, the effects of Reliability of the Aid on the performance measures were not further explored.

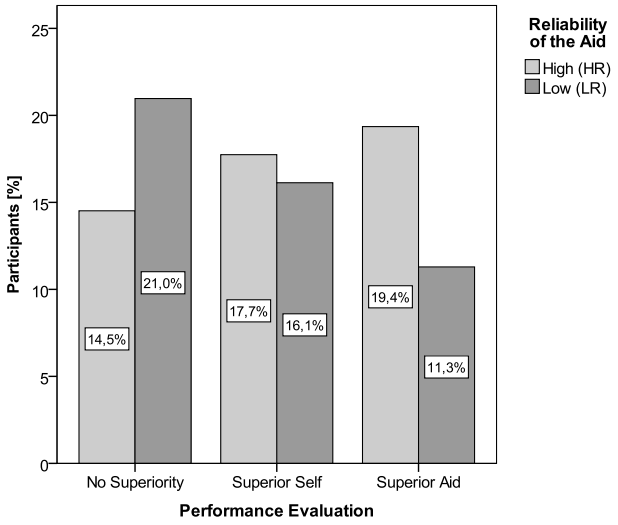


Figure 26: The difference between the HR and LR in the number of participants who evaluated their own performance (Superior Self) or that of the aid as better (Superior Aid) or as equal (No Superiority)

6.6.3. Descriptive data: Agreement with the aid

Descriptive look on the data reveals that almost half of the implemented errors remained undetected and uncorrected by the participants. From the seven errors committed by the aid, the participants in average agreed with almost two misses (out of three; $M = 1.75$, $SD = 1.07$), one sizing error (out of two; $M = .97$, $SD = .83$), and one false alarm (out of two; $M = 1.07$, $SD = .85$), committing therewith both errors of omission and commission. All participants agreed with at least one error of the aid.

6.6.4. Verification behaviour

The verification behaviour was measured by recording the participants' zooming and sizing behaviour. It was recorded that 56.7 % of the participants zoomed and 38.6% sized all 36 indications. Hence, a large portion of the participants were found to be complacent, at least to some extent, as they did not verify all the necessary information in line with the instruction and the inspection procedure.

A Pearson correlation coefficient shows a strong positive correlation between zooming and sizing ($r = .73$, $p < .001$), i.e. frequent zooming is strongly associated with frequent sizing behaviour.

Examining the relationship between agreement with the aid and verification behaviour (second hypothesis), the Independent Samples *T* test indicated a statistically significant difference between those high and low in agreement with misses (omission errors) in both zooming ($t(42) = 5.04$, $p < .001$, $d = 1.56$) and sizing ($t(42) = 2.14$, $p < .05$, $d = .66$). Those that committed more than one omission error (high agreement, $N = 17$) verified the indications *less* frequently than those that committed none or only one omission error (low agreement, $N = 27$).

Zooming behaviour was also different between those high and low in agreement with sizing errors and false alarms (commission errors), $t(42) = 2.06$, $p < .05$, $d = .63$. Those participants

with more than one commission error (high agreement, $N = 24$) zoomed in on the indications *less* frequently than those with only one commission error (low agreement, $N = 20$). The average differences in both zooming and sizing between low and high agreement are depicted in Figure 27.

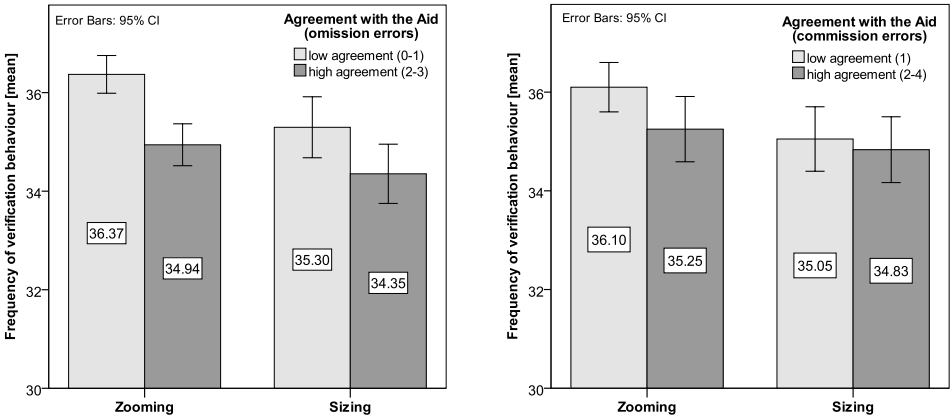


Figure 27: The difference between low and high agreement with the aid in the verification behaviour

6.6.5. Individual differences in risk propensity

To examine the third hypothesis and determine whether there was a difference between high and low risk propensity in the frequency of omission and commission errors, an Independent samples *T* test was conducted. This procedure revealed no significant difference between levels of risk propensity on omission errors. There was, however, a difference in commission errors ($t(63) = 2.45$, $p < .05$, $d = .62$). Low risk propensity was associated with a *higher* rate of commission errors ($M = 2.41$, $SD = 1.31$, $N = 27$), than high risk propensity ($M = 1.66$, $SD = 1.15$, $N = 38$).

6.6.6. Overall performance in the defect detection and sizing task

Apart from agreeing with the aid, the participants committed errors that extend beyond the implemented errors. In average, the identified 95% ($M = .95$, $SD = .03$) and correctly sized 87% of the given indications ($M = .87$, $SD = .09$). Thereby they averagely missed two indications ($M = 2.06$, $SD = 1.50$), incorrectly sized almost five indications ($M = 4.82$, $SD = 3.93$), and committed one to two false alarms ($M = 1.47$, $SD = 2.07$). In total, only eight participants (out of 68) detected all the indications and four sized all the indications accurately.

To determine whether those less likely to take risks exhibited lower performance in the task (fourth hypothesis), an Independent samples *T* test was carried out. No statistically significant differences in the average detection and sizing performance between the levels of Risk Propensity were found.

By comparing the extent of the agreement with the aid (section 6.6.3) with the overall error rates, it appears that the participants not only agreed with some implemented errors, but also committed errors on indications that were correctly detected and sized by the aid. This

suggests that the aid may have been disused, even when its cues were accurate. To explore this assumption, additional exploratory analyses were conducted.

6.6.7. Additional exploratory analyses

Four analyses were carried out to clarify the obtained results in this experiment. First, the assumption about automated aid's disuse was addressed by examining the performance in working with the "correct" automated aid. Second, the order of occurrence of the automated aid's failures was examined to understand why the experimental manipulation might not have been successful. Third, it was explored whether the participants performed differently depending on their experience with the aid, reflected in the post-hoc evaluation of the aid's and own performance evaluations. Finally, the overall performance of the participants in the task was compared to the performance of the aid to ascertain the quality of the performance together with the aid.

6.6.7.1. Performance with the correct automated aid

Next to the seven implemented errors, the automated aid presented with accurate detection and sizing performance for 31 indications ("correct automated aid"). Expectedly, participants should have agreed with the correct results. In that case, the detection and sizing performance should be 100% accurate. However, this was not the case in this study. Figure 28 shows the percentages of the participants that made the correct (agreed with the correct aid) or incorrect decisions (disagreed with the correct aid) on the correct data.

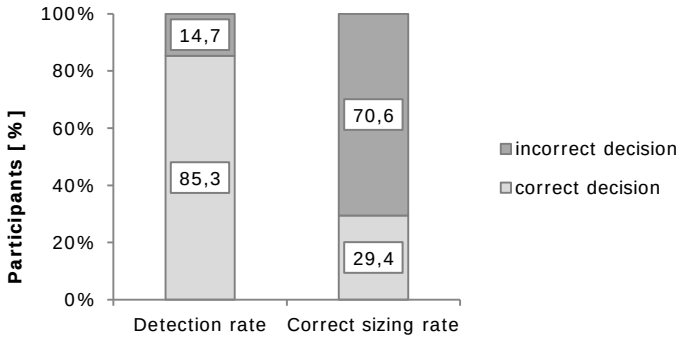


Figure 28: The percentage of the participants that made correct and incorrect decisions in detection and sizing of the correct data

It appears that the participants altered the correct data, especially in terms of indication size.

To examine whether the detection and correct sizing rate are significantly smaller than one (the expected result on the correct data), the mean of the participant sample was compared with the expected value (= 1) of the correct data set using a One-sample *T* test.

The distributions of the errors on the correct data were strongly positively skewed with large kurtosis, as the majority of the participants had a detection rate of 1 (equals 100% detection). These conditions do not satisfy the criteria for the use of a parametric test. However, since the non-parametric alternative reveals the same effects with difficulty in interpretation (a median of the group is found to be significantly different from the same value with which it is compared), a parametric alternative was chosen with restrictions with respect to the significance level. Hence, only those effects with $p < .001$ were held significant.

The results indicate a significant change of the correct data with respect to detection ($t(67) = -2.62$, $p < .05$, $d = -.64$) and sizing rate ($t(67) = -7.27$, $p < .001$, $d = .66$). The participants *reduced* the detection rate ($M_P = .99$, $SD = .03$) and the correct sizing rate ($M_P = .89$, $SD = .12$). Following the afore-mentioned restrictions, only the latter effect is considered significant.

To examine further the nature of aid's disuse, the next question explored was whether there was an association between errors committed on the correct data and the agreement rates in terms of participant and error frequency.

Table 15 shows frequencies of the participants (N), who agreed with the aid in relation to the errors committed on the correct data. Note that since all the participants agreed with at least one implemented error, the frequencies for low (1) and high (2-7) agreement are presented.

Table 15: Association between the agreement with the aid and errors on the correct data

			Frequency of errors on the correct data (yes vs. no)		Total
			no error (0)	error (>0)	
Agreement with the aid	low agreement (1)	Count	0	11	11
		% of Total	0.0%	16.2%	16.2%
	high agreement (2-7)	Count	16	41	57
		% of Total	23.5%	60.3%	83.8%
Total		Count	16	52	68
		% of Total	23.5%	76.5%	100.0%

It was not possible to conduct a 2 (low vs. high agreement with the aid) x 2 (no error vs. error on the correct data) Chi-square test, since the expected frequencies for one cell are less than five (a restriction for using a Chi-square test). However, it may be observed that 23.5% of the participants, who agreed highly with the aid, committed no errors on the correct data, suggesting high reliance on the aid, i.e. misuse. About 16% agreed only with one error of the aid, but committed errors on the correct data, indicating higher reliance on own performance than on the aid. More than 60% of the participants who exhibited high agreement with the aid committed errors on the correct data, suggesting both misuse and disuse. However, the Pearson's test shows very low ($r = -.09$, $p > .05$) insignificant correlation between the agreement with the aid rates and the number of committed errors on the correct data, indicating that the *extent* of misuse of the aid was not associated with the extent of aid's disuse.

A 2x1 Chi-square analysis (no error vs. error on the correct data) was carried out, showing a significant difference between the frequency of participants who committed errors on the correct data ($N = 52$) and those who did not ($N = 16$), $\chi^2(1) = 19.06$, $p < .001$. This indicates that significantly more participants disused rather than appropriately used the correct automated aid.

6.6.7.2. Order of error occurrence

Considering that the reliability of the aid has shown no statistically significant effects on the results, and that the participants exhibited signs of disuse, it was furthermore explored whether the order of occurrence of the aid's failures may have affected the participants' trust in the aid and their consequent behaviour.

The order in which the seven ET images were presented to the participants was randomised. This also meant that the order in which the first aid's error occurred was fully randomised. The order in which the errors occurred along the seven presented images is shown in Figure 29. Note that one of the seven images contained no indications and the rest of them contained from one to fifteen indications per image.

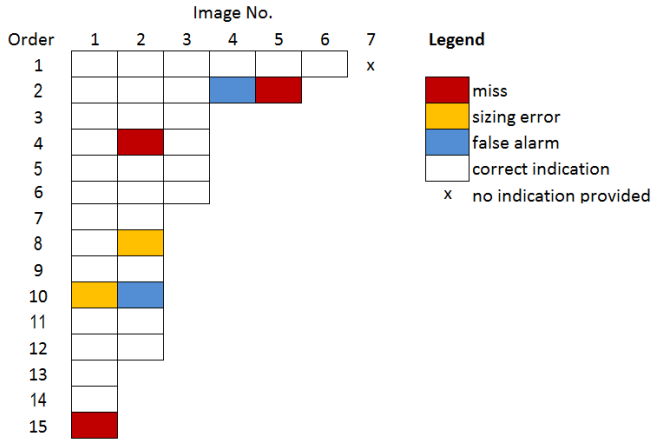


Figure 29: The order of the error occurrence along the seven images. Note: the squared cells represent all indications, and the coloured cells the implemented errors.

This shows that the first aid's error occurred in the first two images, with the first error occurring as second (29.4% of the cases), fourth (50% of the cases), or tenth (20.6% of the cases), among the indications to be evaluated.

However, not all errors were detected by the participants, or at least not reported. Almost 80% of the participants detected the error in the first third of the task (among the first 12 indications), most frequently misses, and sizing errors (Figure 30).

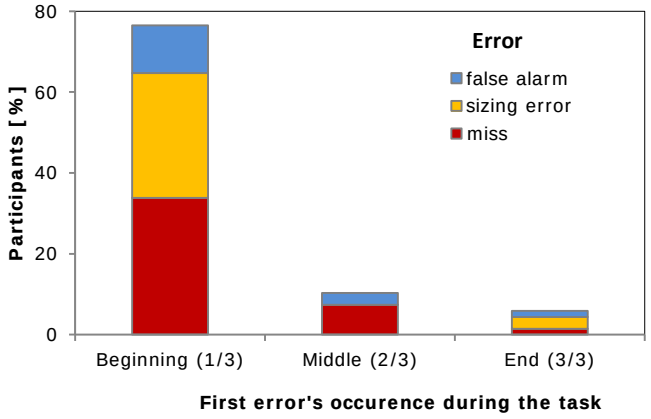


Figure 30: The percentage of the participants detecting the first aid's error with respect to the occurrence during the task and error type

The order of the first error's appearance is found to correlate with the order in which the first aid's error was detected, $r = .67$, $p < .001$ (moderate to strong effect), indicating that early error presentation is associated with early error detection.

Independent samples *T* test was carried out to examine whether there is a difference between early and late detection of the first error (median split) and the agreement with the aid (agreement with all seven implemented errors) and the committed errors on the correct data. It was expected that those who encountered the first failure early in the task would agree less with the aid and possibly disuse the aid, by changing the correct data. The results show that those who detected the first aid's error early in the task (among the first seven indications) agreed significantly *less* with the aid ($M = 2.86$, $SD = 1.56$) than those who detected it later in the task ($M = 4.12$, $SD = 1.56$), $t(61) = -3.20$, $p < .01$, $d = -.82$. No differences were found in the disuse of the aid.

6.6.7.3. Performance evaluation

Another possible effect on the results explored in this analysis was that of the performance evaluation. It was assumed that the participants that evaluated their own performance as superior to that of the aid would be more likely to disuse the aid. However, this assumption was also not supported by the data, since One-way ANOVAs (dependent measure: Errors on the Correct Data) and its non-parametric replacement, i.e. a Kruskal-Wallis Test (dependent measures: Zooming and Sizing) yielded no significant differences between the Performance Evaluation levels in any of the performance measures.

6.6.7.4. Participants' performance vs. the performance of the aid

Taking into account that the participants have shown signs of both misuse (agreeing with the aid's cues or lack thereof) and disuse (changing the correct results of the aid), it finally remained to be established whether the participants, in general, improved the overall performance of the aid or deteriorated it.

The One Sample *T* test was used to establish whether the participants' detection and sizing performance was statistically different from that of the aid. The aid, with its implemented errors had a detection rate of .91 and a correct sizing rate of .94. The results show that the participants improved the performance by exhibiting a higher detection rate ($M = .94$, $SD = .04$) than the aid (DR = .91), $t(67) = 6.75$, $p < .001$, $d = 1.65$. At the same time, the participants deteriorated the sizing performance by having a lower correct sizing rate ($M = .86$, $SD = .12$) than the aid (CSR = .94), $t(67) = -5.75$, $p < .001$, $d = -1.40$. In terms of error frequencies, the participants omitted one indication less ($M = 2.06$, $SD = 1.50$) than the aid ($n = 3$), but committed more sizing errors ($M = 4.82$, $SD = 3.93$) than the aid ($n = 2$).

6.7. Discussion

In this section, the results of the experiment will be summarised and interpreted, accompanied by a critical reflection on the method, with suggestions for improvement, practical implications of the experiment for the NDT practice, and outlook.

6.7.1. Summary and interpretation of the results

The overall aim of this experiment was to explore potential effects of inappropriate automated aid use in NDT. For that purpose, the participants worked with—what was told to them—

either a high or a low reliable automated aid. The reliability of the aid was in fact the same. It was postulated that an induced belief in high reliability of equally reliable aids might affect the perception of the aid's performance. This would in turn have a negative impact on the performance, expressed in higher agreement with the aid as a result of automation bias and higher complacent behaviour. However, this hypothesis was not confirmed.

The participants, however, did misuse the aid to some extent, as visible in their agreement with almost half of the implemented errors and the fact that all participants agreed with at least one error committed by the aid. The participants, who largely agreed with the aid, verified the results of the aid *less* frequently, as it was shown that zooming and, to some extent, sizing behaviour were associated with the tendency to agree with the aid. Zooming and sizing behaviour were also found to correlate strongly, i.e., frequent sizing was associated with frequent zooming. These results indicate relatively high support in favour of the second hypothesis, i.e. that agreement with the errors of the aid is associated with insufficient sampling of information needed to complete the task accurately.

No differences in sizing between those committing only *one* or *more than one* commission error could be explained through different origins of commission errors. Manzey et al.'s (2012, Exp. 2) results suggest three different origins: (1) withdrawal of attention resulting in incomplete crosschecks of information, (2) active discounting of contradictory information, and (3) inattentive processing of contradictory information, analogue to the "looking-but-not-seeing" effect. Hence, it is possible that the participants may have sized an indication, but may have decided to rely on the aid in spite of the discounting evidence or may have not seen that the results are contradictory—possibly due to automation bias.

The third hypothesis was concerned with the individual differences in the propensity to take risks. It was expected that risk-seekers would be less likely to agree with the errors committed by the aid, reflecting in a better performance in the task. This hypothesis was partly confirmed, as risk seekers were found to comply less with the aid, by making significantly fewer *commission* errors. This result partially supports the finding of Madhavan & Wiegmann (2005), who reported that risk-taking is associated with higher disagreement with the aid. However, risk propensity was not found to affect reliance on the aid in terms of omission errors.

The fact that risk propensity had an effect only on commission errors, but not on omission errors, may be explained by different suspected origins of both errors. I.e. whereas commission errors are seen as a result of a failure to take into account all the relevant information and the belief in the superiority of the automated aid, omission errors are believed to occur due to vigilance decrements (Parasuraman & Manzey, 2010; Skitka, Mosier, & Burdick, 2000). Risk behaviour may not be responsible for the latter.

The overall performance in the task was not directly affected by the varied reliability of the aid, performance evaluation or the propensity to take risks, thereby not supporting the fourth hypothesis. However, descriptive data and the subsequent analyses showed a significant decrease in the reliability of NDT, due to both misuse and disuse of the aid.

Even though the effects of automation bias and complacency are typically explored by comparing the measured behaviour to a normative behaviour in the task (e.g. working without the aid), this experiment was built upon an assumption that automation bias and complacency effects *do* happen in interaction with reliable automated aids. For example, those effects have been found also in a field close to NDT, i.e. medical radiology (e.g. Alberdi et al., 2004). Although with lower certainty—since a control group was not present—the results suggest that the aid was indeed misused, as the participants agreed, in average, with almost fifty

percent of the errors committed by the aid (both omission and commission errors were present) and since that agreement was to a large extent associated with the information sampling behaviour. Moreover, all participants agreed with at least one error of the aid. This indicates signs of automation bias and complacency, as the former is typically revealed in omission and commission errors and the latter in insufficient information sampling.

However, it appears that the participants in this experiment did not only misuse the aid, but rather, and to a large extent, disuse it. This was evident from the fact that the participants altered the data, on which the aid was entirely accurate in both detection and sizing performance, thereby reducing the performance of the aid. This was done by almost 71% of the participants. The effect of the change was the most evident in the sizing capability, which was significantly reduced.

Frequency of errors on the correct data was not correlated to the agreement rates. That is, the increase in the agreement with the aid was not associated with a decrease in the frequency of the errors on the correct data, as could be expected. This suggests that there is not a large association between the extent of the aid's misuse and disuse.

Several possible explanations for this type of disuse can be found in the literature. The first refers to the perceived reliability of the aid. If low, operators could become under-dependent on the aid, in a way that they ignore it, even when it may be correct (Dixon & Wickens, 2006). In this study, the participants were confronted with an aid that committed errors on 18% of the task. Previous studies on automation misuse have simulated situations with from 12% up to about 35% of the task containing errors (Alberdi et al., 2004; Madhavan et al., 2004; Parasuraman et al., 1993; Skitka et al., 1999; Skitka, Mosier, Burdick, et al., 2000). This was done so in order to obtain sufficient data points to describe the performance in the task, as well as to simulate different error possibilities (c.f. Alberdi et al., 2004). More recent studies by Manzey and colleagues (Bahner, 2008; Manzey et al., 2012, Exp. 2) show that automation bias and complacency effects can be revealed when working with an aid that commits only *one* error (the studies focused on commission errors). Taking into account that studies with greater error rates have found evidence of misuse rather than disuse, and that disuse is more likely to happen when working with systems with reliability lower than 70 % (Wickens & Dixon, 2007), it implies that the error rate might not have been a direct cause of disuse, or at least not the only one.

The second possible explanation refers to the experience of automation failure. Typically, it will diminish trust. That effect will be enhanced if the participants are led to believe that the aid is near perfect (*positive framing*; Dzindolet et al., 2002). If they expect the automated aid to be near-perfect and an aid commits an error, that expectation is violated. This leads the participants to underestimate the reliability of the aid (Dzindolet, Pierce, Beck, & Dawe, 1999). Continuous errors will furthermore negatively affect trust (Parasuraman & Riley, 1997) allowing for better calibration of the true reliability of the system so that the detection performance of the failures improves (Parasuraman & Wickens, 2008; Rovira et al., 2007). That kind of negative experience has shown to have a much stronger effect on operator trust than positive experience (Manzey et al., 2012, Exp. 2). In those cases, people were found to rely on self, rather than on the aid, as the detection of automation failure boosts one's self confidence (Madhavan et al., 2006). If, in general, the trust in self is higher than the trust in aid, people will disuse the aid (Lee & Moray, 1992, 1994).

Following this rationale, an attempt had been made to explain the misuse and disuse observed in this study by looking more deeply into the potential effects of the first failure and the

experience with the aid. It was expected that those who encountered the first failure early in the task would agree less with the aid and possibly disuse the aid, by changing the correct data.

The first aid's error was presented to the participants very early in the task, which seems to have affected the participants' trust in the aid to some extent. That is, those who detected the first error early were more likely to disagree with the aid. Early error detection was found to be associated with the early error presentation. Still, the fairly early detection of that failure did not seem to affect the disuse.

The evaluation of the performance evaluation after having experience with the aid, i.e. the reported utility of the aid (trust in self [superior self] vs. trust in the aid [superior aid]), has offered no further explanations for the obtained effects in the study, as the evaluations were not related to the observed misuse and disuse of the aid. That is, evaluating own performance as superior to that of the aid did not lead to increased self-reliance (i.e. lesser agreement with the aid and more errors on the correct data), nor did evaluating the aid as superior lead to higher reliance on the aid. It may be that the simple two questions asking to evaluate own and aid's performance may not have been sufficient to describe the complete experience with the aid. Further questions may be necessary, e.g. questions related to the experience of the first and the consequent failures, the evaluation of the difficulty of that failure, the decision to rely or not rely on the aid, assessment of own abilities, etc.

The effects of disuse could not easily be explained by exploring the collected data and by the existing theories. Rather, it seems more likely that some *other* factors may have played a role in disuse, possibly related to the task itself.

As stated earlier, the participants disused the aid by removing critical indications and by altering the indication size. One of the reasons for frequent sizing deviation may be that the participants used a subjective sizing criterion, not complying with the training. The participants were trained to *include* only those pixels in the direct contact (left, right, above, below) with other pixels belonging to the indication and to *exclude* those in diagonal contact (see Figure 18 in section 4.2.2.2 for the sizing criterion). Thus, it may be that the participants violated this rule.

Another possible reason may stem from incorrect establishing of the sizing criterion. Since a different sizing criterion was used for each indication, the participants were instructed to enter the maximum pixel intensity into a pre-designated cell in the reporting spreadsheet in Excel, which was designed to provide them automatically with the sizing criterion for that indication. Had the participants failed to do so, had they mistakenly applied the criterion from another indication, or had they entered the values into the wrong spreadsheet (assigned to another image), this would produce a faulty sizing criterion and, thus, may explain the deviation from the correct size.

Whereas the sizing deviation could be explained by some kind of a subjective or faulty sizing criterion, the fact that some critical indications were *discarded* can only be explained as a possible mistake in judgment. A plausible explanation may be that the participants misjudged an indication (i.e. the one signalling a defect in the material they were instructed to find) for a geometrical indication (i.e. the one that the participants were taught to exclude from the evaluation). The reliability in detection performance was decreased only to a small and insignificant extent—from 1 to .99 (detection rate)—indicating this was a rare occurrence.

In conclusion, this study revealed both misuse and disuse of the detection and sizing aid used for the evaluation of NDT data, which had a significant effect on the reliability of NDT. Whereas the flawed *detection* performance of the aid was improved by the participants, the *sizing* performance was significantly worsened. Individual differences in risk taking were found

to explain the differences in the compliance with the aid and the verification behaviour was largely consistent with the agreement rates. These results provide further evidence for the significance of individual differences in risk taking and complacent behaviour on interaction with automated aids. A large portion of the participants behaved complacently towards the aid, by not sampling all the necessary information to complete the task completely and accurately. Whereas decreased misuse was found to be affected by the early detection of the aid's failure, increased misuse occurred due to insufficient verification of the data (opposite to the instruction) and reveals possible effects of automation bias on working with the aid. Disuse—evident in extensive sizing deviation—was not related to the misuse, which led to the conclusion that it may have stemmed from misunderstanding of the task or due to mistakes in handling reporting sheets.

This study showed that the belief in the reliability of the aid does not affect the perception of the performance of equally reliable aids. This result seems to be in favour of what was suggested by some researchers, that only the *actual* reliability may affect trust in automation and, consequently, the performance (e.g. Riley, 1996). This may be, considering that the participant's own experience with the aid's failures affected the consequent agreement with the aid. However, if the suspected reasons have led to the disuse of the aid (e.g. misunderstood or wrongly applied sizing criterion), it may also be that they unjustly lowered the trust in the aid in both experimental conditions. Hence, based on the results of this study, it is not possible to establish *with certainty* that the induced belief in the high or low reliability of the aid would affect inspectors' interaction with the aid.

6.7.2. Limitations of the study

One of the major problems of this study refers to the order of errors' occurrence, as the agreement with the aid was significantly affected by the early appearance of aid's failures. This occurred because the indications were spread along only seven images. These images did indeed stem from an actual inspection of seven different welds, but the welds contained a number of defects, which was untypically large for the practice (as the defects were purposefully produced during the manufacturing process for the purposes of the development of NDT methods and techniques). This caused the first aid's failure to occur very early in the task and affect trust in the aid, regardless of the experimental instruction. Future studies should opt for spreading the indications over a larger number of images, a delayed presentation of the first failure (e.g. Bahner et al., 2008), and consider using the paradigm of confronting the participants only with one failure, as suggested by recent studies (Manzey et al., 2012, Exp. 2; Parasuraman & Manzey, 2010).

The reason for choosing several failures of different type was to simulate different conditions that can occur during an evaluation of NDT practice with a faulty aid. However, it is not likely that all error types would occur in the same evaluation, as those types of errors may probably stem from faulty sensitivity settings, which may then systematically produce similar error types, i.e. either misses and sizing errors, *or* sizing errors and false alarms. Future studies may focus separately on these different error possibilities.

Furthermore, the training may have negatively framed the participants into distrusting the aid. During the training, they are taught *how* to control and what they should do *in case* the aid commits an error. Not only the training, but also the experimental instruction, may have been unsuccessful in inducing high trust in the reliability of the aid, as the participants were not told that the aid is *always* correct, but instead that it was *almost always* correct, following suggestions from the practitioners.

The *potential* problem with the reporting sheets, i.e. the resulting faulty decision criterion, may not have happened in reality, since the sizing criterion would have been established in a different way. Hence, further efforts should be invested into avoiding problems with handling spreadsheets.

One important shortcoming in terms of field research refers to the complexity of the task, which may have been oversimplified in this study. The task was simplified by simulating the NDT task using simpler software, i.e. software not requiring extensive training, experience, and skills. Since the use of automation-aided defect detection is not yet that wide-spread, acquiring experienced and qualified participants is limited to only a few. More importantly, the currently used software differs significantly among the users, making it even harder to acquire participants. Lack of qualified participants with specific knowledge and skills is a major problem for studies in the field of NDT. Using NDT trainees as participants may be a better alternative to sampling students, as they share the sense of responsibility and understanding of the severity of consequences if the NDT inspection shows to be insufficiently reliable, as well as understand the complexity of the field applications. However, this study could profit more from experienced personnel.

Including measures of trust, attitudes towards automation, and perceived reliability of the aid *before* the experimental task, more reliable measures of risk propensity, and experience with the aid *after* the task, may additionally clarify issues of inappropriate use of automated aids in this application.

6.7.3. Implications of the study for the NDT practice

NDT is expected to provide with completely accurate results, i.e. all critical defects need to be detected, their size correctly estimated and, preferably, the number of false alarms should be kept at a minimum. The obtained results suggest not only that the performance in the NDT data evaluation task deviated from the expected flawless performance required in the practice (to satisfy the safety requirements of the components and to comply with high reliability standards expected from the utilised NDT methods), but also that the participants committed *more* errors in the task than expected.

This study revealed different opportunities for inappropriate use of automated detection and sizing aids in the evaluation of NDT data. While disuse can be assigned to possible shortcomings in the experimental design, the fact that the participants agreed with almost half of the implemented errors, even in light of early error occurrence in substantial quantity, raises concern for the reliability of NDT.

First, this study opens a possibility that aids can be misused. The fact that the agreement with the errors was to a large extent associated with complacent behaviour emphasises the need for following of the inspection procedure, which has shown not always to be used as it should (e.g. McGrath, 2008). An inspector that fully complies with the procedure will follow all steps and be less likely to be complacent. However, compliance with the procedure is dependent on the quality of the procedure, i.e. on the understanding if its content and its usability (Bertovic & Ronneteg, 2014). However, this alone may not be sufficient to decrease inappropriate reliance on the aid.

Just informing inspectors that the automated aids can commit errors has not shown to be successful enough in decreasing automation bias (Bahner et al., 2008; Skitka, Mosier, Burdick, et al., 2000). However, direct individual experience with the failures during training yielded a decrease in trust (Dzindolet et al., 2003) and, consequently, an increase in the verification behaviour and time spent identifying failures (Bahner et al., 2008). Dzindolet et al. (2003)

recommend realising that people can be biased towards automated aids and providing with instructions and experience (training) as means for establishing appropriate trust in the aid. Thus, acknowledging that automation bias might be a problem in NDT is the first step toward tackling the issue. And the possibility of the aid's failure should be implemented in the training of the NDT personnel.

Moreover, providing inspectors with information about *why* the automated system might err may lead to a more appropriate trust and selection of better discriminating strategies—by allowing them to better discriminate between reliability of the aid and their own reliability (Dzindolet, Pierce, et al., 2001; Lee & See, 2004; Madhavan et al., 2006). In addition, the knowledge about *how* the automation works leads to a more appropriate automation use (Parasuraman & Riley, 1997).

Since there are large individual variations in how people use automation, policies and procedures should highlight the importance of taking specific considerations into account (Parasuraman & Riley, 1997). The found effect that risk takers are less likely to rely on the aid, suggests that forms of risk taking should be encouraged in NDT. However, this should not refer to violations of the procedure. Instead, the inspectors should be encouraged to speak openly about inconsistencies during the task and be discouraged from blindly relying on the inspection procedure or on the suggestions of the aid. This should be considered during the personnel training.

This study may be criticised for employing insufficiently experienced NDT apprentices instead of professional personnel, as NDT community typically relies on the qualification and the experience of the inspectors, when expecting high levels of performance. However, studies have shown that reliance on experience of the inspecting personnel will not guarantee immunity to automation bias, as shown by Mosier et al. (1996), who observed that experience and expertise led to a *greater* tendency of relying on automated cues. Moreover, automation bias effects have been observed in both students' and professionals' samples.

On another note, automated detection and sizing aids are not widespread in NDT practice yet, which may make NDT inspectors initially suspect the reliability of the aid, rather than to trust it blindly. This, in combination with the afore-mentioned factors (e.g. early error occurrence, easy errors, experience with faulty aid, etc.), may lead to unnecessary disuse of the aid even when its recommendations are correct. As observed in this study, increased reliance on own judgement instead of that of the aid, may lead to excessive sizing errors. Self-reliance is not always desired, especially if the aid is accurate and the person errs, as observed in this study.

Although the task may not have been that easy for the participants, the detection and characterisation of data collected with eddy current method is generally seen as an easy task in NDT. However, errors on easy tasks, as suggested by Madhavan et al.'s (2004, 2006) "easy error hypothesis", lead to underestimation of the aid's reliability, lowered trust in the aid, and higher self-reliance (c.f. Dzindolet et al., 2003).

Even though detection and sizing aids in NDT are not designed to be blindly relied on, i.e. the inspector is typically instructed to control all data (especially the size of the indications); the type of disuse encountered in this study—in terms of unmerited excessive sizing deviation—represents a further threat to the reliability of the evaluation. Inspectors should be trained to consult others or control the equipment settings in case numerous cases of aid's failures occur. Combining different types of feedback of aid's performance has shown to additionally decrease automation disuse (Dzindolet et al., 2000).

In conclusion, this study reveals that interaction with automated systems may lead to risks, previously not considered by the designers of NDT detection and sizing aids. Encouraging participants not to agree with the aid in case of contradictory information and encouraging them to speak openly about problems encountered during the evaluation, placing emphasis on building appropriate trust towards the aid based on experience with the aid, rather than information about it, and generally being aware that automated systems may fail and that people may fail to notice it, are some of the suggestions for NDT community to consider.

6.7.4. Outlook

NDT is a field, in which automation plays an increasingly important role. Hence, the interaction with automated systems and automated aids in NDT is a topic worth further exploration. Future studies should venture more into the field and find ways to examine interaction with actual aids that carry numerous complexities with respect to the task and indication interpretation.

Belief in high reliability could, furthermore, be put to test using a more strict method, i.e. by lowering the number of aid's failures, by raising the task complexity, by better distinguishing between high and low reliability of the aid in terms of experimental instruction and, moreover, by delayed occurrence of aid's failures.

Eye tracking may provide with additional insights into the data evaluation process and decision-making. This would also allow for distinguishing origins of omission and commission errors, as well as give further insights into the difficulties encountered during sizing, which appears to place higher cognitive demands on the inspectors than detection.

7. General discussion

Non-destructive testing is regarded as one of the key elements in ensuring quality control of engineering systems and their safe use. Reliability of NDT, typically assessed in relation to the technical capability of the system, is known to also depend on human factors, an element in the reliability chain that has not yet been sufficiently understood. Substituting manual NDT with mechanised NDT—a form of automation-assisted inspection—is generally seen as a good method to decrease variability in the inspection results and, thus, to increase the reliability of NDT. However, the potential risks involved in this application have never been investigated.

The overall aim of the presented work was to obtain insights into the potential problems of applying mechanised testing in NDT (with emphasis on the nuclear domain) and find ways of mitigating their effects on the inspection performance. In doing so, the aim was to address some of the current challenges of the NDT field, primarily the missing knowledge.

To address the first two objectives, i.e. to identify and analyse potential risks in mechanised NDT and devise measures against those risks, a risk assessment technique was employed (Study 1). The results of this assessment served as a starting point to address the second two objectives, i.e. to critically assess the preventive measures and suggest ways for their implementation. For that purpose, two additional empirical studies were conducted. In the first (Study 2), the human redundancy, a suggested measure of error recovery and in the second (Study 3), the use of automated aids in the evaluation of NDT data was put into focus.

The first study revealed the potential for failure in the application of mechanised testing during both the acquisition and the following evaluation of data, stemming from technological shortcomings and potential technical failures, but also from the individual errors (mainly unintentional), and the organisation (e.g. the working environment, inspection procedures, etc.). The tasks assigned the highest risk priority, in both acquisition and evaluation of data, were tasks associated with detection and sizing of indications, during which the human inspector plays the key role. Whereas some of the errors can be detected through consecutive steps, errors in the evaluation of data (e.g. missing defects) can slip through the net of the existing barriers, presenting, therewith, with the highest risk for the reliability of NDT. Improved inspection procedures, human redundancy, automation-aided indication detection, interpretation aids (i.e. defect catalogue), and attention to training and to hiring of experienced personnel were the preventive measures suggested to aid in error avoidance or error recovery in the evaluation. The conclusion of the study was that before they can be implemented and

expected to serve their purpose, those preventive measures need to be carefully considered with respect to *new* potential risks, thus serving as a starting point for further empirical study.

The second study was concerned with potential decrements in the performance that may result from applying sequential human redundancy in the evaluation of NDT data. The potential performance costs in sequential redundancy were addressed only by a handful of scientists (Clarke, 2005; Conte & Jacobs, 1997; Swain & Guttmann, 1983) and not given sufficient attention in the NDT field. However, instead of decrements, as hypothesised, the study revealed that the first redundant inspector, led to believe his partner will perform the task after him, was shown to provide the same amount of effort, as when working alone. The second, i.e. the redundant checker, compensated for a low reliable predecessor. The information about the high experience of the first redundant inspector was not successful in inducing a belief in his superior performance due to its actual performance being very poor – a result that could be assigned to the experimental design. These results were not in line with the hypotheses, which expected performance and reliability decrements due to the effect of social loafing (Clarke, 2005; Karau & Williams, 1993; Manzey et al., 2013; Marold, 2011; Skitka, Mosier, Burdick, et al., 2000). Both results suggest that the outcome of the evaluation task may have been highly valued by the participants, and that they may have felt their input to be instrumental to achieving that end, which may explain their motivation and effort in the task, consistent with the Collective Effort Model by Karau & Williams (1993).

In the third study, the use of automated aids in NDT was explored. Motivated by the studies that suggested that in interaction with highly reliable aids people are likely to rely on them due to the bias towards automation, resulting in aids' misuse (e.g. Parasuraman & Riley, 1997); this study addressed a prevalent belief held by the NDT community in the high reliability of automation. The perception of equally reliable aids was not affected by the induced belief in high or low reliability of the aid, as hypothesised. The suggested explanation was found in the early occurrence of aid's failures (as the majority was presented and detected by the participants in the first third of the task). The results, however, revealed signs of both use and misuse of the aid, that can affect the reliability with which inspections are carried out. Misuse was partly explained with propensity to take risks (as low risk-seekers were found to agree more with the aid) and, largely, by decreased verification behaviour. The displayed disuse was observed in the fact that the participants significantly altered even the data correctly evaluated by the aid, thereby significantly decreasing the overall performance in terms of accurate sizing. This type of disuse was not associated with misuse, indicating that different mechanisms were responsible for the behaviour of the participants who tended to misuse or disuse the aid. Whereas explanations for misuse can be found in the automation bias and complacency research (e.g. Parasuraman & Manzey, 2010), the disuse was assigned to problems in establishing the criterion for indication sizing or mishandling of the reporting sheets that were designed to provide the sizing criterion.

These empirical studies—even though not all entirely successful in confirming their hypotheses—showed yet again that variability between the inspectors might affect NDT reliability. Due to the experimental setup, they revealed increases in the performance, but those conclusions need to be taken with care. The common limitation of these two empirical studies was that both the first redundant inspector and the automated detection and sizing aid were insufficiently reliable. In the daily practice, it is unlikely that an experienced inspector or a well-designed automated aid would commit many errors. Since trust in automation and trust in other people develop over longer periods of time working with a highly reliable counterpart, future studies need to involve periods of reliable performance to establish a corresponding level of trust. Considering this limitation, it is not possible to reject entirely the

hypotheses made in these studies and to conclude that redundant checkers will compensate for their predecessor or that a strong belief in the reliability of the aid will not affect performance with high reliable aids. Interestingly, even though working with a fairly unreliable aid and being confronted with the errors early in the task, the participants still complied with the aid (resulting in lower detection and sizing performance) and in turn, decreased the overall reliability.

As frequent in human factors studies employing students, the question of ability to generalise the obtained results can be raised. The latter two studies were carried out predominantly with NDT trainees (86.4%; 13.6 % trainers and researchers). Even though they differ from regular students insofar that, they have the knowledge and, to some extent, practical experience in NDT, they still do not represent the population of experienced NDT personnel very well. In addition, the participants were not experienced in data evaluation, which was the experimental task. Albeit not completely representative of the experienced field inspectors, NDT trainees carry benefits over populations with no experience in NDT. These benefits refer to understanding of the context for purposes of which the study is carried out and to the relatively high motivation for the positive outcome of the task. The suspected motivation for the outcome of the task, and consequent effort invested in its completion (both social compensation and automation disuse suggest extra effort had been invested in the task) suggest NDT trainees' awareness of the dangers lurking behind unreliable performance and willingness to aid in further development that can probably not be observed in regular students' populations.

In the first study, the participants were not certified regular inspectors, but rather developers of NDT methods in question. Taking into account that only a few people are familiarised with the investigated application, this was the only approach possible.

Further limitation of the empirical studies refers to the simplification of the task, which may make NDT community question how this can apply to their task. Instead of working with the actual evaluation software, which carries along various complexities, the task was narrowed down to simple signal detection and pixel counting. As indicated in the previous chapters, the wide variety of used software and the prerequisite for successful completion of the task in terms of knowledge, training and experience is the major limitation to human factors study in the data evaluation process. Future studies may opt for a qualitative study of the data evaluation process. However, issues of trust in people and in automation span over a variety of tasks and applications, and the complexity of the data evaluation may introduce new variability related to handling of new software, other than the variability stemming from trust—in focus of the study—in a more familiar "realistic" task.

The identification of risks in mechanised NDT was conducted in the field of spent nuclear fuel management. The methods discussed were not fully developed and are only used for inspections in the development of the manufacturing processes. Considering that the spent nuclear fuel disposal is set to start with operation in the future, the prospective approach to risk assessment was the only one possible, and the outcomes of the analysis served their purpose. However, another question with respect to generalisability can be whether the findings can be applied to other methods and applications. Although the way NDT is applied can differ depending on the conditions under which it is carried out (suggesting that further study of risks involving mechanised NDT in other active applications may be necessary for a more general understanding of the existing risks), some of the identified risks, their causes and consequences, can also be transferred to other applications. Even physically completely different methods, such as ultrasonic testing and radiography, were found to have similarities with respect to analysed risks and their properties, as presented in this dissertation.

According to Moray & Huey (1988), access to realistic settings—to facilities and to people, such as the experienced operators—is one important barrier to effective human factors research. However, as McGrath (2008) pointed out, even when the extreme conditions typical for the practice were removed or minimised, the studies of human factors in NDT conducted under controlled conditions (e.g. laboratory or 'mock-up' industrial environments) still observed considerable variation in the performance (present study included, e.g. Study 3).

In spite of their limitations, the carried out studies successfully raised the question of reliable mechanised NDT that can be affected by factors previously never considered by the designers or by the managers in NDT. This also refers to raising awareness that the variability in the data evaluation can be assigned to factors other than inexperienced or unmotivated personnel, inappropriate procedure, or unreliable equipment, but also to interactions between people and interactions with technology.

The current research indicates that human factors are embedded in a system of task, technology, and organisation. In this system, the human element plays a key role, as only humans are able to classify events, anticipate risks, and develop and implement adequate preventive measures (Badke-Schaub et al., 2012). However, reports on human error rates and their contributions to events and accidents regard human as a source of risk, rather than as a problem solver, which frequently leads to increased use of automation, implementation of redundancy, and more strict procedures. Albeit useful in fostering safety, these measures can backfire if not implemented with care (e.g. Bainbridge, 1987; Parasuraman & Riley, 1997; Reason, 1997; Sagan, 2004). For example, redundancy may not only be counterproductive due to social loafing and shirking of responsibility but also (a) because it increases complexity and can make systems prone to common-cause failures; (b) because it makes systems more opaque, making individual failures likely to remain unnoticed, accumulate over time, and become latent; and (c) because it creates a false impression of safety, which can lead to gradual degradation of safety margins in pursuit of efficiency and profits (Kettunen, Reiman, & Wahlström, 2007; Rasmussen, 1997; Sagan, 2004). Automation, on the other hand, changes what people do. This can create new demands, e.g. demands for knowledge, skills, and training (e.g. Sarter, Woods, & Billings, 1997), and lead to new classes of errors than those it was designed to prevent (Skitka et al., 1999). And, finally, more strict procedures will not necessarily be used as designed, putting the issue of their usability ever more in focus (Bertovic & Ronneteg, 2014; McGrath, 2008).

What does this mean for the NDT practice? How are redundancy and automated aids to be implemented without the costs associated with social influences and over trust in automation? The first step is to raise awareness that defences may fail. This awareness can be fostered by the regular use of risk identification methods, by further research, and by introducing human factors training into the regular training of the NDT personnel. This especially applies to the potential influences of bias towards automation. The conducted study indicated that only informing inspectors that the aid is not very reliable would not necessarily lead to a lesser reliance on the aid. Instead, actual experience with the aid that may err and the understanding of why the aid may err constitute as better strategies in decreasing uncritical reliance. The studies on human redundancy, on the other hand, stress that profits may only be expected if the redundant operators are independent of each other (e.g. Clarke, 2005; Sagan, 2004). In the practice, this may not always be easy to achieve, since some familiarity is bound to be present among the redundant individuals in small organisations, such as the spent nuclear fuel management. Thus, valuable strategies include increased individual accountability for the outcome of the redundant team, awareness of the evaluation potential by the supervisors and other inspectors, and including both redundant inspectors in the decisions made after the

inspection. Important is also that the previous inspection results are not provided, and that redundant inspections are not conducted only on areas where potential failures may be expected. Instead, larger areas should be re-inspected, therewith increasing the chance of identifying defects previous inspector may have overseen. Instead of implementing redundancy at all times, which may lead to common-cause failures and increased impression of safety, applying redundancy at irregular intervals, changing its frequency in dependence on the recovered errors, and taking human diversity into account, may be useful approaches.

The observed benefits of risk taking tendencies in decreased reliance on the aid (Study 3) do not indicate that risky behaviour should be supported in NDT. Instead, this result shows that risk takers are more likely to question the existing procedures and not comply with them if they feel something is wrong. Even though in cases of its inappropriateness, violations of the procedure have sometimes shown to lead to a more reliable outcome (Reason, 1997), in the majority of the cases procedure violations are undesired, as they diminish the purpose they were designed for, i.e. to prevent errors. Since following the inspection procedure is a behaviour that is demanded in NDT and upon which reliable inspections rely, inspectors should be encouraged to speak up about inconsistencies with their supervisors and their remarks should be carefully considered. Although the inspection procedures may contain all the relevant information and all the necessary steps that need to be taken, procedures not understood, or not usable, will most likely be counterproductive to their purpose.

All this suggests that human factors need to be carefully considered in the design of the inspection process and the inspection procedures. Moreover, the consideration of human factors in the NDT inspection for the purposes of spent nuclear fuel management showed that addressing human factors can be valuable not only for methods, which are in service, but also aid in the development of NDT methods and procedures. The consideration of human factors and their effects on the reliability of NDT aids in the development and in validation of the quality of the manufacturing and welding processes. The gained knowledge can also serve as a starting point for establishing the working environment and the working practices when the repository starts with operation.

Human redundancy and automated aids may not be the only preventive measures that can backfire. Further study into the potential performance-degrading factors in both acquisition and evaluation of data and their potential defences is necessary. Future topics worth exploring in this domain include, e.g. further identification of hazards, their prioritization, and development of strategies to minimize their effects. The focus should also be put on interactions between individuals and other elements of the system, i.e. the organisation. Moreover, the joint decision-making in combining results of several methods to assess the condition of the component should be considered, as well as the usability of the evaluation software and the inspection procedures.

Considering that the NDT community is still primarily technology-oriented, and—with respect to human factors—person-oriented, further transfer of knowledge from social sciences to NDT is needed in order to raise awareness of influences not previously considered by the designers of the system. Developing human factors training strategies for the inspection personnel, NDT developers, software designers, and the management is a direction worth pursuing. Most importantly—even though it extends beyond capabilities of a scientific study—ways have to be found how findings from scientific studies can be implemented in the NDT practice, therewith closing the communication and implementation gaps (Bertovic et al., 2014).

Although human factors are commonly seen as opposite to the technical factors and they are typically addressed separately, it is exactly the interaction between those factors that is of highest relevance. High technical and high human reliability do not necessarily lead to a reliable system (Giesa & Timpe, 2002). In extension, NDT reliability analyses focusing only on the intrinsic capability may not reveal the system's true capability when applied in the field, as the interaction with other factors is neglected. The Modular Reliability Model (Müller et al., 2013), suggesting reliability of NDT depends on the intrinsic capability, application factors, and human and organisational factors, was conceptualised as a starting point by indicating which factors need to be considered in the reliability assessment. Future direction reliability analyses should take may be to strive to integration of the concepts and focus on interactions between the modules. Due to a lack of understanding of human factors, reliability assessments are ever so more relying on simulations to determine the "true" capability of their system by reducing the costs of expensive experiments and by *excluding* the "human factor" (Chapuis et al., 2014). However, the actual performance of a system cannot exceed the capability for which it was designed, but it may be diminished by an interplay of people with systems and the specific conditions under which inspection is conducted, which suggest simulated data may not be fully applicable for the use of the system in the field. The human factors' perspective on the need to equally develop technology and people was eloquently summarised by Timpe (1993): "*In psychology it is theoretically impossible to prove that human and machines can be equated, so from the psychological point of view there is little use in attempting to foster reliability or safety of the entire person-machine system by reducing the "human factor*" (p. 119).

References

Albanese, R., & Van Fleet, D. D. (1985). The free riding tendency in organizations. *Scandinavian Journal of Management Studies*, *2*(2), 121–136. doi:10.1016/0281-7527(85)90003-9

Alberdi, E., Ayton, P., Povyakalo, A. A., & Strigini, L. (2005). Automation bias and system design: a case study in a medical application. In *Paper presented at "The IEE and MOD HFI DTC Symposium on People & Systems--Who are We Designing For?", 16-17 November 2005, London, UK.* (pp. 53–60).

Alberdi, E., Povyakalo, A. A., Strigini, L., & Ayton, P. (2004). Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology*, *11*(8), 909–918. doi:10.1016/j.acra.2004.05.012

Alberdi, E., Povyakalo, A. A., Strigini, L., Ayton, P., Hartswood, M., Procter, R., & Slack, R. (2005). Use of computer-aided detection (CAD) tools in screening mammography: a multidisciplinary investigation. *The British Journal of Radiology*, *78*, 31–40. doi:10.1259/bjr/37646417

Algedri, J., & Frieling, E. (2001). *Human-FMEA*. Munich: Carl Hanser Verlag.

Ali, A.-H., Balint, D., Temple, A., & Leevers, P. (2012). The reliability of defect sentencing in manual ultrasonic inspection. *NDT & E International*, *51*, 101–110. doi:10.1016/j.ndteint.2012.04.003

Annett, J. (2004). Hierarchical task analysis. In N. Stanton, A. Hedge, K. Brookhuis, E. Salas, & H. Hendrick (Eds.), *Handbook of Human Factors and Ergonomics Methods*. Boca Raton, FL: CRC Press.

Annis, C., & Gandossi, L. (2012). *ENIQ TGR technical document: Influence of Sample Size and Other Factors on Hit / Miss Probability of Detection Curves [ENIQ Report No. 47]*. Petten, The Netherlands: European Commission, Joint Research Centre, Institute for Energy.

Arabian-Hoseynabadi, H., Oraee, H., & Tavner, P. J. (2010). Failure Modes and Effects Analysis (FMEA) for wind turbines. *International Journal of Electrical Power & Energy Systems*, *32*(7), 817–824. doi:10.1016/j.ijepes.2010.01.019

Asch, S. E. (1955). Opinions and Social Pressure. *Scientific American*, *193*(5), 31–35. doi:10.1038/scientificamerican1155-31

Badke-Schaub, P., Hofinger, G., & Lauche, K. (2012). Human Factors. In P. Badke-Schaub, G. Hofinger, & K. Lauche (Eds.), *Human Factors. Psychologie sicheren Handelns in Risikobranchen. 2. Auflage* (pp. 3–20). Berlin Heidelberg: Springer.

Bahner, J. E. (2008). *Übersteigertes Vertrauen in Automation: Der Einfluss von Fehlererfahrungen auf Complacency und Automation Bias*. (Doctoral dissertation. Technische Universität Berlin). Retrieved from https://opus4.kobv.de/opus4-tuberlin/files/1944/bahner_jennifer.pdf

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, *66*(9), 688–699. doi:10.1016/j.ijhcs.2008.06.001

Bainbridge, L. (1987). Ironies of Automation. In J. Rasmussen, K. Duncan, & J. Leplat (Eds.), *New Technology and Human Error* (pp. 271–283). Chichester, UK: John Wiley & Sons.

Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2007). Automation Usage Decisions: Controlling Intent and Appraisal Errors in a Target Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(3), 429–437. doi:10.1518/001872007X200076

Behravesh, M., Karimi, S., & Ford, M. (1989). Human factors affecting the performance of inspection personnel in nuclear power plants. In D. Thompson & D. Chimenti (Eds.), *Review of Progress in Quantitative Nondestructive Evaluation* (Vol. 8B, pp. 2235–2242). New York, NY: Plenum Press. Retrieved from http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=2510&context=qnde

Bell, A., Munley, G., Rowley, K., McGrath, B., & Bainbridge, H. (2012). Personality traits and cognitive abilities associated with manual ultrasonic operator performance. In J. Wilson, A. Mills, T. Clarke, J. Rajan, & N. Dadashi (Eds.), *Rail human factors around the world* (pp. 696–706). Leiden, The Netherlands: CRC Press.

Bennett, N. (2004). Withholding Effort at Work: Understanding and Preventing Shirking, Job Neglect, Social Loafing, and Free Riding. In R. Kidwell & C. Martin (Eds.), *Managing Organizational Deviance* (pp. 113–130). Thousand Oakes, CA: SAGE Publications, Inc. doi:10.4135/9781452231105.n5

Bento, J. (2002). *Procedures as a Contributing Factor to Events in the Swedish Nuclear Power Plants [SKI Report 02:63]*. Nyköping, Sweden: Swedish Nuclear Power Inspectorate (SKI). Retrieved from http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/42/022/42022489.pdf

Berens, A. P. (1989). NDE Reliability Data Analysis. In *ASM Metals Handbook, Volume 17, Nondestructive Evaluation and Quality Control* (9th Ed., pp. 689–701). Materials Park, OH: American Society of Metals International.

Bertovic, M. (2014). *Identifying and Managing Risks in Mechanized NDT: A Human Factors Study [SKB document ID No. 1427252, available per request]*. Oskarshamn, Sweden: Svensk Kärnbränslehantering AB.

Bertovic, M., Calmon, P., Carter, L., Fischer, J., Forsyth, D., Holstein, R., … Selby, G. (2014). Summary of the "Open Space Technology Discussions" on Reliability of NDE. *Materials Testing*, *56*(7-8), 602–606. doi:10.3139/120.110604

Bertovic, M., Gaal, M., Müller, C., & Fahlbruch, B. (2011). Investigating human factors in manual ultrasonic testing: testing the human factors model. *Insight*, *53*(12), 673–676. doi:10.1784/insi.2011.53.12.673

Bertovic, M., & Ronneteg, U. (2014). *User-centred approach to the development of NDT instructions [SKB Report R-14-06]*. Oskarshamn, Sweden: Svensk Kärnbränslehantering AB. Retrieved from www.skb.se/publication/2718046/R-14-06 .pdf

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Auflage.). Berlin-Heidelberg: Springer.

Brickner, M. A., Harkins, S. G., & Ostrom, T. M. (1986). Effects of personal involvement: Thought-provoking implications for social loafing. *Journal of Personality and Social Psychology*, *51*(4), 763–769. doi:10.1037/0022-3514.51.4.763

Carroll, J. S. (2004). Redundancy as a design principle and an operating principle. *Risk Analysis : An Official Publication of the Society for Risk Analysis*, *24*(4), 955–7. doi:10.1111/j.0272-4332.2004.00498.x

Carter, L., & McGrath, B. (2013). We Know How To Improve Inspection Reliability - Why Don't We Do It? In *DGZfP Proceedings BB 116-CD: 4th European-American Workshop on Reliability of NDE, 24-26 June 2009, Berlin, Germany* (pp. 1–8). Retrieved from http://www.ndt.net/article/reliability2013/papers/lecture19.pdf

Carvalho, A. A., Rebello, J. M. a., Souza, M. P. V., Sagrilo, L. V. S., & Soares, S. D. (2008). Reliability of non-destructive test techniques in the inspection of pipelines used in the oil industry. *International Journal of Pressure Vessels and Piping*, *85*(11), 745–751. doi:10.1016/j.ijpvp.2008.05.001

Cassanelli, G., Mura, G., Fantini, F., Vanzi, M., & Plano, B. (2006). Failure Analysis-assisted FMEA. *Microelectronics and Reliability*, *46*, 1795–1799. doi:10.1016/j.microrel.2006.07.072

Chapuis, B., Jenson, F., Calmon, P., DiCrisci, G., Hamilton, J., & Pomié, L. (2014). Simulation supported POD curves for automated ultrasonic testing of pipeline girth welds. *Welding in the World*, *58*(4), 433–441. doi:10.1007/s40194-014-0125-z

Clarke, D. M. (2005). Human redundancy in complex, hazardous systems: A theoretical framework. *Safety Science*, *43*(9), 655–677. doi:10.1016/j.ssci.2005.05.003

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences* (Revised Ed.). New York, NY: Academic Press.

Conte, J. M., & Jacobs, R. R. (1997). Redundant Systems Influences on Performance. *Human Performance*, *10*(4), 361–380. doi:10.1207/s15327043hup1004_3

Cotte, N., Meyer, J., & Coughlin, J. F. (2001). Older and Younger Drivers' Reliance on Collision Warning Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *45*(4), 277–280. doi:10.1177/154193120104500402

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In *Proceedings of the AIAA 1st Intelligent Systems Technical Conference, 20-22 September 2004, Chichago, Illinois*. Retrieved from http://web.mit.edu/aeroastro/labs/halab/papers/CummingsAIAAbias.pdf

Dekker, S. W. (2002). *The field guide to human error investigations*. Aldershot, England: Ashgate.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*(3), 629–636. doi:10.1037/h0046408

Deutsch, V., Platte, M., Schuster, V., & Deutsch, W. A. K. (2006). *Die Verfahren der ZfP [engl. NDT methods]*. Wuppertal: Castell-Verlag.

Dhillon, B. (1999). *Design Reliability: Fundamentals and Applications*. Boca Raton, FL: CRC Press.

Dhillon, B. (2003). Methods for performing human reliability and error analysis in health care. *International Journal of Health Care Quality Assurance*, *16*(6), 306–317. doi:10.1108/09526860310495697

Dhillon, B. (2007). *Human reliability and error in transportation systems. eBook*. Springer Verlag. doi:10.1007/978-1-84628-812-8

Dickens, J., & Bray, D. (1994). Human performance considerations in nondestructive testing. *Materials Evaluation*, *52*(9), 1033–1041.

DIN EN 1330-4. (2010). Non-destructive testing - Terminology - Part 4: Terms used in ultrasonic testing; Trilingual version EN 1330-4:2010. Berlin: DIN Deutsches Institut für Normung e.V.

DIN EN ISO 9712. (2012). *Non-destructive testing – Qualification and certification of NDT personnel (ISO 9712:2012); English version EN ISO 9712:2012, English translation of DIN EN ISO 9712:2012-12*. DIN Deutsches Institut für Normung e.V., Berlin.

Dixon, S. R., & Wickens, C. D. (2006). Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(3), 474–486. doi:10.1518/001872006778606822

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the Independence of Compliance and Reliance: Are Automation False Alarms Worse Than Misses? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(4), 564–572. doi:10.1518/001872007X215656

Dzindolet, M. T., Beck, H. P., & Pierce, L. G. (2000). Encouraging Human Operators to Appropriately Rely on Automated Decision Aids. In *Proceedings of the 2000 Command and Control Research and Technology Symposium, Monterey, CA* (pp. 1–10). Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a462375.pdf

Dzindolet, M. T., Dawe, L. A., Beck, H. P., & Pierce, L. G. (2001). *A framework of automation use [Report No. ARL-TR-2412]*. Aberdeen Proving Ground, MD: Army Research Laboratory. Retrieved from http://www.arl.army.mil/arlreports/2001/ARL-TR-2412.pdf

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*, 697–718. doi:10.1016/S1071-5819(03)00038-7

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (1999). Misuse and Disuse of Automated Aids. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *43*(3), 339. doi:10.1177/154193129904300345

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. a. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors*, *44*(1), 79–94. doi:10.1518/0018720024494856

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting Misuse and Disuse of Combat Identification Systems. *Military Psychology*, *13*(3), 147–164. doi:10.1207/S15327876MP1303_2

Echeverria, D., Barnes, V., & Bittner, A. . (1991). The Impact of Environmental Exposures on Industrial Performance of Tasks. In W. Karwowski & J. W. Yates (Eds.), *Advances in Industrial Ergonomics and Safety III* (pp. 629–636). Philadelphia, PA: Taylor & Francis.

Enkvist, J. (2003). *A human factors perspective on non-destructive testing (NDT). Detection and identification of cracks [Doctoral dissertation]*. Stockholm University, Stockholm, Sweden.

Enkvist, J., Edland, A., & Svenson, O. (1999). *Human factors aspects of non-destructive testing in the nuclear power context. A review of research in the field [SKI report 99:8]*. Stockholm, Sweden: Swedish Nuclear Power Inspectorate (SKI). Retrieved from http://www.stralsakerhetsmyndigheten.se/Global/Publikationer/SKI_import/010806/13184331069/99-8.pdf

Enkvist, J., Edland, A., & Svenson, O. (2001a). Effects of operator time pressure and noise on manual ultrasonic testing. *Insight*, *43*(11), 725–730.

Enkvist, J., Edland, A., & Svenson, O. (2001b). *Effects of time pressure and noise on non-destructive testing [SKI report 01:48]*. Stockholm: Swedish Nuclear Power Inspectorate (SKI). Retrieved from http://www.stralsakerhetsmyndigheten.se/Global/Publikationer/SKI_import/020227/4a29504cde409b07dda1c235bc958dbc/01-48.pdf

Enkvist, J., Edland, A., & Svenson, O. (2001c). Operator performance in a blind test piece trial. *Materials Evaluation*, *59*(4), 531–536.

Erhard, A. (2013). Non-destructive Evaluation. In H. Czichos (Ed.), *Handbook of Technical Diagnostics* (pp. 161–174). Berlin: Springer Verlag.

Ezer, N., Fisk, A. D., & Rogers, W. A. (2008). Age-Related Differences in Reliance Behavior Attributable to Costs Within a Human-Decision Aid System. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(6), 853–863. doi:10.1518/001872008X375018

FAA. (2000). *System Safety Handbook*. Federal Aviation Administration (FAA). Retrieved from http://www.faa.gov/library/manuals/aviation/risk_management/ss_handbook/

FAA. (2009). *Risk management handbook*. Federal Aviation Administration (FAA). Retrieved from https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/risk_management_handbook/media/risk_management_handbook.pdf

Fahlbruch, B. (2009). Integrating Human Factors in Safety and Reliability Approaches. In *DGZfP Proceedings BB 116-CD: 4th European-American Workshop on Reliability of NDE, 24-26 June 2009, Berlin, Germany* (pp. 1–7). NDT.net. Retrieved from http://www.ndt.net/article/reliability2009/Inhalt/th4a1.pdf

Fahlbruch, B., & Wilpert, B. (1999). System safety - an emerging field for I/O psychology. In C. Cooper & I. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Volume 14., pp. 55–93). New York, NY: John Wiley & Sons Ltd.

Felsenthal, D., & Fuchs, E. (1976). Experimental Evaluation of Five Designs of Redundant Organizational Systems. *Administrative Science Quarterly*, *21*(3), 474–488. Retrieved from http://www.jstor.org/stable/2391855

Field, A. (2013). *Discovering statistics using SPSS* (4. Edition.). London, UK: SAGE Publications.

Forsyth, D., Komorowski, J., Gould, R., & Marincak, A. (1999). Automation of enhanced visual NDT techniques. In *Proceedings of the 1st Pan American Conference for Nondestructive Testing, 14-18 September 1998, Toronto, Canada* (Vol. 4). Retrieved from http://www.ndt.net/article/pacndt98/18/18.htm

Fücsök, F., & Müller, C. (2000). Human Factors: The NDE Reliability of Routine Radiographic Film Evaluation. In *Proceedings of the 15th WCNDT, 15-21 October 2000, Rome, Italy* (pp. 1–7). Retrieved from http://www.ndt.net/article/wcndt00/papers/idn740/idn740.htm

Fücsök, F., Müller, C., & Scharmach, M. (2002). Reliability of Routine Radiographic Film Evaluation–An Extended ROC Study of the Human Factor. In *Proceedings of the 8th European Conference on Non Destructive Testing, June 17-21 2002, Barcelona, Spain*. Retrieved from http://www.ndt.net/article/ecndt02/429/429.htm

Gaal, M., Bertovic, M., Zickler, S., Fahlbruch, B., Spokoiny, V., Schombach, D., … Cramer, H.-J. (2009). *Untersuchungen zum Einfluss menschlicher Faktoren auf das Ergebnis von zerstörungsfreien Prüfungen, Möglichkeiten zur Minimierung dieses Einflusses und Bewertung der Prüfergebnisse*. Salzgitter, Germany: Bundesamt für Strahlenschutz. Retrieved from http://doris.bfs.de/jspui/bitstream/urn:nbn:de:0221-2009111107/1/BfS_2009_BfS-RESFOR-25-09.pdf

Gandossi, L., & Annis, C. (2010). *ENIQ TGR technical document: Probability of Detection Curves: Statistical Best-Practices [ENIQ report No 41]*. Petten, The Netherlands: European Commission, Joint Research Centre, Institute for Energy. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Probability+of+Detection+Curves+:+Statistical+Best-Practices#0

George, J. M. (1992). Extrinsic and intrinsic origins of perceived social loafing in organizations. *Academy of Management Journal*. doi:10.2307/256478

Giesa, H., & Timpe, K.-P. (2002). Technisches Versagen und menschliche Zuverlässigkeit [Technical failure and human reliability]. In K.-P. Timpe, T. Jürgensohn, & H. Kolrep (Eds.), *Mensch-Maschine-Systemtechnik: Konzepte, Modellierung, Gestaltung, Evaluation [Human-Machine Systems Engineering: Concepts, modeling, design, evaluation]* (pp. 63–106). Düsseldorf: Symposion Publishing.

Haapanen, P., & Helminen, A. (2002). *Failure Mode and Effects Analysis of software-based automation systems [Report No. STUK-YTO-TR 190]*. Helsinki: Radiation and Nuclear Safety Authority (STUK). Retrieved from http://www.stuk.fi/julkaisut/tr/stuk-yto-tr190.pdf

Harkins, S. G. (1987). Social loafing and social facilitation. *Journal of Experimental Social Psychology*, *23*(1), 1–18. doi:10.1016/0022-1031(87)90022-9

Harkins, S. G., & Jackson, J. M. (1985). The Role of Evaluation in Eliminating Social Loafing. *Personality and Social Psychology Bulletin*, *11*(4), 457–465. doi:10.1177/0146167285114011

Harkins, S. G., Latané, B., & Williams, K. D. (1980). Social loafing: Allocating effort or taking it easy? *Journal of Experimental Social Psychology*, *16*(5), 457–465. doi:10.1016/0022-1031(80)90051-7

Harkins, S. G., & Petty, R. E. (1982). Effects of task difficulty and task uniqueness on social loafing. *Journal of Personality and Social Psychology*, *43*(6), 1214–1229. doi:10.1037/0022-3514.43.6.1214

Harkins, S. G., & Szymanski, K. (1988). Social loafing and self-evaluation with an objective standard. *Journal of Experimental Social Psychology*, *24*, 354–365. doi:10.1016/0022-1031(88)90025-X

Harris, D. (1988). *Human performance in NDE inspections and functional tests [EPRI report NP-6052]*. Santa Barbara, CA: Electric Power Research Institute (EPRI). Retrieved from http://www.epri.com/abstracts/Pages/ProductAbstract.aspx?ProductId=NP-6052

Harris, D. (1990). Effect of human information processing on the ultrasonic detection of intergranular stress-corrosion cracking. *Materials Evaluation*, *48*, 475–480.

Harris, D. (1992). *Effect of Decision Making on Ultrasonic Examination Performance [EPRI report TR-100412]*. Palo Alto, CA: Electric Power Research Institute (EPRI). Retrieved from http://www.epri.com/abstracts/Pages/ProductAbstract.aspx?ProductId=TR-100412

Harris, D., & Chaney, F. (1969). *Human factors in quality assurance*. New York, NY: John Wiley & Sons, Inc.

Harris, D., & McCloskey, B. (1990). *Cognitive correlates of ultrasonic inspection performance [EPRI report NP-6675]*. Palo Alto, CA: Electric Power Research Institute (EPRI). Retrieved from http://www.epri.com/abstracts/Pages/ProductAbstract.aspx?ProductId=NP-6675

Hatfield, J., & Fernandes, R. (2009). The role of risk-propensity in the risky driving of younger drivers. *Accident; Analysis and Prevention*, *41*(1), 25–35. doi:10.1016/j.aap.2008.08.023

Hellier, C. J. (2013). *Handbook of Nondestructive Evaluation.* (2nd Ed.). New York, NY: McGraw-Hill.

Herr, J., & Marsh, G. (1978). NDT reliability and human factors. *Materials Evaluation*, *36*(13), 41–46.

Hollnagel, E. (1993). The phenotype of erroneous actions. *International Journal of Man-Machine Studies*, *39*(1), 1–32. doi:10.1006/imms.1993.1051

Hollnagel, E. (2008a). Risk + barriers = safety? *Safety Science*, *46*(2), 221–229. doi:10.1016/j.ssci.2007.06.028

Hollnagel, E. (2008b). The changing nature of risks. *Ergonomics Australia Journal*, *22*(1), 33–46. Retrieved from http://hal-ensmp.archives-ouvertes.fr/docs/00/50/88/58/PDF/Changingnatureofrisks.pdf

Hollnagel, E., & Amalberti, R. (2001). THE EMPEROR ' S NEW CLOTHES Or Whatever Happened To "Human Error"? In *Invited keynote presentation at 4th International Workshop on Human Error, Safety and System Development, 11–12 June 2001, Linköping, Sweden* (pp. 1–18). Retrieved from http://www.scribd.com/doc/181044462/The-emperor-s-new-clothes-pdf#scribd

Holstein, R., Bertovic, M., Kanzler, D., & Müller, C. (2014). NDT Reliability in the Organizational Context of Service Inspection Companies. *Materials Testing*, *56*(7-8), 607–610. doi:10.3139/120.110601

HSE. (1999). *Reducing error and influencing behaviour. Human Factors and Ergonomics*. HSE Books. Retrieved from http://www.hse.gov.uk/pubns/books/hsg48.htm

Hudson, P. T. W., Reason, J. T., Bentley, P. D., & Primrose, M. (2013). Tripod Delta: Proactive Approach to Enhanced Safety. *Journal of Petroleum Technology*, *46*(1), 58–62. doi:10.2118/27846-PA

IAEA. (2011). *Viability of Sharing Facilities for the Disposal of Spent Fuel and Nuclear Waste [Report No. IAEA-TECDOC-1658]*. Vienna: International Atomic Energy Agency (IAEA). Retrieved from http://www-pub.iaea.org/MTCD/Publications/PDF/TE1658_web.pdf

IEC/ISO 31010. (2009). *Risk management - Risk assessment techniques*. Geneva, Switzerland: ISO/IEC.

Ingham, A., Levinger, G., Graves, J., & Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. *Journal of Experimental Social Psychology*, *10*, 371–384. Retrieved from http://www.sciencedirect.com/science/article/pii/002210317490033X

ISO 31000. (2009). *Risk management — Principles and guidelines*. Geneva, Switzerland: International Organization for Standardization (ISO).

ISO Guide 73. (2009). *Risk management — Vocabulary*. Geneva, Switzerland: International Organization for Standardization (ISO).

JRC-IE. (2010). *ENIQ recommended practice 10: Personnel Qualification [ENIQ report No 38]*. Petten, The Netherlands: European Commission, Joint Research Centre, Institute for Energy.

Judge, T., & Chandler, T. (1996). Individual-level determinants of employee shirking. *Relations Industrielles/Industrial Relations*, *51*(3), 468–487. Retrieved from http://www.timothy-judge.com/Judge & Chandler.pdf

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*(4), 681–706. doi:10.1037/0022-3514.65.4.681

Karau, S. J., & Williams, K. D. (1997). The effects of group cohesiveness on social loafing and social compensation. *Group Dynamics: Theory, Research, and Practice*, *1*(2), 156–168. doi:10.1037/1089-2699.1.2.156

Kettunen, J., Reiman, T., & Wahlström, B. (2007). Safety management challenges and tensions in the European nuclear power industry. *Scandinavian Journal of Management*, *23*(4), 424–444. doi:10.1016/j.scaman.2007.04.001

KTA 3201.4. (2010). Safety Standards of the Nuclear Safety Standards Commission (KTA) - KTA 3201.4 (2010-11): Components of the Reactor Coolant Pressure Boundary of Light Water Reactors - Part 4: In-service Inspections and Operational Monitoring. Salzgitter, Germany: Bundesamt für Strahlenschutz (BfS). Retrieved from http://www.kta-gs.de/e/standards/3200/3201_4_engl_2010_11.pdf

KTA 3221.4. (2013). Safety Standards of the Nuclear Safety Standards Commission (KTA) - KTA 3221.4 (2013-11) Pressure and Activity Retaining Components of Systems Outside the Primary Circuit; Part 4: Inservice Inspections and Operational Monitoring. Salzgitter, Germany: Bundesamt für Strahlenschutz (BfS). Retrieved from http://www.kta-gs.de/e/standards/3200/3211_4_engl_2013_11.pdf

Landau, M. (1969). Rationality, Redundancy, and the Problem of Duplication and Overlap. *Public Administration Review*, *29*(4), 346–358. Retrieved from http://www.jstor.org/stable/973247

LaPorte, T., & Cansolini, P. (1991). Working in Practice But Not in Theory: Theoretical Challenges of "High-Reliability Organizations." *Journal of Public Administration and Theory*, *1*, 19–47.

Latané, B., Williams, K. D., & Harkins, S. G. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, *37*(6), 822–832. doi:10.1037/0022-3514.37.6.822

Latorella, K. A., & Prabhu, P. V. (2000). A review of human error in aviation maintenance and inspection. *International Journal of Industrial Ergonomics*, *26*(2), 133–161. doi:10.1016/S0169-8141(99)00063-3

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–70. doi:10.1080/00140139208967392

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. doi:10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. doi:10.1518/hfes.46.1.50_30392

Lerner, A. W. (1986). There is more than One Way to be Redundant: A Comparison of Alternatives for the Design and Use of Redundancy in Organizations. *Administration & Society*, *18*(3), 334–359. doi:10.1177/009539978601800303

Leveson, N. G. (2011). *Engineering a safer world: Systems thinking applied to safety. eBook*. Cambridge, MA: MIT Press. Retrieved from https://mitpress.mit.edu/sites/default/files/titles/free_download/9780262016629_Engineering_a_Safer_World.pdf

Liao, T. W., & Li, Y. (1998). An automated radiographic NDT system for weld inspection: Part II—Flaw detection. *NDT & E International*, *31*(3), 183–192. doi:10.1016/S0963-8695(97)00042-X

Lingvall, F., & Stepinski, T. (2000). Automatic detecting and classifying defects during eddy current inspection of riveted lap-joints. *NDT & E International*, *33*(1), 47–55. doi:10.1016/S0963-8695(99)00007-9

Madhavan, P., & Wiegmann, D. A. (2004). A New Look at the Dynamics of Human-Automation Trust: Is Trust in Humans Comparable to Trust in Machines? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(3), 581–585. doi:10.1177/154193120404800365

Madhavan, P., & Wiegmann, D. A. (2005). Cognitive Anchoring on Self-Generated Decisions Reduces Operator Reliance on Automated Diagnostic Aids. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *47*(2), 332–341. doi:10.1518/0018720054679489

Madhavan, P., & Wiegmann, D. A. (2007). Effects of Information Source, Pedigree, and Reliability on Operator Interaction With Decision Support Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(5), 773–785. doi:10.1518/001872007X230154

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2004). Occasional automation failures on easy tasks undermines trust in automation. In *Proceedings of the 112th Annual Meeting of the American Psychological Association* (pp. 1–6). Retrieved from https://www.researchgate.net/publication/242179092_OCCASIONAL_AUTOMATION_FAILURES_ON_EASY_TASKS_UNDERMINES_TRUST_IN_AUTOMATION

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(2), 241–256. doi:10.1518/001872006777724408

Manzey, D. (2012). Systemgestaltung und Automatisierung [System design and automation]. In P. Badke-Schaub, G. Hofinger, & K. Lauche (Eds.), *Human Factors. Psychologie sicheren Handelns in Risikobranchen. 2. Auflage* (pp. 333–352). Berlin Heidelberg: Springer.

Manzey, D., Boehme, K., & Schöbel, M. (2013). Human Redundancy as Safety Measure in Automation Monitoring. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 369–373. doi:10.1177/1541931213571080

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87. doi:10.1177/1555343411433844

Marold, J. (2011). *Sehen vier Augen mehr als zwei? Der Einfluss personaler Redundanz auf die Leistung bei der Überwachung automatisierter Systeme*. (Doctoral dissertation. Technische Universität Berlin). Retrieved from https://opus4.kobv.de/opus4.../marold_juliane.pdf

McGonnagle, W. J. (1975). *Nondestructive testing* (2nd Ed.). London, UK: McGraw-Hill.

McGrath, B. (1999). *Programme for the Assessment of NDT in Industry*. HSE report on CD-ROM.

McGrath, B. (2008). *Programme for the Assessment of NDT in Industry, PANI 3 [Report No. RR617]*. Health and Safety Executive. Retrieved from http://www.hse.gov.uk/research/rrpdf/rr617.pdf

McGrath, B., Wheeler, J., & Bainbridge, H. (2009). PANI and the Role of the Written NDT Procedure. In *DGZfP Proceedings BB 116-CD: 4th European-American Workshop on Reliability of NDE, 24-26 June 2009, Berlin, Germany* (pp. 1–7). Retrieved from http://www.ndt.net/article/reliability2009/Inhalt/th5a1.pdf

McGrath, B., Worrall, G., & Udell, C. (2004). *Programme for the Assessment of NDT in Industry, PANI 2*. HSE report on CD-ROM.

Metzger, U., & Parasuraman, R. (2005). Automation in Future Air Traffic Management: Effects of Decision Aid Reliability on Controller Performance and Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *47*(1), 35–49. doi:10.1518/0018720053653802

Meyer, J. (2001). Effects of Warning Validity and Proximity on Responses to Warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 563–572. doi:10.1518/001872001775870395

Meyer, J. (2004). Conceptual Issues in the Study of Dynamic Hazard Warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(2), 196–204. doi:10.1518/hfes.46.2.196.37335

MIL-STD 1629A. (1980). Procedures for performing a Failure Mode, Effects and Criticality Analysis. Military standard. Washington, DC: U.S. Department of Defense. Retrieved from http://sre.org/pubs/Mil-Std-1629A.pdf

Moray, N. (2003). Monitoring, complacency, scepticism and eutactic behaviour. *International Journal of Industrial Ergonomics*, *31*(3), 175–178. doi:10.1016/S0169-8141(02)00194-4

Moray, N., & Huey, B. (Eds.). (1988). *Human factors research and nuclear safety*. Washington, D.C.: National Academy Press.

Mosier, K. L., & Skitka, L. J. (1996). Human Decision Makers and Automated Decision Aids: Made for Each Other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications* (pp. 201–220). Mahwah, New Jersey: Lawrence Erlbaum Associates Ltd., Publishers.

Mosier, K. L., Skitka, L. J., Burdick, M. D., & Heers, S. T. (1996). Automation Bias, Accountability, and Verification Behaviors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *40*(4), 204–208. doi:10.1177/154193129604000413

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation Bias: Decision Making and Performance in High-Tech Cockpits. *The International Journal of Aviation Psychology*, *8*(1), 47–63. doi:10.1207/s15327108ijap0801_3

Müller, C., Bertovic, M., Pavlovic, M., Kanzler, D., Ewert, U., Pitkänen, J., & Ronneteg, U. (2013). Paradigm Shift in the Holistic Evaluation of the Reliability of NDE Systems. *Materials Testing*, *55*(4), 261–269. doi:10.3139/120.110433

Müller, C., Pavlovic, M., Takahashi, K., Ewert, U., Rosenthal, M., Brekow, G., … Pitkänen, J. (2007). POD Evaluation of NDE Techniques for Canister-Components for Risk Assessment of Nuclear Waste Encapsulation. In *Proceedings of the 6th International Conference on NDE in Relation to Structural integrity for Nuclear and Pressurized Components, October 12-14 2007, Budapest, Hungary* (pp. 1–12). Retrieved from http://www.ndt.net/article/jrc-nde2007/papers/08_11-22.pdf

Murgatroyd, R. (1992). Assuring human reliability for effective inspection. In W. Gardner (Ed.), *Improving the effectiveness and reliability of non-destructive testing*. Oxford: Pergamon Press Ltd.

Murgatroyd, R., Chapman, R., Crutzen, S., Seed, H., Willets, A., & Worrall, G. (1994). *Human Reliability in Inspection, Final Report on Action 7 in the PISC III Programme [PISC III Report n° 31]. Nuclear Energy Agency*. JRC & NEA.

National Transportation Safety Board. (1989). *Aircraft accident report - United Airlines Flight 232*. Washington, D.C.: National Transportation Safety Board.

Nichols, R., & Crutzen, S. (Eds.). (1988). *Ultrasonic inspection of heavy steel components: the PISC II final report*. Essex: Elsevier Applied Science Publishers Ltd.

Nockemann, C., & Fortunko, C. (1997). Summary of the workshop. In *Proceedings of the European-American Workshop: Determination of Reliability and Validation Methods on NDE, 18-20 June 1997, Berlin, Germany* (pp. 11–17). Berlin: DGZfP.

Nockemann, C., Heidt, H., & Thomsen, N. (1991). Reliability in NDT: ROC study of radiographic weld inspections. *NDT & E International*, *24*(5), 235–245. doi:10.1016/0963-8695(91)90372-A

Norros, L. (1998). Human and organisational factors in the reliability of non-destructive testing (NDT). In J. Solinm, M. Sarkimo, M. Asikainen, & A. Avall (Eds.), *RATU2: The Finnish Research Programme on the Structural Integrity of Nuclear Power Plants. Synthesis of achievements 1995 – 1998* (pp. 271–280). Espoo: VTT Technical Research Centre of Finland. Retrieved from http://www.vtt.fi/inf/pdf/symposiums/1998/S190.pdf

Norros, L., & Kettunen, J. (1998). *Analysis of NDT-inspectors working practices [Report No. STUK-YTO-TR 147][Abstract]*. Helsinki: STUK (Finnish Centre for Radiation and Nuclear Safety). Retrieved from https://www.etde.org/etdeweb/servlets/purl/294845-eRPdZv/webviewable/294845.pdf

Oakley, B., Mouloua, M., & Hancock, P. (2003). Effects of Automation Reliability on Human Monitoring Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*(1), 188–190. doi:10.1177/154193120304700139

Parasuraman, R., & Manzey, D. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 381–410. doi:10.1177/0018720810376055

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *The International Journal of Aviation Psychology*, *3*(1), 1–23. doi:http://dx.doi.org/10.1207/s15327108ijap0301_1

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. doi:10.1518/001872097778543886

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still Vital After All These Years of Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 511–520. doi:10.1518/001872008X312198

Pitkänen, J. (2013). *Inspection of Disposal Canisters Components [Report POSIVA 2012-35]*. Eurajoki, Finland: Posiva Oy. Retrieved from http://www.posiva.fi/files/3514/POSIVA_2012-35.2.pdf

Pitkänen, J., Bertovic, M., Brackrock, D., Brekow, G., Ewert, U., Kanzler, D., & Müller, C. (2014). Reliable Evaluation of Acceptability of Weld for Final Disposal Based on the Canister Copper Weld Inspection Using Different NDT Methods. *Materials Testing*, *56*(9), 748–757. doi:10.3139/120.110609

Pitkänen, J., Lipponen, A., Lahdenpera, K., & Kiselmann, I. (2009). The Eddy Current Inspection for Detection of Surface and Near Surface Defects in Copper Components and an Electron Beam Weld. In *Proceedings of the JRC-NDE 2009*. Retrieved from http://www.ndt.net/article/jrc-nde2009/papers/147.pdf

Pitkänen, J., Salonen, T., Bertovic, M., Müller, C., & Pavlovic, M. (2011). NDT Reliability in Risk Minimization during Manufacturing and Welding of Spent Nuclear Fuel Disposal Components - A Realistic Tool for Reliable Inspections. In *DGZfP Proceedings BB 116-CD: 4th European-American Workshop on Reliability of NDE, 24-26 June 2009, Berlin, Germany* (pp. 1–21). DGZfP. Retrieved from http://www.ndt.net/article/reliability2009/Inhalt/fr1a2.pdf

Pond, D., Donohoo, D., & Harris, Jr, R. (1998). *An evaluation of human factors research for ultrasonic inservice inspection [NUREG/CR-6605]*. Washington, DC: U.S. Nuclear Regulatory Commission. Retrieved from http://www.osti.gov/energycitations/product.biblio.jsp?osti_id=582236

Posiva Oy. (2010). Safe Final Disposal of Spent Nuclear Fuel. Olkiluoto, Eurajoki: Eura Print Oy.

Posiva Oy. (2015). *Annual report 2014*. Eurajoki, Finland: Posiva Oy.

Povyakalo, A. A., Alberdi, E., Strigini, L., & Ayton, P. (2013). How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. *Medical Decision Making : An International Journal of the Society for Medical Decision Making*, *33*(1), 98–107. doi:10.1177/0272989X12465490

Price, K. H. (1987). Decision Responsibility, Task Responsibility, and Social Loafing. *Organizational Behavior and Human Decision Processes*, 330–345.

Raj, B., & Venkatraman, B. (2013). Overview of Diagnostics and Monitoring Methods and Techniques. In H. Czichos (Ed.), *Handbook of Technical Diagnostics* (pp. 43–68). Berlin: Springer Verlag.

Rasband, W. (2010). ImageJ (Version 1.43u) [Computer software]. National Institutes of Health, USA. Retrieved from http://imagej.nih.gov/ij/index.html

Rasmussen, J. (1980). What can be learned from human error reports? In K. Duncan, M. Gruneberg, & D. Wallis (Eds.), *Changes in Working Life*. London: Wiley.

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. *Safety Science*, *27*(2-3), 183–213. doi:10.1016/S0925-7535(97)00052-0

Reason, J. (1990). *Human error*. New York: Cambridge University Press.

Reason, J. (1993). Managing the management risk: New approaches to organisational safety. In B. Wilpert & T. Qvale (Eds.), *Reliability and Safety in Hazardous Work Systems: Approaches to Analysis and Design* (pp. 7–22). Hove: Lawrence Erlbaum Associates Ltd., Publishers.

Reason, J. (1995). A systems approach to organizational error. *Ergonomics*, *38*(8), 1708–1721. doi:10.1080/00140139508925221

Reason, J. (1997). *Managing the Risks of Organizational Accidents*. Farnham, Surrey: Ashgate.

Reason, J. (2000). Human error: models and management. *BMJ*, *320*, 768–770. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1117770/

Reason, J., & Hobbs, A. (2003). *Managing maintenance error: a practical guide*. Aldershot, England: Ashgate.

Reason, J., Shotton, R., Wagenaar, W., Hudson, P. T. W., & Groeneweg, J. (1989). *TRIPOD, A Principled Basis for Safer Operations*. The Hague: Shell Internationale Petroleum Maatschappij.

Reichenbach, J., Onnasch, L., & Manzey, D. (2011). Human Performance Consequences of Automated Decision Aids in States of Sleep Loss. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(6), 717–728. doi:10.1177/0018720811418222

Rhee, S., & Ishii, K. (2003). Using cost based FMEA to enhance reliability and serviceability. *Advanced Engineering Informatics*, *17*(3-4), 179–188. doi:10.1016/j.aei.2004.07.002

Rice, S., & Keller, D. (2009). Automation reliance under time pressure. *Cognitive Technology*, *14*(1), 36–44. doi:10.1518/001872007X215656

Riley, V. (1996). Operator Reliance on Automation: Theory and Data. In R. Parasuraman & M. Mouloua (Eds.), *Automation and Human Performance: Thoery and Applications* (pp. 19–35). Mahwah, New Jersey: Lawrence Erlbaum Associates Ltd., Publishers.

Rohrmann, B. (2005). *Risk Attitude Scales: Concepts, Questionnaires, Utilizations*. Melbourne: University of Melbourne. Retrieved from http://www.rohrmannresearch.net/pdfs/rohrmann-ras-report.pdf

Rosado, L. S., Santos, T. G., Piedade, M., Ramos, P. M., & Vilaça, P. (2010). Advanced technique for non-destructive testing of friction stir welding of metals. *Measurement*, *43*(8), 1021–1030. doi:10.1016/j.measurement.2010.02.006

Rosenthal, R. (1991). *Meta-analytic procedures for social research (rev. ed.). Applied social research methods series, Vol. 6*. Thousand Oakes, CA: SAGE Publications.

Röttger, S., Bali, K., & Manzey, D. (2009). Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. *Ergonomics*, *52*(5), 512–23. doi:10.1080/00140130802379129

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(1), 76–87. doi:10.1518/001872007779598082

Sagan, S. D. (2004). The problem of redundancy problem: why more nuclear security forces may produce less nuclear security. *Risk Analysis : An Official Publication of the Society for Risk Analysis*, *24*(4), 935–46. doi:10.1111/j.0272-4332.2004.00495.x

Sambath, S., Nagaraj, P., & Selvakumar, N. (2010). Automatic Defect Classification in Ultrasonic NDT Using Artificial Intelligence. *Journal of Nondestructive Evaluation*, *30*(1), 20–28. doi:10.1007/s10921-010-0086-0

Santos, J., & Perdigão, F. (2001). Automatic defects classification — a contribution. *NDT & E International*, *34*(5), 313–318. doi:10.1016/S0963-8695(00)00043-8

Sarter, N. B., Woods, D. D., & Billings, D. R. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of Human Factors & Ergonomics* (2. Edition., pp. 1926–1943). New York, NY: Wiley.

Sasou, K., & Reason, J. (1999). Team errors: definition and taxonomy. *Reliability Engineering & System Safety*, *65*(1), 1–9. doi:10.1016/S0951-8320(98)00074-X

Schmitz, H.-P., & Mißmann, J. (2009). *Dictionary of Non-Destructive Testing*. Essen: Vulkan-Verlag GmbH.

Schöbel, M., & Manzey, D. (2011). Subjective theories of organizing and learning from events. *Safety Science*, *49*(1), 47–54. doi:10.1016/j.ssci.2010.03.004

Shafeek, H. I., Gadelmawla, E. S., Abdel-Shafy, A. A., & Elewa, I. M. (2004). Automatic inspection of gas pipeline welding defects using an expert vision system. *NDT & E International*, *37*(4), 301–307. doi:10.1016/j.ndteint.2003.10.004

Sheridan, T. B. (2008). Risk, Human Error, and System Resilience: Fundamental Ideas. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 418–426. doi:10.1518/001872008X250773

Sheridan, T. B., & Parasuraman, R. (2005). Human-Automation Interaction. *Reviews of Human Factors and Ergonomics*, *1*(89), 89–129. doi:10.1518/155723405783703082

Singh, A., Tiwari, T., & Singh, I. L. (2009). Effects of automation reliability and training on automation-induced complacency and perceived mental workload. *Journal of the Indian Academy of Applied Psychology*, *35*(Special Issue), 9–22. Retrieved from http://medind.nic.in/jak/t09/s1/jakt09s1p9.pdf

Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation- Induced "Complacency": Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology*, *3*(2), 111–122. doi:10.1207/s15327108ijap0302_2

SKB. (2008). *Encapsulation. When, where, how and why?* Stockholm: Svensk Kärnbränslehantering AB.

SKB. (2013). *RD&D Programme 2013. Programme for research, development and demonstration of methods for the management and disposal of nuclear waste [SKB Report TR-13-18]*. Stockholm: Svensk Kärnbränslehantering AB.

Skitka, L. J., Mosier, K. L., & Burdick, M. D. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991–1006. doi:10.1006/ijhc.1999.0252

Skitka, L. J., Mosier, K. L., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, *52*(4), 701–717. doi:10.1006/ijhc.1999.0349

Skitka, L. J., Mosier, K. L., Burdick, M., & Rosenblatt, B. (2000). Automation bias and errors: are crews better than individuals? *The International Journal of Aviation Psychology*, *10*(1), 85–97. doi:10.1207/S15327108IJAP1001_5

Spanner, J. (1999). *Swedish Human Factors Study of NDE [Report TP-114671]*. Charlotte, NC: Electric Power Research Institute (EPRI). Retrieved from http://www.epri.com/abstracts/Pages/ProductAbstract.aspx?ProductId=TP-114671

Spanner, J., & Harris, D. (1999). Human Factor Developments in Computer Based Training and Personnel Qualification. In *Proceedings of the First International Conference on NDE in Relation to Structural Integrity for Nuclear and Pressurised Components, 20 - 22 October 1998, Amsterdam, The Netherlands.* (pp. 149–165). Abington, Cambrige: Woodhead Publishing Limited.

Spanner Sr., J. C. (1986). Human Reliability Impact on In-Service Inspection. In D. Stahl (Ed.), *Proceedings of the 8th International Conference on NDE in the Nuclear Industry, 17-20 November 1986, Kissimmee, Florida, USA* (pp. 89–95). Orlando, FL: American Society for Metals.

Sun, Y., Bai, P., Sun, H., & Zhou, P. (2005). Real-time automatic detection of weld defects in steel pipe. *NDT & E International*, *38*(7), 522–528. doi:10.1016/j.ndteint.2005.01.011

Swain, A., & Guttmann, H. (1983). *Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report [No. NUREG/CR-1278; SAND-80-0200]*. Washington, D.C.: U.S. Nuclear Regulatory Commission.

Swets, J. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: collected papers.* Mahwah, NJ: Lawrence Erlbaum Associates.

Szymanski, K., & Harkins, S. G. (1987). Social loafing and self-evaluation with a social standard. *Journal of Personality and Social Psychology*, *53*(5), 891–897. doi:10.1037/0022-3514.53.5.891

Taylor, T., & Nockemann, C. (1999). Summary of American-European Workshop on NDE Reliability. In *Paper summaries book - American-European Workshop on Nondestructive Inspection Reliability, 21-24 September 1999, Boulder* (pp. 7–8). Columbus, OH: The American Society for Nondestructive Testing, Inc.

Taylor, T., Spanner, Sr., J. C., Heasler, P., Doctor, R., & Deffenbaugh, J. D. (1989). An Evaluation of Human Reliability in Ultrasonic ln-Service Inspection for Intergranular Stress-Corrosion Cracks through Round-Robin testing. *Materials Evaluation*, *47*(3), 338–344.

Timpe, K.-P. (1993). Psychology's Contributions to the Improvement of Safety and Reliability in the Man-Machine System. In B. Wilpert & T. Qvale (Eds.), *Reliability and Safety in Hazardous Work Systems: Approaches to Analysis and Design* (pp. 119–132). Hove, UK: Lawrence Erlbaum Associates Ltd., Publishers.

Trampus, P. (2013). NDT challenges and responses - an overview. In *Proceedings of The 12th International Conference of the Slovenian Society for Non-Destructive Testing »Application of Contemporary Non-Destructive Testing in Engineering«, 4-6 September 2013, Portorož, Slovenia* (pp. 1–11). Retrieved from http://www.ndt.net/article/ndt-slovenia2013/papers/1.pdf

U.S. NRC. (2012). *North Anna Power Station - NRC integrated inspection report 05000338/2012003, and 05000339/2012003*. U.S. Nuclear Regulatory Commission. Retrieved from http://pbadupws.nrc.gov/docs/ML1221/ML12213A637.pdf

van Leeuwen, J. F., Nauta, M. J., de Kaste, D., Odekerken-Rombouts, Y. M. C. F., Oldenhof, M. T., Vredenbregt, M. J., & Barends, D. M. (2009). Risk analysis by FMEA as an element of analytical validation. *Journal of Pharmaceutical and Biomedical Analysis*, *50*(5), 1085–7. doi:10.1016/j.jpba.2009.06.049

Wall, M. (2013). Evaluating POD in Real Situations and the " Delta " Factor. In *DGZfP Proceedings BB 144-CD: 5th European-American Workshop of NDE", 7-10 October 2013, Berlin, Germany* (pp. 1–16). Retrieved from http://www.ndt.net/article/reliability2013/papers/lecture33.pdf

Wall, M., Burch, S., & Lilley, J. (2009). Human factors in POD modelling and use of trial data. *Insight - Non-Destructive Testing and Condition Monitoring*, *51*(10), 553–561. doi:10.1784/insi.2009.51.10.553

Wang, G., & Liao, T. W. (2002). Automatic identification of different types of welding defects in radiographic images. *NDT & E International*, *35*(8), 519–528. doi:10.1016/S0963-8695(02)00025-7

Wassink, C. H. P. (2012). *Innovation in Non Destructive Testing*. (Doctoral dissertation. Technische Universität Delft). Retrieved from http://repository.tudelft.nl/view/ir/uuid:cd0e041c-7e1f-4c28-bc34-b81f0e00e55b/

Wetterneck, T. B., Skibinski, K., Schroeder, M., Roberts, T. L., & Carayon, P. (2004). Challenges with the Performance of Failure Mode and Effects Analysis in Healthcare Organizations: An IV Medication Administration HFMEATM. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(15), 1708–1712. doi:10.1177/154193120404801517

Wheeler, W., Rankin, W., Spanner, J., Budalmente, R., & Taylor, T. (1986). *Human factors study conducted in conjunction with a mini-round robin assessment of ultrasonic technician performance [Report NUREG/CR-4600]*. Richland, WA: U.S. Nuclear Regulatory Commission.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. doi:10.1080/14639220500370105

Wickens, C. D., Lee, J., Liu, Y. D., & Gordon Becker, S. E. (2004). *An introduction to human factors engineering* (2. ed.). New Jersey: Pearson Prentice Hall.

Williams, K. D., Harkins, S. G., & Latané, B. (1981). Identifiability as a deterrant to social loafing: Two cheering experiments. *Journal of Personality and Social Psychology*, *40*(2), 303–311. doi:10.1037/0022-3514.40.2.303

Williams, K. D., & Karau, S. J. (1991). Social loafing and social compensation: The effects of expectations of co-worker performance. *Journal of Personality and Social Psychology*, *61*(4), 570–581. doi:10.1037/0022-3514.61.4.570

Wilpert, B., & Fahlbruch, B. (1998). Safety related interventions in interorganisational fields. In A. Hale & M. Baram (Eds.), *Safety Management: The challenge of change* (pp. 235–247). Oxford, UK: Elsevier Science Ltd.

Wilpert, B., & Miller, R. (1999). Introduction. In J. Misumi, B. Wilpert, & R. Miller (Eds.), *Nuclear Safety: A Human Factors Perspective* (pp. xxi–xxiv). London, UK: Taylor & Francis Ltd.

Woods, D. D., Dekker, S. W., Cook, R., Johannesen, L., & Sarter, N. (2013). *Behind Human Error. Second edition*. Farnham, Surrey: Ashgate.

Zeller, B., Küfner, C., & Neumann, F. (2012). *Untersuchung zu neuen und modernisierten Berufsprofilen und einem Berufsgruppenprinzip für prüftechnische Berufe. [Abschlussbericht; unpublished]*. Nürnberg: Forschungsinstitut Betriebliche Bildung (f-bb) GmbH.

Glossary

A-scan presentation	Display of the ultrasonic signal in which the X-axis represents the time and the y-axis the amplitude
Acceptance level	Prescribed limits, below which a component is accepted
Amplitude (UT)	Absolute or relative measure of a sound wave's magnitude
Automated UT	A method by which an object is tested ultrasonically and the results are analysed without human intervention
Automatic scanning	Automatic displacement of the probe
B-scan presentation	Image of the results of UT showing a cross section of the test object perpendicular to the scanning surface and parallel to the reference direction
C-scan presentation	Image of the results of UT showing a cross section of the test object parallel to the scanning surface
Calibration	A process of establishing the sensitivity of the measurement system
Characterisation	Classifying the size and shape of an indication so it may be identified
Couplant (UT)	A medium interposed between the probe and the object under examination to enable the passage of ultrasonic waves between them
Critical defect	Discontinuity in the material large enough to cause concern of structural failure
Decision (or sizing) threshold	A threshold above which all pixels in the direct contact with the pixel exceeding the reporting level are judged as belonging to the indication
D-scan presentation	Image of the results of UT showing a cross section of the test object perpendicular to the scanning surface and perpendicular to the projection of the beam axis on the scanning surface (D-scan is typically perpendicular to B-scan)
Defect	A component discontinuity that has shape, size, orientation, or location, such that it is detrimental to the useful service of the part
Detection	Establishment of the presence of a discontinuity
Discontinuity	Detectable change in the material (also known as inhomogeneity)

Echo (UT)	Ultrasonic pulse reflected to the probe
Evaluation	Assessment of indications revealed by NDT against a predefined level
Failure (FMEA)	The failure of an item, which would result in failure of the system and is not compensated for by redundancy or alternative operational procedure
Failure cause (FMEA)	The physical or chemical processes, design defects, quality defects, part misapplication, or other processes which are the basic reason for failure or which initiate the physical process by which deterioration proceeds to failure
Failure mode (FMEA)	The manner by which a failure is observed (Generally, it describes the way the failure occurs and its impact on equipment operation)
Failure effect (FMEA)	The consequence a failure mode has on the operation, function, or status of an item
Indication	Representation or a signal from a discontinuity in the format typical for the method used
Geometrical indication	A non-relevant indication of a signal arising from an interaction of the energy sent through the material (e.g. ultrasonic beam) and the component geometry (e.g. edges).
Kurtosis	Pointyness of the distribution
Localization	Determining the location of an indication in the component
Manual scanning	Manual displacement of the probe
NDT instruction	Written description of the precise steps to be followed in testing to an established standard, code, specification or NDT procedure
NDT method	Discipline applying a physical principle in non-destructive testing, e.g., ultrasonic testing
NDT procedure	Written description of all essential parameters and precautions to be applied when non-destructively testing products in accordance with standard(s), code(s), or specification(s)
NDT technique	Special way of utilising an NDT method
NDT training	Process of instruction in theory and practice in the NDT method in which certification is sought, taking the form of training courses to a syllabus approved by the certification body
NDT reliability	The degree that an NDT system is capable of achieving its purpose regarding detection, characterisation, and false calls
Noise (signal)	Randomly distributed signals in the screen image, due to reflections from the structure of the material or the equipment
Probe (UT)	Electro-acoustical device, usually incorporating one or more transducers intended for transmission and/or reception of the ultrasonic waves
Qualification	Demonstration of physical attributes, knowledge, skill, training, and experience required to perform NDT tasks properly
Reporting	Amplitude of the echo above (or below) which every echo is reported

threshold/level (UT)	or recorded
Sensitivity (UT)	A measure of the smallest ultrasonic signal, which will produce a discernible indication on the display of an ultrasonic system
Signal-to-noise ratio, SNR (UT)	Ratio of the amplitude of the signal arising from a discontinuity in a material to the amplitude of the average background noise
Sizing	Determination of the dimensions of discontinuities or indications for evaluation
Skewness	Lack of symmetry of the distribution
Winzorising	A procedure of exclusion of outliers by replacing the outliers with the last value that is not an outlier

Note: The definitions were quoted from: Ali, Balint, Temple, & Leevers, 2012; DIN EN 1330-4, 2010; DIN EN ISO 9712, 2012; Field, 2013; Hellier, 2013; ISO 31000, 2009; MIL-STD 1629A, 1980; Nockemann & Fortunko, 1997; and Schmitz & Mißmann, 2009.

Abbreviations

ET	Eddy Current Testing
FMEA	Failure Modes and Effects Analysis
FMECA	Failure Modes and Effects and Criticality Analysis
NDT	Non-Destructive Testing
NDE	Non-Destructive Examination, Non-Destructive Evaluation
NDI	Non-Destructive Inspection
POD	Probability of Detection
RPN	Risk Priority Number
RT	Radiographic Testing
rVT	Remote Visual Testing
UT	Ultrasonic Testing

Statistical abbreviations

ANOVA	Analysis of Variance; a statistical procedure
d	Cohen's *d*; a measure of the effect size
df	Degrees of freedom
F	F-ratio; a test statistics used in ANOVA
M	Mean; a measure of central tendency
Mdn	Median, a measure of central tendency
n	The size of the population
N	The size of the sample or a group
p	Probability; also the statistical significance level
R	Pearson's correlation coefficient; also a measure of effect size for non-parametric statistical procedures
SD	Standard deviation; a measure of data dispersion
t	Test statistics for a Student's *t*-test
U	Test statistics for a Mann-Whitney *U* test
z	Data point expressed in standard deviation units
χ^2	Chi-square; refers to the test statistic, as well as to a distribution of data

Acknowledgements

I would like to use this opportunity to thank all those who have supported me throughout working on this dissertation. First, I would like to thank Prof. Dietrich Manzey for his openness and kindness, for the empowerment and support, and for guiding me scientifically and morally through this work. I extend this gratitude to the second advisor, Dr. Gerd-Rüdiger Jaenisch, for his most insightful comments, valuable discussions, and his constructive guidance. I would also like to remember my first advisor, the deceased Prof. Bernhard Wilpert, who introduced me to the field of human factors with such style, friendliness, and charm.

My highest gratitude goes to Dr. Christina Müller, who set me on this path, guided, and supported me in every possible way—morally, spiritually, and professionally. I thank her for the strength she gave me, for the ways she has shown me, for her patience and understanding, and for giving me the opportunity to do this work. My further gratitude goes to Dr. Babette Fahlbruch, for her scientific guidance, relentless support, and continuous motivation—as a mentor and as a friend.

This work was carried out during my employment at the Federal Institute for Materials Research and Testing (BAM), which funded my work and provided the resources necessary for its completion. The work presented in this dissertation is a result of a yearlong scientific cooperation between BAM, Swedish Nuclear Fuel and Waste Management Co (SKB), and the Finnish Posiva Oy. I feel privileged for being able to work on these projects and learn so much from my project partners. Their vision, interest, and openness to new challenges laid the foundations upon which this work was built. On that note, I would like to thank Ulf Ronneteg from SKB not only for his professional guidance, but moreover for his friendship and unlimited support. My gratitude extends to the recently deceased Jorma Pitkänen from Posiva, whose too early departure has left us saddened and whose absence is felt in more than one way. I would like to use this opportunity to thank him for his wondering mind and for all the valuable lessons on how NDT is carried out.

Furthermore, I would like to thank my colleagues from BAM for their support, time, for their help when needed, for the preparation of the studies, and for the participation in the studies. Foremost, I thank Daniel Kanzler for the help in the preparation and execution of the studies, for his consult, and for the continuous moral support. Furthermore, I thank Martina Rosenthal and Steffen Milsch for aiding in the preparation of the data for the study. I also thank Matthäus Stöhr, Julia Lakämper, and Christopher Borko for aiding in the preparation for the statistical analyses, and Dr. Inga Meyer for statistical consult.

A special gratitude goes to Dr. Ralf Holstein and the DGZfP, for providing with solutions when they were most needed and appreciated.

The studies presented in this work owe gratitude to all the participants from DGZfP, BAM, Siemens, Lise-Meitner School of Science, W. S. Werkstoff Service, and foremost, to the NDT experts from Sweden and Finland for their participation in the FMEA: Thomas Grybäck, Barend van den Bos, Johan Persson, Robert Risberg, David Åkerman, Matti Sarkimo, Aarne Lipponen, Raimo Paussu, Tommi Saastamoinen, and Petteri Raak.

I am also extremely grateful to Dr. Katja Krol and Claudia Möhring for proofreading the thesis and to all my friends for their continuous support.

Finally, I would like to express gratitude to my family. Without their love, support, and sacrifice, I would never have gotten to where I am now.

Declaration of academic integrity and deviations from the original manuscript

I hereby declare that this dissertation titled "Human Factors in Non-Destructive Testing: Risks and Challenges of Mechanized NDT" is solely a product of my independent scholarly work and that all sources from others (i.e. journal publications, books, dissertations, reports, and personal communications) have been appropriately acknowledged, i.e. cited, paraphrased, and quoted; and that the necessary permissions have been obtained. In extension, I declare that my own contributions to this work, even though published in co-authored publications, have been demonstrated truthfully.

Furthermore, I declare that the content of this version of the dissertation does not deviate from the version submitted to the Technische Universität Berlin, except in the following:

- The names of third-party figures, for which permissions from the publishers needed to be obtained (Figure 3, Figure 5, Figure 6, Figure 7, Figure 8, Figure 11), have all been re-named to comply with publishers' requirements and to maintain consistency. E.g. from "adapted from" into "From" and from "Copyright YYYY *Publisher*" into "Reprinted with permission of *Publisher*".
- For Figure 11, an additional footnote was added (Footnote No. 4).
- In Figure 7, the word *stohasic* was changed into *stochastic*.
- In Table 10, the word *Error-enforc. conditions* was replaced by *Error-enforcing conditions*.

Since this dissertation has been published by the Technische Universität Berlin (October 2015) and until it has been published by the Bundesanstalt für Materialforschung und -prüfung (BAM), Chapter "6. Empirical Study 3: Use of automated aids in the evaluation of NDT data" has been published in a journal:

> Bertovic, M. (2016). A human factors perspective on the use of automated aids in the evaluation of NDT data. *42st Annual Review of Progress in Quantitative Nondestructive Evaluation. AIP Conference Proceedings*, *1706*, 020003–1—020003–16. doi:10.1063/1.4940449

The content of this publication equals the original manuscript almost in its entirety, which was stated in the footnote of the publication. According to the copyright agreement with the American Institute of Physics signed 11.11.2015, the author has the nonexclusive right to republish the article (in print and online) without obtaining permission, as long as the a) Publisher-prepared version is not used, b) the content is not published in another conference proceedings or journal, and c) no fee is being charged. Hereby, no copyright infringement is intended and the publication is acknowledged.