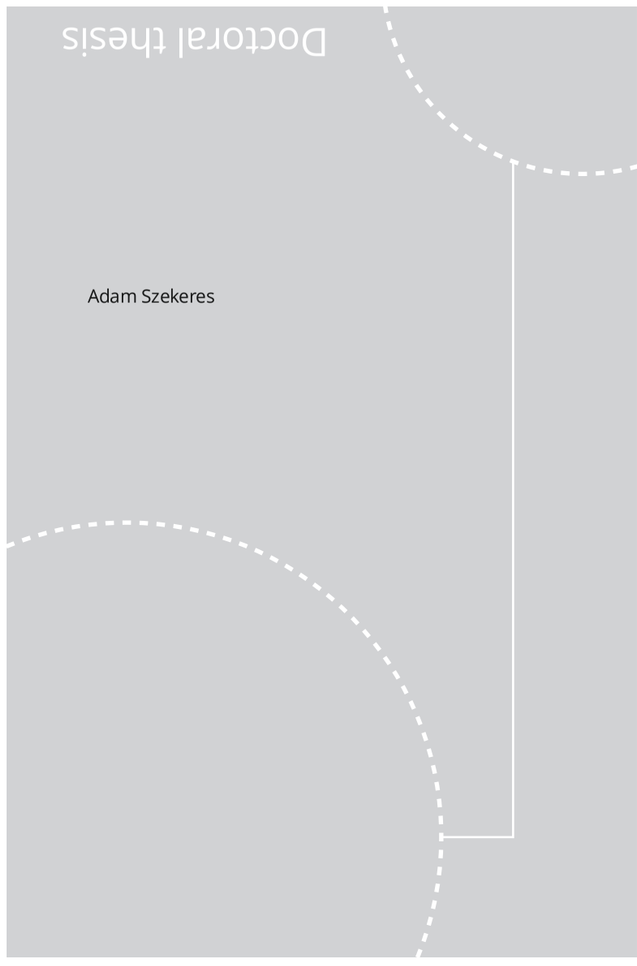
Doctoral theses at NTNU, 2020:373

Adam Szekeres

Human Motivation as the Basis of Information Security Risk Analysis

Doctoral thesis



ISBN 978-82-326-5092-7 (printed ver.)
ISBN 978-82-326-5093-4 (electronic ver.)
ISSN 2703-8084 (online)
ISSN 1503-8181 (trykt utg.)

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical
Engineering
Dept. of Information Security and
Communication Technology

Doctoral theses at NTNU, 2020:373

NTNU

NTNU
Norwegian University of
Science and Technology

Adam Szekeres

Human Motivation as the Basis of Information Security Risk Analysis

Thesis for the Degree of Philosophiae Doctor

Gjøvik, December 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

ISBN 978-82-326-5092-7 (printed ver.)
ISBN 978-82-326-5093-4 (electronic ver.)
ISSN 2703-8084 (online)
ISSN 1503-8181 (printed ver.)

Doctoral theses at NTNU, 2020:373

Printed by NTNU Grafisk senter

To all my loved ones:
small and vast,
present and departed.

Declaration of Authorship

I, Adam Szekeres declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed:

Adam Szekeres

Date:

Abstract

The electric grid represents a critical infrastructure which has an essential role in supporting societies. Therefore, it is important to identify, analyse and mitigate undesirable events that may disrupt the reliable operation of the grid. The traditional electric infrastructure is undergoing a radical transformation by the large-scale introduction of internet of things (IoT) technologies turning it into a Smart Grid (SG). Even though it is characterized by high levels of automation, people are responsible for the decisions that affect its development, operation and security. The importance of human decision-making is highlighted by the fact that the concept of security exists for a fundamental reason: stakeholder incentives can be misaligned i.e. there may exist a person who would benefit from causing a loss to another entity. While conscious attacks may take several forms and use various methods, they all require at least one motivated individual. On the other hand, there exists another class of affairs known as negative externalities which are not motivated by the explicit desire to do harm but represent undesirable side effects of conscious decisions to which another entity is exposed. The previously established Conflicting Incentives Risk Analysis (CIRA) method was built from game-theoretic and economic concepts to analyse risks due to misaligned incentives, in which the strength of human motivation plays a key role in characterizing risks. As the purpose of risk analysis is to make predictions about potential future events to guide resource allocations, CIRA relies on predictions about the behavior of key stakeholders in the future. The method's real-world applicability depends on the accuracy with which strategic stakeholder decisions can be predicted. Therefore, there is a need for the reliable and valid assessment of human motivation underlying observable behaviour. However, CIRA lacks a foundation in psychological theories which could enhance its practical utility. This thesis contributes to the literature of information security risk analysis by investigating the predictability of human

behavior and by integrating a major motivational theory into CIRA's existing framework. The work is guided by the Design Science Research (DSR) paradigm, which emphasizes that design artefacts and knowledge about their performance can be obtained by iterating through build-evaluate cycles. The behavior prediction problem is divided into two sub-problems using a person-situation (P-S) interactionist framework, which proposes that assessment of personal and situational attributes is necessary to enable improved predictions. When addressing the person side, this work assumes highly restricted environments with adversarial stakeholders who may be inaccessible for traditional psychological assessment methods and non-cooperative with an analyst, which requires the use of unobtrusive methods for inferring relevant motivational profile information about stakeholders. The thesis proposes and evaluates methods for constructing personal and situational profiles and evaluates the P-S framework to assess its practical feasibility by taking into account expected analyst performance. Furthermore, a model is proposed and evaluated which establishes a connection between CIRA and the Smart Grid infrastructure to facilitate a common understanding among stakeholders involved in the development and risk analysis of SG scenarios, and to improve risk communication. Limitations related to the specific artefacts and their implications for the general problem of human behavior prediction are identified and directions for further work are discussed with the goal of providing a better understanding about the connection between basic human motivations and the resulting risks which may pose a threat to the safety and security of societies.

Acknowledgements

This research has been conducted at NTNU i Gjøvik as part of the IoTSec project sponsored by the Research Council of Norway.

I am truly grateful to my supervisor Einar Arthur Snekkenes for inviting me into this mysterious field, for his constant support and optimism, for his patience as big as Hardangervidda and for the transformative weekly discussions bringing about and dispelling the Clouds.

I would like to thank the following people for their dedicated work which enabled me to focus on the research activities: Hilde Bakke, Nils Kalstad, Stine Terese Ruen Nymoen, Kathrine Huke Markengbakken, Urszula Nowostawska, Jingjing Yang, Marina Shalaginova. I thank Steven Furnell, Mariëlle Stoelinga for joining the evaluation committee and Basel Katt for being the administrator of the committee.

I also thank all the people who remained anonymous to me, but contributed to the successful accomplishment of the multitude of sub-tasks encompassed by the research. When things go as expected it is easily forgotten how our complex social system relies on millions of people doing their job and making effort constantly. Upon closer inspection it seems almost magical that things are not failing constantly. Thus, I acknowledge the contribution of people who made planes and luggage arrive on time, who spent time completing my questionnaires, who baked bread every day, who wrote the software I relied on and who had done a lot of invisible work to make sure disasters are avoided.

I am grateful for the support and company of friends and colleagues. I would like to thank Laura Sarnyai for her cosmic kindness and countless helpful suggestions; Reza Lashkarivand and his family for their hospitality and for providing a second home from time to time; Vivek Agrawal for his help and for making the complex

appear simple; Pankaj Wasnik, Shao-Fang Wen Steven, Gaute Wangen, Romina Muka, Mazaher Kianpour, Hareesh Mandalapu, Ali Khodabakhsh, Aland Aguirre Mendoza and others for inspiring discussions, various extracurricular activities, ski and squash sessions.

Finally and most importantly, I am deeply thankful to all my family members for their encouragement and support. I thank my sister and her little family for the delicious dinners and for her wisdom about what is important in life. I am immensely grateful to my parents for their unconditional, uninterrupted trust in me and for their love which was a great driving force at all times.

Contents

List of Tables

List of Figures

List of Abbreviations

BHV	Basic human values
BSC	Balanced Scorecard
CEO	Chief executive officer
CIRA	Conflicting Incentives Risk Analysis
DSO	Distribution system operator
DSR	Design Science Research
ESS	European Social Survey
GDPR	General Data Protection Regulation
ICT	Information communication technology
IoT	Internet of things
IS	Information security
ISO	International Organization for Standardization
MAUT	Multi-attribute utility theory
NIST	National Institute of Standards and Technology
P-S	Person-situation
PI	Personality Insights
PVQ-21	Portrait Value Questionnaire with 21 items

SG Smart Grid

SGAM Smart Grid Architecture Model

TPB Theory of planned behavior

Part I

Overview

Chapter 1

Introduction

The concept of security (of information or otherwise) exists because there is a potential for misaligned incentives. That is, at least one person may exist who would benefit from creating a loss for a particular system or entity. Complex societal systems like democracy can be disrupted by various means: influencing people by targeted ads delivered on their social networks (e.g. Cambridge Analytica case [6, 36, 83]), by hacking the voting machines [12], by bribing decision-makers or by the key decision-makers themselves (e.g. Watergate scandal [188]). Any complex system or organization has multiple attack surfaces from the level of the physical hardware including information and communication technologies through the biological wetware up to the functional level of the entity.

The electric grid is both a technical and a social system, which fulfils a fundamental role in modern societies: provides a stable and reliable supply of electricity which is a pre-requisite for all aspects of life. Economic incentives, requirements for modernisation to meet the demands of the future create a situation for critical infrastructures in which novel risks and opportunities are tightly coupled together. IoT devices are expected to become commonplace in several critical infrastructures, but as operations move to the public internet the attack surface increases significantly. Opening critical infrastructures to potential cyber-attacks represents a risk to societies due to the decisions of a small number of key decision-makers whose short-term goals may not be in alignment with the goals and values of a society.

The task of assessing whether this is an inaccurate perception of reality (i.e. paranoia) fuelled by the rhetoric of fear, uncertainty and doubt [128] or something of a real concern is the task of the risk assessment procedure. Overestimation of risks as well as underestimation has consequences like resources wasted on the wrong

controls, or increased vulnerability to unknown threats. Attacks may take various forms depending on the choice of methods. Sabotage happens at the physical level, cybercrime (i.e. sophisticated attacks or high-tech crimes) and cyber-enabled crimes (i.e. traditional criminal activity facilitated by technology [86]) represent attacks on the communication level. Gambling and manipulating prices happens at the economic level [44], while targeted influencing happens at the psychological level. Whatever the choice of method and the point of entry, attacks do not happen without a motivated person or groups having and interest in the outcomes. Technological solutions are not responsible for initiating campaigns against vaccination or against the adoption of new technologies. People are. Humans are often blamed for their cognitive limitations, biases, susceptibilities and other vulnerabilities within information security (IS). On the other hand, people are responsible for developing products, for administering systems, for the establishment of contracts and for the creation of security technologies as well as for hacking the systems [68].

Human individuals are complex systems motivated by greed, revenge, fun, sense of responsibility, purpose and several other goals. The behaviour of complex systems is of interest for a variety of disciplines. Psychology is primarily interested in understanding the behaviour (observable or mental) of humans across all domains of life. The prediction of human behaviour has great practical utility in applied and commercial settings and represents the greatest challenge for a scientific theory or discipline. The purpose of this thesis is to present the research work which aimed at investigating how the Conflicting Incentives Risk Analysis (CIRA) method can be enhanced with behaviour predictive capabilities from the field of psychology. The work is motivated by the need to increase CIRA's real-world applicability in operational IS risk analysis settings.

The following sections provide a brief introduction about the key concepts related to the research project encompassed in this thesis. Section 1.1 focuses on the fundamental research problem and motivation; Section 1.2 presents the goals of the project and the specific research questions which were investigated to tackle the main problem. Next, the articles addressing each research question are listed in Section 1.3, along with the list of additional publications in Section 1.4. The introduction concludes by defining the scope of the thesis in Section 1.5 and its structure in Section 1.6.

1.1 Research Problem and Motivation

The CIRA method is a novel approach to IS risk analysis [159] which re-conceptualizes risks. Risks in CIRA are defined as misaligned incentives between human stakeholders. Risks are characterized by the extent of disagreement between stakeholders about the desirability of certain actions. Unlike traditional risk assessment and

analysis methods which define risk as a combination of the probability of potential negative events and their consequences (i.e. probability $\times$ consequence), the risks in CIRA are entirely attributed to conscious human behaviour, where one stakeholder may lose or gain in terms of utility due to exposure to the actions or inactions of another stakeholder. As conscious intentional human behaviour is at the center of the risk concept in CIRA, the method's utility depends on its capability to predict stakeholder behaviour by assessing the desirability of certain actions from the perspective of the person capable of implementing the actions. To illustrate the concept of misaligned incentives, Table 1.1 presents a classification of human threat types to IS. Except for unintentional errors, all other categories can be characterized by the misalignment of incentives between stakeholders. However, each category comprises of a variety of psychologically and motivationally distinct behaviours. A comprehensive behaviour prediction method needs to account for all the categories which are represented by the shaded cells.

Table 1.1: Categorization of threats to information security attributed to human actions. Shaded areas are within scope of CIRA due to the strategy owner's awareness/consciousness about the potential consequences of the actions (i.e. intentionality). The strategy owner is the individual whose behavior needs to be predicted.

	Misalignment of incentives	**Strategy owner aware about consequences for self**	**Strategy owner aware about consequences for other(s)**	**Intention to cause harm**
Human error	No	No	No	No
Non-compliance	Yes	Yes/No	Yes/No	No
Motivated attacker (hackers, crackers, insiders, disgruntled employees, etc.)	Yes	Yes	Yes	Yes
Externalities	Yes	Yes	Yes/No	No

CIRA is a method under development, and its theoretical framework was constructed by using ideas from game-theory, economics and decision theory. However, it lacks foundation is psychology which is a discipline primarily interested in theorizing and investigating real-world human behaviours. Therefore, the research project aimed at integrating theories from psychology to improve CIRA's real-world applicability and capability to predict stakeholder behaviour. The research work was part of the IoTsec project funded by the Research Council of Norway. The IoTsec project focused on the security aspects of the emerging SG infrastructure. Since the SG is a highly complex, dynamic, emerging system which lacks historical data for traditional risk assessment methods, it represents a suitable test case for the CIRA method.

1.2 Research Objectives and Questions

The fundamental objective of the research project was to investigate potential approaches for improving the CIRA method with psychological theories to establish its behaviour prediction capabilities. The key motivation is to enable its application in real-world scenarios involving real human stakeholders. Furthermore, another objective was to extend the method's scope of applicability to the domain of SGs. The objective related to CIRA's enhancement was divided into two sub-problems by recognizing that in order to predict stakeholder behaviour in constrained environments, detailed information is needed about at least two objects: the person making a decision and the situation in which a decision is made.

Therefore, the first phase of the research project investigated various approaches for characterizing inaccessible and potentially adversarial stakeholders, who would be reluctant to reveal their psychological profiles to a risk analyst. This phase focuses on the development and evaluation of unobtrusive data collection methods for the construction of stakeholder motivational profiles. The second phase related to the first research goal investigated approaches for characterizing situational aspects which influence decision-makers, representing a person-situation interactionist approach to the prediction problem. The third phase aimed at integrating the findings into a useful artefact, which could facilitate the work of real-world risk analysts in SG eco-systems. Based on the main research problems and the identified objectives, the following research questions were formulated to guide the entire research project:

- **RQ 1: Which psychological theory can be integrated into CIRA to enable practically useful characterization of individual stakeholders?** Article 1 addressed this research question.
- **RQ 2: Which unobtrusive data collection methods can be utilized for building stakeholder motivational profiles, taking into account the limited access to subjects during risk analysis?** Article 1, 2, 3 investigated various approaches for building stakeholder motivational profiles.
- **RQ 3: What situational features need to be considered with respect to risk types identified in CIRA?** Article 4 aimed at establishing the connection between CIRA and empirical results from the field of moral decision-making.
- **RQ 4: To what extent does a person-situation interactionist framework improve predictive capabilities?** Article 5 assessed the framework's performance.

- **RQ 5: How to increase CIRA's applicability to Smart Grid scenarios?** Article 6 investigated how human decision-makers can be represented in a well-established architecture model of the SG.

Figure 1.1 provides a summary of the connection between research activities, research questions and the articles addressing each research question.

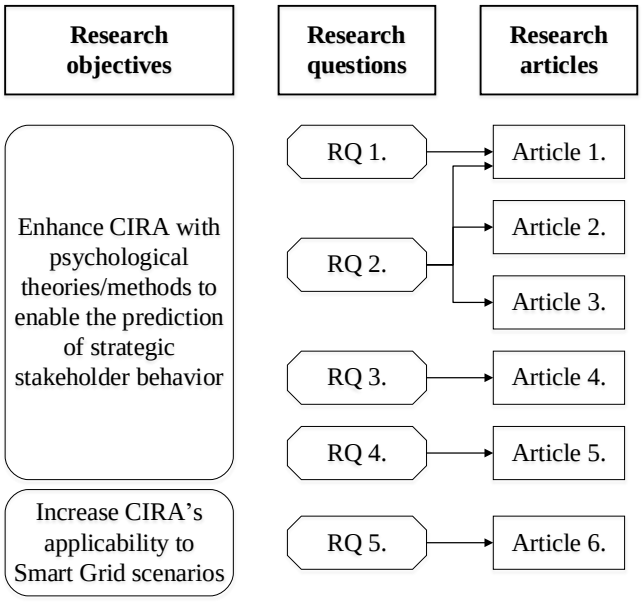


Figure 1.1: Connection between research objectives, research questions and articles.

1.3 List of research articles

This section presents the research articles addressing specific research questions within the project.

- **Article 1. [168]:** Adam Szekeres and Einar Arthur Snekkenes. Predicting CEO Misbehaviour from Observables: Comparative Evaluation of Two Major Personality Models. In: *E-Business and Telecommunications. ICETE 2018. Communications in Computer and Information Science. Vol. 1118*. Springer, Cham. 2019, pp. 135–158.

- **Article 2. [172]:** Adam Szekeres, Pankaj Shivdayal Wasnik and Einar Arthur Snekkenes. Using Demographic Features for the Prediction of Basic Human Values Underlying Stakeholder Motivation. In: *Proceedings of the 21st*

International Conference on Enterprise Information Systems - Volume 2: ICEIS. SciTePress. 2019, pp. 377–389.

- **Article 3. [167]:** Adam Szekeres and Einar Arthur Snekkenes. Construction of Human Motivational Profiles by Observation for Risk Analysis. In: *IEEE Access, Vol. 8*. IEEE. 2020, pp. 45096–45107.
- **Article 4. [166]:** Adam Szekeres and Einar Arthur Snekkenes. A Taxonomy of Situations within the Context of Risk Analysis. In: *Proceedings of the 25th Conference of Open Innovations Association FRUCT*. FRUCT Oy, Helsinki, Finland. 2019, pp. 306–316.
- **Article 5. [169]:** Adam Szekeres and Einar Arthur Snekkenes. Prediction of threat and opportunity risks: evaluation of a psychological approach using attributes of persons and situations. Under review In: *Risk Analysis: An International Journal*. Wiley-Blackwell. 2020.
- **Article 6. [170]:** Adam Szekeres and Einar Arthur Snekkenes. Representing decision-makers in SGAM-H: the Smart Grid Architecture Model Extended with the Human Layer. Accepted for publication In: *The Seventh International Workshop on Graphical Models for Security*. Springer, Cham. 2020.

1.4 List of additional publications

This section presents additional research articles published during the research project, which are not included in the thesis.

- **Article 7. [3]:** Vivek Agrawal and Adam Szekeres. CIRA Perspective on Risks Within UnRizkNow - A Case Study. In: *2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*. IEEE. 2017, pp. 121–126.
- **Article 8. [171]:** Adam Szekeres and Einar Arthur Snekkenes. Unobtrusive Psychological Profiling for Risk Analysis. In: *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications - Volume 1: SECRYPT, INSTICC*. SciTePress. 2018, pp. 210–220.

1.5 Scope of the thesis

Figure 1.2 depicts the scope and main contributions of the thesis within the disciplines of psychology and IS, specifying which steps of the risk assessment process are addressed.

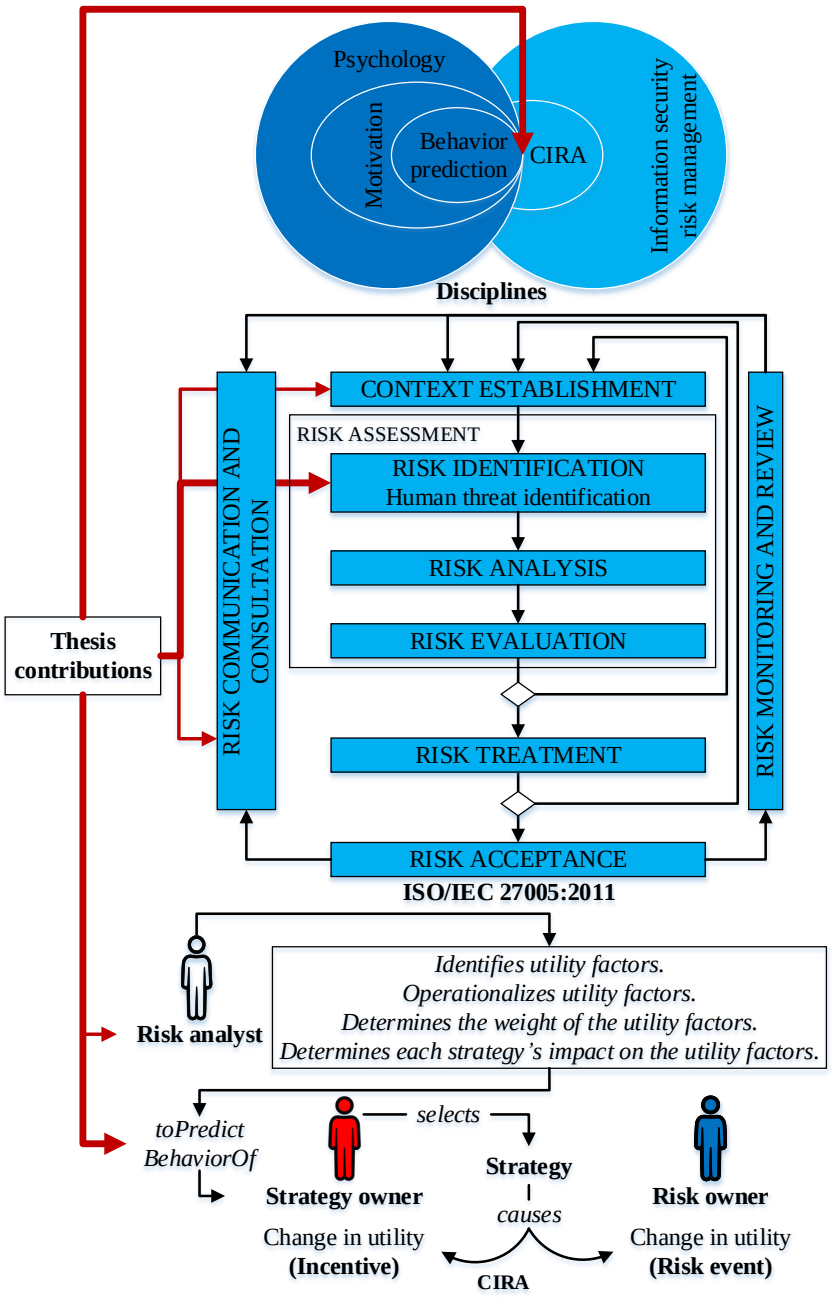


Figure 1.2: Scope and key contributions of the thesis across disciplines and within risk management: a motivation-based psychological approach to behaviour prediction applied within the domain of IS risk management.

This thesis focuses on the prediction of conscious human behaviour in adversarial settings, where traditional (i.e. direct or reactive) psychological assessment methods would be inapplicable for the purpose of characterizing subjects (i.e. strategy owners in CIRA terms) due to subject unavailability and/or subject's motivation to game the analysis.

The thesis proposes and evaluates a psychological approach to behaviour prediction which aims at combining personal and situational attributes with the mechanical behavior prediction approach to improve prediction accuracy. Furthermore, unobtrusive profiling methods have been proposed and evaluated considering the environmental restrictions on applicable methods for assessing personal attributes in real-life risk analysis settings. Situational features corresponding the risk types distinguished by CIRA are identified and organized in a taxonomy. Finally, the enhanced CIRA is integrated into the industry's most well-known architecture model to increase CIRA's applicability to SG scenarios.

1.6 Structure of the thesis

The thesis is composed of two main parts: Part I gives an overview about the research activities; Part II presents the collection of six research articles addressing the overall objectives of the thesis in separate chapters.

Part I is organized into eight chapters. Chapter 2 or the Background chapter provides a detailed overview about psychological approaches to behaviour prediction in Section 2.1 and about relevant theories of human motivation in Section 2.2. The chapter concludes with a summary of the background considering the requirements of the project in Section 2.3. An overview of the related work is presented in Chapter 3 including the description of the CIRA method in Section 3.1, the theory of basic human values (BHV) chosen to be integrated into CIRA in Section 3.2, the SG which provides the research work with connection to a highly relevant real-world problem in Section 3.3. The chapter also discusses key topics related to the use of psychology within IS in Section 3.4, the psychology of risk and situational aspects of decision-making in Section 3.5. An overview of unobtrusive profiling methods is provided in Section 3.6 and the chapter concludes with a summary of the related work in Section 3.7. Chapter 4 presents the methodology and specific methods utilized throughout the research project. The Design Science Research (DSR) is described, which was used an organizing framework for the project. Chapter 5 provides a summary of the research articles and Chapter 6 explains key contributions of the thesis. Chapter 7 overviews the limitations and provides ideas for further work. Chapter 8 concludes Part I of the thesis. Part II comprises of six research articles addressing the main research problems identified in the thesis.

Chapter 2

Background

The purpose of this chapter is to provide an overview about two main topics which are especially relevant for the objectives of the thesis (i.e. CIRA's improvement) but are scattered across the literature and a comprehensive discussion is lacking. The first section discusses key theoretical, practical and methodological issues related to human behaviour prediction and the second section provides an overview about theories of human motivation from the field of psychology. The key questions which guide this chapter are as follows: **What behaviour is to be predicted?** Behaviours within scope include individuals' one-time behaviour (i.e. no historical track record available about specific behaviour), resulting in potential threat and opportunity risks as defined by CIRA. Out of scope are predictions made by humans about future events not attributed to human behaviour (e.g. bets on the horse track); or attributed to the behaviour of a large number of people; prediction of force majeure and other natural events. **Who/what makes the prediction?** This is discussed by using the clinical-mechanical prediction approach distinction from the literature. **How data is combined?** The four potential answers include implicit relationships representing expert judgment and subjective evaluation; inference from class membership; self-reference to subject's behaviour (not applicable to one-time behaviours, with no historical record from subject's past); empirical explanatory models on cause-effect relationships (i.e. theories of motivation). **What data is combined?** Data types include personal attributes, situational attributes, person-situation attributes together, causal determinants of behavior (i.e. motivational theories).

2.1 Prediction of Human Behaviour

The purpose of this section is to enumerate and analyse existing and potential approaches to the problem of human behaviour prediction extracted from the

psychological literature to provide a better understanding about the methodological considerations behind the research work.

Risky but correct predictions in science can be considered as the goals (i.e. test of a theory [151]), while for risk management predictions are a means to an end (i.e. to modify predicted events in desirable ways). Predictions in classical mechanics require two basic components: a mathematical formula (capturing the theory or laws in an exact format which specifies the relationship between the parameters/variables of interest), and measurements of the quantities associated with the parameters/variables. Additionally, the errors associated with each measurement should be quantified since "experience has shown that no measurement, however carefully made can be completely free of uncertainties. Because the whole structure and application of science depends on measurements, the ability to evaluate these uncertainties and keep them to a minimum is crucially important" [175, p. 3].

Most psychological theories lack the exact formalism found in classical mechanics and most measurements in psychology are indirect and relate to theoretical constructs, which are "some postulated attribute of people, assumed to be reflected in test performance" [29]. Since accurate predictions are possible only if the fundamental assumptions upon which predictions are based (i.e. description, explanation of phenomena) are correct, it is important to investigate these assumptions in the literature of basic and applied research.

First, the behaviour of interest determines the choice of theory or theories which specify the most relevant variables and their relationships within scope. "The key to successful behaviour prediction is the identification of critical pieces of information that are indicative of particular behaviours..." [81]. In order to achieve human behaviour predictions in a similar fashion to that of classical mechanics, four requirements must be fulfilled, which are introduced in a clinical context: "1. the criterion to be predicted must be subject to definition and measurement; 2. experience tables or regression equations must be available beforehand in order to make statistical predictions; 3. the individuals whose behaviour is to be predicted must have had at least one clinical interview; and 4. in addition to the statistically determined variables, other data which are presumably associated with the criterion must be made available" [142]. The discussion aimed at clarifying a controversial topic, which was/is prevalent in predictions of human behaviour and relates to who or what performs the predictions. Theoretically, in the field of classical mechanics it is irrelevant who or what makes the predictions: if the same formula (theory) is used by a person or a computer, both will arrive at the same predictions (apart from potential unintended errors). However, in several applied settings (e.g. clinical settings, parole decisions, criminology, etc.) predictions about human behaviour are performed by human experts either because the relevant

statistical and theoretical relationships do not exist or because they are not utilized. Thus, approaches to human behaviour prediction can be distinguished based on the entity making the predictions. The literature distinguishes between **clinical** and **statistical/mechanical** approaches to behaviour prediction [116].

Clinical prediction

Clinical prediction approaches generally refer to subjective judgements made by an expert about a subject's (e.g. offender, psychiatric patient, candidate, etc.) future behaviour (e.g. recidivism, violence, academic/job performance, etc.). Clinical predictions can rely on empirically proven inputs (e.g. established rules between class membership and behavioural outcome) combined with the expert's judgement making the prediction. The expert may consider the behaviour and personal history of the subject, thus focusing on causal effects between inputs and outputs of prediction [39, 116]. The key feature of the clinical approach is that these predictions assume interaction between personal and situational variables (i.e. "a patient with a given set of personal characteristics will behave in a particular manner when placed in a certain kind of situation" [52]).

Mechanical prediction

Mechanical prediction approaches include statistical methods (using explicit equations), actuarial methods (similar to insurance companies' actuarial tables), and algorithmic methods (e.g., a software emulating expert judges) [66]. Mechanical approaches may utilize empirically established relationships between predictors (e.g. psychometric test scores, past behaviours, interviews) and the outcome, explicit formulas, actuarial predictions (based on class membership) and formulas to combine the parameters. The mechanical approach outputs a probability figure which is an empirically determined relative frequency.

Clinical vs mechanical approaches

Since the beginning of focused investigations [116] into clinical-mechanical approaches to behaviour prediction, a great amount of empirical evidence has been collected, enabling meta-analytic studies to compare the performance of the two approaches. Based on a meta-analysis of 136 studies in the domain of human health, superiority of the mechanical approach to behaviour prediction has been observed [65]. The evidence shows that statistical approaches are superior to clinical judgments in a wide variety of other contexts (e.g. mental health, academia, finance and business, stock picking, military training success, parole violation, violence, recidivism, advertising, marketing, personnel selection/training, etc.) [162, 32]. Thus, while both approaches can combine a variety of input data (e.g. demographics; behavioural observations; clinical interview; psychological inventories and test scores;

population base rates based on group membership like gender, age, race, etc.) and rely to some extent on empirical observations, mechanical approaches have a clear advantage over the clinical approach: they are reliable and reproducible, whereas human experts are affected by various biases and are prone to inconsistency [162].

In a conceptual analysis of the differences between clinical and mechanical (statistical) approaches, the clinical approach is described as a high-risk strategy which tries to predict all variance of behaviour, but the probability of success is low [39]. The clinical approach focuses on causal relationships between variables, relies on a deterministic assumption and strives for perfect predictions since most predictions are directly linked to high importance real-world decisions. In contrast, statistical approaches accept error, assuming that events are uncertain, where probabilistic knowledge is the best to hope for, since errors due to randomness cannot be reduced by greater knowledge and accepts that models are simplifications of reality which inherently produce errors [39].

A similar divide in psychology can be identified in the nomothetic-idiographic opposition [111] which is potentially motivated by similar concerns (i.e. what is the proper way of generating valid knowledge about human behaviour, and what is the right trade-off between knowing a person in detail vs generating statistical information?). The approaches represent two different perspectives and methodologies for research. The nomothetic (nomos-law) approach is associated with seeking general laws and utilizes procedures accepted in the exact sciences (e.g. group-centered, standardized and controlled environmental contexts, and quantitative methodologies). The idiographic (idios-peculiar) approach is associated with understanding particular events and focuses on the uniqueness of individuals by using procedures established in the social sciences (e.g. individual-centered, naturalistic environmental contexts, and qualitative methodologies) [111]. With respect to predictions, a purely nomothetic approach can be used for generating mechanical predictions and the idiographic approach is more compatible with the clinical prediction approach.

Lack of adoption of mechanical approaches

Despite the accumulated evidence clearly favouring mechanical approaches, there is a general reluctance to use mechanical predictions and there is a prejudice against algorithmic predictions. These have been attributed to the expert's illusion of validity and skill [92]; and a lack of true understanding about the interpretation of frequentist probabilities, which is often problematic even among trained professionals [76]. The difficulties of interpreting frequentist probabilities are illustrated by the following three possibilities in [15]: "1. An actual sequence of repetitions may be available; for example, a sequence of coin tosses or a sequence of independent

measurements of the same quantity. 2. A sequence of repetitions may be available in principle but not likely to be carried out in practice; for example, the polio experiment of 1954 involving a sample of over a million children. 3. A unique event which by its very nature can never be replicated, such as the outcome of a particular historical event; for example, whether a particular president will survive an impeachment trial. The conditions of this experiment cannot be duplicated. The frequentist concept of probability can be applied in cases 1. and 2. but not in the third situation." [15]. Additional factors which may contribute to non-adoption of mechanical predictions include notions like "dehumanizing", misconceptions such as "statistics do not apply to individuals", and aversion to lack of certainty explicitly stated by mechanical prediction approaches [32].

It should be noted that when information about an extraordinary event becomes known to the expert, it is reasonable to rely on the expert (known as the broken-leg rule) instead of the mechanical prediction [116], if both approaches are available. However, experts too often think that they possess extraordinary information and try to consider complex combinations of features [92]. Several investigations have explored the cognitive biases (e.g. overconfidence in subjective judgements [46] and factors (e.g. lack of benefit from experience [162], representativeness [93] etc.) producing inconsistencies in the expert's judgment hampering the reliability of the predictions. Thus, mechanical predictions are superior to human judgements in low-validity, noisy environments because algorithms can detect weakly valid cues and maintain modest accuracy through consistency [92, p. 241].

The assumptions and potential pitfalls of using aggregated, group-centered approaches to the study of individuals are presented in [161] along with corrective suggestions with relevance across the field of psychology. The single-subject approach with time series analysis is suggested as an alternative way of generating knowledge about behaviour at the individual level [72]. The method was designed to reveal the relationships of one or more variables to themselves and each other over time within an individual who is modelled as a stochastic system varying over time. This approach enables the investigation of within-person variability, which is often ignored in group-centered studies aimed at generalizing to the population.

Even though the issues associated with the clinical prediction approach (i.e. expert judgment) have been explored in detail and the non-inferiority and (in most contexts) the superiority of the mechanical approach have been extensively established, it is important to consider, potential reasons why mechanical predictions are still only slightly better than experts (i.e. not very accurate in general). Thus, why it is that when experts' inconsistencies are controlled by the introduction of mechanical methods, the improvement can be still characterized as "a progress from completely useless to moderately useful" [92, p. 230]. Expressed numerically, why is it the that

"the best statistical models appear to have maximal predictabilities expressed by correlation coefficients of 0.3 or 0.4"? [32].

Challenges for mechanical prediction approaches

Since "a mechanical formula is only as good as the input into the formula, which ideally is based on good research and adequate study of the empirical relations between identified predictors and the criterion" [162], it is important to analyse the major challenges to mechanical predictions which may include: predictor-criterion contamination (using the same measurement to assess predictors and criterion), criterion validity and fuzzy constructs (poor reliability and validity of criterion, ambiguity of constructs), low base rate problem (predicting rare events from population data), difficulty in identifying the best predictors for a given behaviour [162]. Researchers are not immune to certain biases like the fundamental attribution error (over-emphasizing the importance of dispositional explanations and under-emphasizing the role of situational influences when observing the behaviour of others, while the reverse is true when describing own behaviour) [136]. Algorithms developed by humans may also demonstrate biases which can be introduced unintentionally and may reflect existing practices, attitudes, etc. [54].

Explanation vs prediction

The epistemological asymmetry between explanation and prediction in the inexact sciences is due to a few factors. The most obvious difference is that the hypothesis of explanation concerns the past, the hypothesis of prediction concerns the future. Scientific conclusions (explanations and predictions) can be reached using three types of laws: general laws, statistical laws and quasi-laws (restricted generalizations) [75]. Quasi-laws are dominant in the inexact sciences and conclusions based on quasi-laws have a different type of uncertainty than conclusions based on statistical laws. General or universal laws allow conclusions to be reached using logical deductions with certainty. Statistical laws assert the presence of some attribute in a certain percentage of cases. Quasi-laws state the presence of an attribute in all cases for which an exceptional status cannot be claimed. While the hypothesis of an explanation needs to establish itself as more credible than its negation, the hypothesis of a prediction needs to establish itself more credible than any comparable alternative. However, in the absence of information which could narrow down the immense variety of future possibilities to a manageable size, the a priori likelihood of any particular event is extremely small. Predictions are difficult because of the critical causal importance of chance events which are essentially unknown and grant an exceptional status to several cases [75]. Thus, while explanations may have internal validity, predictions of future events lack internal validity due to lack of control over the independent variables (i.e. several alternative hypotheses can

arise).

Furthermore, challenges can be identified at the critical level where a distinction between basic and applied science is made. The goal of basic science is to develop theories that describe and explain how the world works (i.e. specify causal relationships), while the goal of applied science is to make empirical predictions and subsequent modifications in the world [157]. Statistical modelling is the established and accepted method both for testing a theory in basic science (i.e. checking the existence of the hypothesized causal relationships), as well as in applied science where predictions relate to finding a model which is best at predicting new or future observations. Models that demonstrate a high explanatory power are often assumed to automatically possess high predictive power. However, this is not necessarily the case. The field of statistics has not provided a clear distinction between explanations and predictions which resulted in a widespread confusion about explanatory and predictive modelling in several scientific fields [154].

The reason why good explanatory theories and statistical models are not necessary good predictive models, is due to several factors: the type of uncertainty is fundamentally different for explanations than for predictions and the measurable data which operationalize a theoretical construct are not accurate representations of the underlying construct. Operationalization of theories and abstract constructs into statistical models and data results in a discrepancy between the ability to explain phenomena at a semantically meaningful level and to produce predictions at the measurable level [154]. Basic science often disregards predictive modelling and considers it as of little scientific use, due to its theory agnosticism and lack of transparency for human interpretation. The conflation of explanation and prediction produces several far-reaching undesirable consequences within the affected scientific fields (i.e. inexact sciences): use of inappropriate statistical methods for a specific goal, loss of the ability to test the practical utility of theories, lost opportunities to discover new causal mechanisms, incorrect scientific and practical conclusions and a fundamental gap between research and practice. However, basic science could also benefit from embracing the predictive modelling approach which could assist in quantifying the potential predictability of a phenomenon (e.g. mobility of mobile users [160]), which in turn could result in the development of practically more useful theories [154].

The assumption that explanation by default facilitates prediction is highly prevalent in the field of psychology according to [192]. A methodological shift is recommended to solve the problems associated with a lack of predictive validity of explanatory models. Such a shift would put a greater emphasis on predictions by utilizing principles from machine learning (where the goal is to predict out-of-sample data accurately, i.e. minimize prediction error by avoiding over-fitting). The tension

between explanation and prediction can be attributed to the difference between simple models that are theoretically elegant and conceptually understandable and highly complex models (e.g. deep neural networks) which operate with representations that are incompatible with the semantically meaningful level of analysis - the level at which researchers operate and develop theoretical constructs [192].

Inferential statistics

Certain problems can arise when using statistical inference. Ecological fallacy refers to the false assumption that correlations which are established at the group (aggregate) level are the same at the level of individuals, whereas the only valid assumption is that an ecological correlation is almost certainly not equal to the correlation at the individual-level [135]. Simpson's paradox refers to the phenomenon where the direction of an association at the population level may disappear or reverse when subgroups of the population are analysed [100]. These factors seriously limit the possibility of using most of the existing research results which are reported at the aggregate level, to be applied for the prediction of individual's behaviour.

Mechanical prediction approaches utilizing inference from group membership, rely on the assumption of similarity between members of the same class. The reference class problem refers to the non-triviality and the difficulties of defining the appropriate class which is most representative of the individual whose behaviour is to be predicted. The problem arises since the number of observable properties or attributes of a subject is indefinite, therefore an indefinite number of potential reference classes can be identified each of which may provide different predictions [70]. On the other hand, the similarity assumption fails when a sufficient number of attributes are combined for a subject, thus the reference class contains only one element (n=1). When no suitable reference class can be identified, the subject's past behaviour may be the only useful predictor of the future behaviour.

One time behaviors

A simple method for behaviour prediction could be created by observing the behaviour of interest over a period of time to predict the same behaviours. While a prediction method which uses data about instances of specific past behaviours (e.g. regularity of physical exercise) in a time-series fashion can be accurate for predicting the same behaviour, it would have limited utility for predicting any other type of behaviour since its construct validity is restricted to the observed behaviour. Thus, the transferability and predictive validity of such a measurement depends on how well "regularity of physical exercise" operationalizes a specific psychological construct like death anxiety. Furthermore, studies focusing on the consistency of behaviour over time showed that it is usually not possible to predict single instances

of behaviour, but it is possible to predict behaviour averaged over a sample of situations or occasions [41]. Since in any one particular instance, behaviour is determined largely by the immediate situation, a cross-situational inconsistency can be observed. The problem is attributed to a high component of error of measurement and a narrow range of generality which characterize single instances of behaviour. When behavioural measures are averaged over a larger number of occasions behavioural stability coefficients increase to higher levels for observable behaviours and an actuarial prediction of behaviour can be achieved from a larger sample of similar behaviours to achieve better than random predictions. However, individuals are not equally predictable (i.e. within-subject behavioural consistencies show great variance between individuals) [41].

Intention-behavior gap

Intention data is frequently used to predict the behaviour of subjects. For example the Theory of Planned Behaviour (TPB), which is the most widely used framework in psychology for the prediction of conscious deliberate human behaviour uses attitudes, subjective norms and perceived behaviour control (collected directly from subjects) to predict behavioural intention, which is the immediate determinant and best predictor of real-world behaviours. Meta-analyses of the TPB show that the model on average explains 40%-50% of the variance in intentions, which drops to 19%-38% when real-world behaviour is also investigated [165]. The model uses the formalism of mechanical approaches, but since subjects provide self-evaluations about their future behaviour, it resembles the clinical approach, due to the subjectivity of self-assessments and due to the potential errors prevalent in human judgment processes. It has been shown that certain features of the environment (e.g. existence of deadlines [8], sexual arousal [7]) can have a strong impact on judgment and decision-making, demonstrating the importance of situational factors on preferences and illustrating subjects' inability to predict their own behaviour accurately.

An alternative explanation for the divergence between intentions and behaviour is provided in [113], which opposes the suggestion that individuals are poor predictors of their own future behaviour due to their inaccurate assessments. It is suggested that the reason for the intention-behaviour gap is that the amount of information available to respondents at the time of the assessment is more limited than the information they possess when the behaviour is determined. Thus, even if intentions are the best predictors of future behaviour, their utility is limited by events not yet realized at the time of conducting the assessment [113].

Existing and potential variants of mechanical approaches

There are three approaches which can be devised depending on the data used for predicting a criterion [52]. All three approaches require measurements of the criterion from the subject's past on several occasions (i.e. behavior of interest is not a one-time unobservable behavior). In the first approach predictor variables could be personal characteristics that vary over time and correlate with the criterion (e.g. mood as predictor of performance). The second approach utilizes an idea similar to Herbert Simon's, stating that the "advantage of dividing outer from inner environment in studying an adaptive or artificial system is that we can often predict behaviour from knowledge of the system's goals and its outer environment, with only minimal assumptions about the inner environment" [155]. Thus, a different class of variables could be developed based on characteristics of the situations. In the second approach predictor variables would be ratings of situational variables collected on several occasions from the subject's past which correlate with the criterion (e.g. presence of stressors in the environment as predictor of performance). The third approach would utilize both personal and situational characteristics for predicting the criterion. Therefore, the third approach assumes interaction between personal and situational characteristics [52]. It could specify how a subject with a particular set of personal attributes will behave when exposed to a particular set of situational characteristics (e.g. mood and presence of stressors in the environment as predictors of performance). However, if there are no measurements available from the subject's past about the criterion, or the behaviour to be predicted is different from the one about which measurements are available, reference classes must be used (e.g. subjects with similar personal features and situations with similar features).

Results showing that the interaction of individuals and situations account for more variance than either source of variance alone [41] motivate explorations of the utility of including situational attributes in mechanical predictive models. Additionally, the investigations focusing on expert judgment and intuition analysed attributes of the environment to identify the conditions which are necessary for the development of true expertise and the conditions which inhibit the development of intuitions. Reliable intuitions can develop when predictable regularities exist in the environment; subjects have the opportunity to learn the regularities through practice; and immediate feedback on the performance is available [92].

Based on the surveyed literature two basic methods can be constructed within the category of mechanical approaches to behaviour prediction using empirical observations. The first category of methods may use empirical relationships established in other people, while the second class of methods may use the subject's past behaviour to overcome the reference class problem.

Figure 2.1 presents a classification of existing and potential mechanical methods which arise from various combinations of the attributes used for the predictions (i.e. attributes of persons, situations, both). In general, only the first variant is utilized (i.e. attributes of persons). Relevant assumptions of this approach are as follows: Y (outcome of interest) is one-time behaviour of the subject, no historical data from subject's past is available, but outcome is observable in case of other people; X (predictors) are stable attributes; similarity between entities (i.e. reliance of reference classes).

Potential within-subject mechanical approaches to behaviour prediction based on empirical observations established in the same individual are depicted in Figure 2.2. The relevant assumptions are as follows: Y (outcome of interest) is observable for the subject, (i.e. historical data from subject's past is available); X (predictors) vary over time; self-similarity of subject, (i.e. reference classes not required). Within both categories the 3rd variant assumes interaction between attributes of persons and situations.

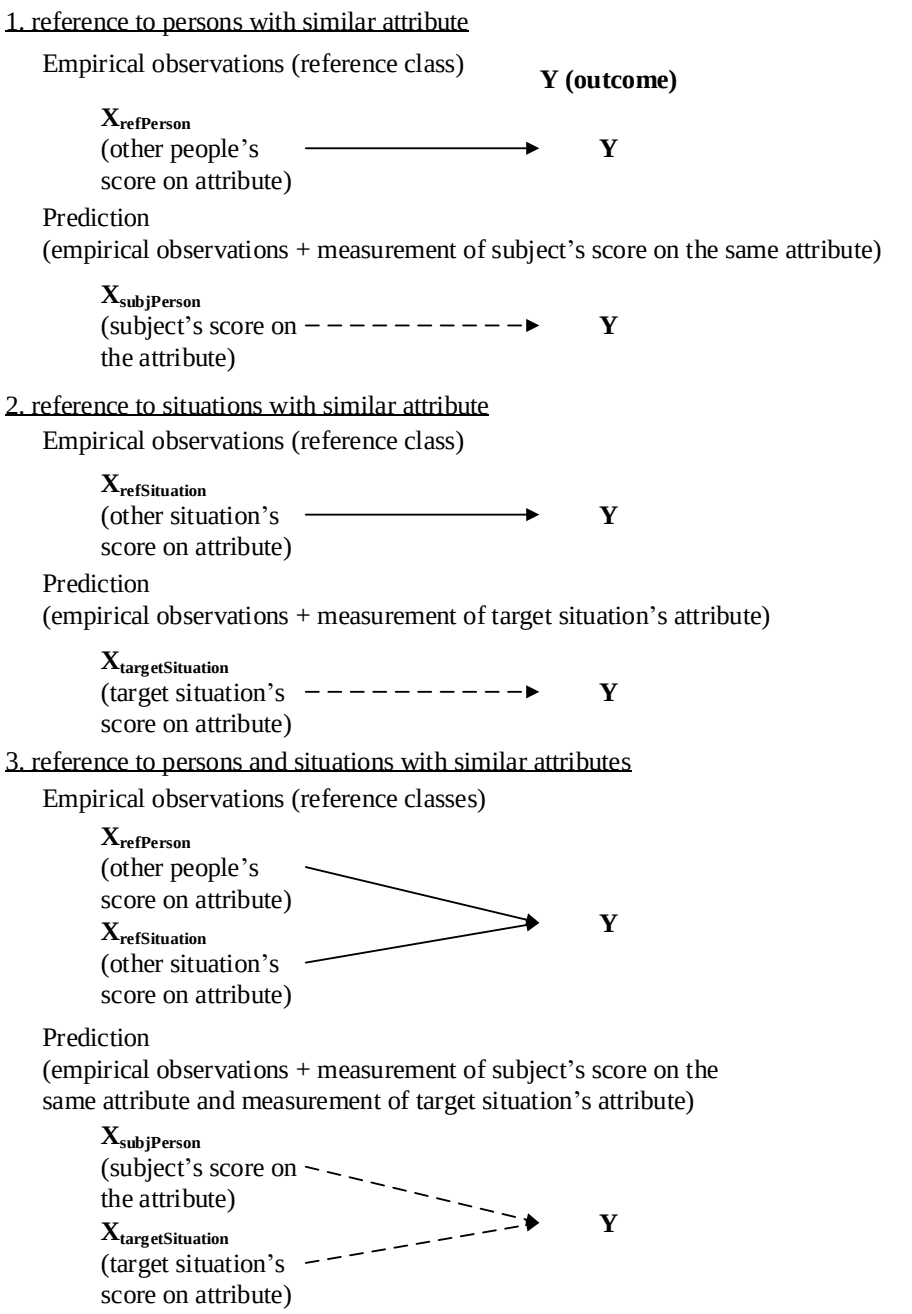


Figure 2.1: Potential approaches for behaviour prediction using attributes of persons, situations and combination of both. This approach of behaviour prediction relies on reference classes.

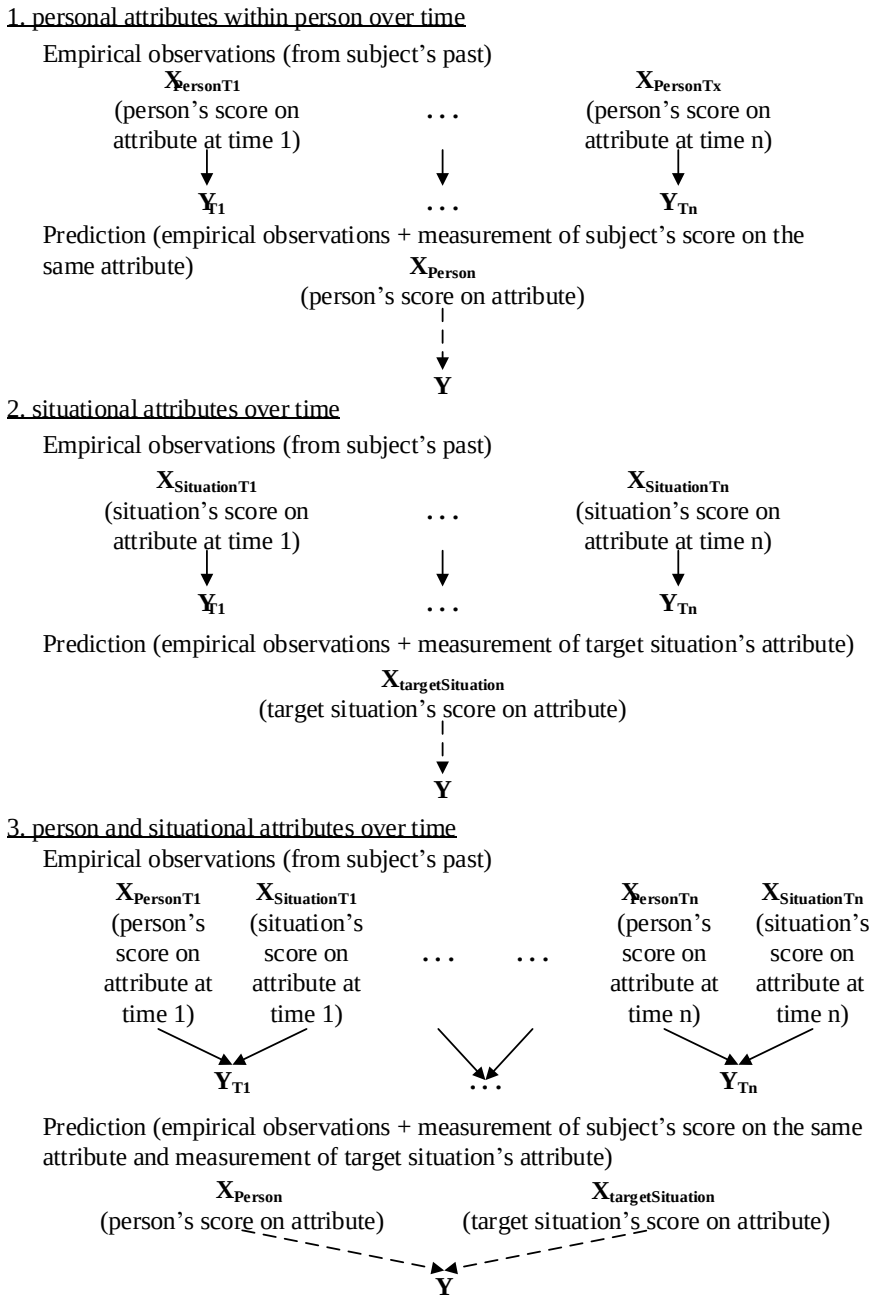


Figure 2.2: Within-subject approaches for behaviour prediction could overcome the problem of reference classes, however, assume that the behaviour of interest (outcome) can be observed in the subject's past.

A third approach which can be utilized for behaviour prediction focuses on causal determinants of behaviour, specifying the contents or processes which gives rise to motivated actions (i.e. theories of motivation). Such approaches may potentially overcome some of the challenges associated with the reviewed approaches (e.g. reference class problem and lack of historical data from the subject's past), since they require a one-time measurement on the relevant variables to make predictions about future performance or behaviour. An overview about theories of human motivation is provided in the following section.

2.2 Theories of Human Motivation

The purpose of this section is to provide a overview about motivational theories considering the objectives of the overall research project: prediction of stakeholder behaviour assuming no direct interaction between risk analyst and the adversarial subjects. Due to the vastness of the knowledge base and several other challenges (discussed below) the purpose is not to provide a comprehensive discussion on the topic of human motivation, but to introduce the literature giving a high-level overview. Therefore, the survey starts with presenting recent studies which aim at structuring, systematizing the field. Next, major theories of motivation are briefly surveyed. To this end, two of the most recent systematizing articles (i.e. [50, 16] were used to identify relevant theories, and a literature search was conducted on all the theories included in these two articles to identify meta-analyses of the theories. Due to a lack of empirical comparisons between theories, meta-analyses can provide a picture about the extent to which theories provide practically useful results, given that sufficiently high number of empirical investigations have utilized them.

Motivation (and emotion derived from Latin *moveo, movere* means to move) is an often-used concept in psychology as well in many applied settings with several meanings (e.g. *Willingness of action especially in behaviour. The action of motivating. Something which motivates. An incentive or reason for doing something.* [120]). Thus, motivation may refer to desired end-states, determinants of behaviour, selection of courses of actions, maintenance of performance, actions aimed at increasing performance etc. Human activity spans across various contexts, levels of consciousness (i.e. from dreaming to intentional goal-directed behaviour) throughout the lifetime of any individual. At the most general level, the study of motivation aims at describing, explaining, predicting and modifying what humans do. The constructs and theories vary in terms of their method of development (e.g. naturalistic observations, projective tests, analysis of biographies, factor analysis, lexical sources, subjective considerations, etc.), level of analysis (e.g. instincts, biologically determined drives, needs, social and cognitive motivations, conscious/non-conscious goals, values, etc.) and scope (e.g. general principles vs.

task-specific motivations) [9].

The combination of ambiguity of the term and breadth of behaviours has lead to a knowledge base which is characterized by a vast proliferation of theories and constructs developed over the past in order to address various aspects of human behaviour. The field demonstrates a high degree of disorder which is often mentioned by the few researchers attempting to reconcile discrepancies and to systematically organize the literature. Observations about the challenges of the literature include diversity in use of terminology and methods, confusion of constructs, abundance of micro-theories, imbalance between knowledge generation and application [10].

In order to answer the general question of **"Why people do what they do?"** a useful distinction is to divide the literature based on the questions **"What motivates people?"** referring to the **content** of theories and **"How motivated action takes place within individuals?"** referring to the **processes** as presented in [10, 177]. It should be noted that the following articles utilize arbitrary criteria for organizing their subject matter, which may result in incomplete or incompatible classification systems.

A huge number of theories and constructs were included in a review which focuses on goals (i.e. internal representations of desired states spanning from biological needs to conscious desired outcomes). Goal constructs are analysed in terms of their structure (i.e. properties, organization and dimensions of goals), process (i.e. establishing, striving toward and revising goals) and content (i.e. classification of outcomes or states that people approach or avoid), across the cognitive, personality and motivational domains at the individual level [10]. Goals are most often conceptualized in a hierarchical structure, in which sub-goals are grouped under various higher order goals. Goal processes encompass the behaviours and cognitions related to establishing, planning, striving for, and revising goals. Goal contents have been most often explored using a mixture of idiographic and nomothetic self-report methods. Early conceptualizations of goal contents focused on instincts and needs (biological and acquired), then research interest shifted toward self-concept, values, life tasks and personality factors.

A total of 135 goal concepts were used to construct a taxonomy of high-level goals in [24]. The taxonomy contains several goal-like constructs (e.g. needs, goals, values) extracted from the literature which were sorted by subjects based on similarity judgments, thus the taxonomy is based on empirical data (i.e. consensus among raters). The sorted goal lists were subjected to cluster analysis, giving rise to 30 conceptually meaningful clusters. At the top level, a distinction between interpersonal (social) goals and intrapersonal (individual) goals emerged. The benefit of

such a taxonomy is that it provides a comprehensive, structured description and a common terminology for further investigations.

The lack of systematic integration of the literature motivated the work presented in [50]. The unified model is based on theoretical analysis of the literature and builds on the assumption that all motivated actions aim at change. Thus, two questions could be asked which provide the organizing principle for the model: "Change where?" and "What type of change?". Answers to these questions give rise to a 3 by 3 matrix where rows correspond to three answers to the "Change where?" question (i.e. within the self, within the material world, within the social world) and columns correspond to answers to the "What type of change?" question (i.e. potential/expectations for life, process/experience of living, and outcomes/evaluation of life activities). The model organizes a total of 106 motivational constructs developed by 46 theorists from the 20th century and gives rise to nine distinct motivational domains (i.e. security, identity, mastery, empowerment, engagement, achievement, belonging, nurturance, esteem), excluding biological, physiological drives, avoidance motives and aggression [50].

A narrower theoretical work is presented in [16], which develops a theory of task-specific motivation defined by the level of readiness to take a specific action. A total of six theories are integrated. It is noted that explicit or implicit overlap between concepts and theories, conflicting views represent major challenges in the field, making it difficult to draw conclusions or to combine the results of various studies. The work conceptualizes intrinsic and extrinsic motivation (i.e. distinction between person and context/situation), identifies task specific immediate antecedents of motivation and distinguishes between approach and avoidance motivations. The model uses valence expectation to capture the interaction between positive and negative, affective and cognitive valences expected by choosing an action. Furthermore, a distinction is proposed between personal (benefit for the self) and non-personal valences (benefit for other people or entities). The combination of the six main motivational theories gives rise to the model in which an action's immediate antecedent is readiness for action and 10 other factors have influence on readiness for action, three of them directly: subjective norms, feasibility expectations and valence expectations. Valence expectations consist of two sub-categories: affective valences (positive or negative) and cognitive valences (positive or negative); personal or non-personal. Both valences (affective and cognitive) are influenced by four factors: subjective norms, sense of personal relatedness, feasibility expectations and sense of personal autonomy. Subjective norm has one antecedent: sense of personal relatedness. Feasibility expectations are influenced by two factors: sense of personal competence and perceived external support. Sense of personal autonomy has one antecedent: perceived freedom of action. The model proposes a structure of

concepts derived from previous theories, but its evaluation in terms of validity and performance is lacking [16].

The following paragraphs briefly overview the theories ($k_{\text{initial}} = 106 + 6$) identified from the previous two reviews (i.e. [50, 16], which have been utilized in sufficiently high number of empirical studies to enable evaluations using meta-analytic techniques ($k_{\text{final}} = 19$). This way of scoping enables structuring the literature based on the theories' practical utility. Where possible the name of the original theory's/construct's developer is provided to avoid construct confusions.

The **cognitive evaluation theory** (Edward L. Deci) aimed at explaining the observation that extrinsic rewards have a detrimental effect on intrinsic motivation. Based on 45 independent studies and 88 effect sizes the existence of the phenomenon has been confirmed in studies which operationalized the key construct of the theory adequately [138].

Self-efficacy expectations (Albert Bandura) are assumed to influence behaviours through three ways: choice of behaviours, effort expenditure and persistence, state of physiological arousal. Based on 21 studies, it has been shown that self-efficacy expectations are related to task-performance and choice of behaviours. Effects were higher in laboratory setting, than in field studies [140].

Expectancy theory (Victor Vroom) states that a criterion (e.g. performance, effort, intention, preference, choice) is a function of three factors: valence (affective orientation toward outcomes), instrumentality (probability of obtaining an outcome) and expectancy (subjective probability that effort leads to an outcome). Two sets of analyses used 77 independent studies to test within-subjects and between-subjects associations between criterion and the overall model. Results show low average correlations, suggesting a poor of validity for the model [180].

Path goal theory (Robert J. House) was built on expectancy theories of motivation and proposes that the primary role of organisational leaders is to enhance employee expectancy, instrumentality and valence by coaching, guidance, support and other rewards. Based on the analysis of 103 articles covering 120 studies, results show major methodological limitations (e.g. lack of established instruments, conceptual deficiencies, etc.) and only weak partial support for the theory's propositions [189].

Protection motivation theory (Ronald W. Rogers) was developed in the field of healthcare and proposes that actions to protect oneself against a threat are a function of: perceived threats; the desire to avoid negative outcomes; a cost-benefit analysis which weighs the cost of precautionary actions and the benefits of the actions; and the perceived effectiveness of precautionary actions. The theory aims to capture two cognitive processes which operate when people evaluate threats (threat-appraisal)

and select a coping alternative (coping-appraisal). Based on 65 studies, the results show a strong support for the theory's validity and utility in practical settings for modifying subjects' health-related behaviours [49].

The **transtheoretical model of behaviour change** (James O. Prochaska and Carlo Di Clemente) was initially developed to help subjects change their addictive behaviours. The model proposes five stages of change in which subjects have different intentions and behaviours: precontemplation (lack of intention), contemplation (thinking about change), preparation (small changes implemented), action (changes implemented recently), maintenance (change implemented over 6 months). Based on the analysis of 71 articles (91 independent samples) from the physical activity domain, the model is supported by empirical evidence (i.e. relevant constructs differ in various stages and in the predicted directions). However, there is a need for standardized, reliable instruments [114].

The **theory of planned behaviour** (Icek Ajzen) proposes that a (volitional) behaviour's best predictor is an individual's stated intention to engage in the behaviour. Intention mediates the effect of attitude, subjective norm and perceived behavioural control on the behaviour. The analysis of 72 studies from a physical activity context provide a strong support for the proposed relationships among the theory's constructs and self-efficacy has significant, independent contribution to the theory for explaining behavioural intentions and behaviours [69].

The **theory of goal setting** (Edwin A. Locke and Gary P. Latham) has expanded gradually over the years and aims at facilitating high-levels of performance (modifying behaviour). The theory is based on the observation that difficult goals produce the highest levels of effort and performance. Furthermore, it emphasizes the importance of feedback, commitment (enhanced by self-efficacy and goal importance), task complexity and situational constraints as moderator variables [109]. An analysis of 11 studies with 16 effect sizes investigated the effect of goal setting only and the combined effect of goal setting with feedback on performance. Results confirm that goal setting plus feedback increases performance for difficult tasks [123].

Need for achievement was one of the three core needs (need for power, need for affiliation) formulated by David McClelland. It was hypothesized that individuals scoring high on need for achievement are more likely to engage in innovative, entrepreneurial activities involving some risks, than people low on this attribute. Based on 41 studies, occupational choice and performance is significantly related to the level of need for achievement, in support of the original theory [26].

Locus of control (Julian B. Rotter) refers to an individual's self-evaluation regarding the perceived control over their circumstances. Individuals with an internal

locus of control believe they are masters of their fate, feel confident and perceive a strong link between their actions and circumstances, whereas individuals with an external locus of control believe they do not have control over their fate and tend to attribute outcomes to luck or external factors. Analysis of 222 articles showed that internal locus of control is positively associated with job satisfaction, work commitment, task motivation, expectancy, instrumentality, job involvement, self-efficacy and problem-focused coping, supporting the concept's construct validity and practical utility for interventions [125].

Terror management theory (Jeff Greenberg, Sheldon Solomon, Tom Pyszczynski) proposes that people embrace cultural beliefs, symbolic systems, values and try to maintain self-esteem in order to cope with the awareness of their unavoidable mortality. Experimental investigations focus on priming subjects with their own mortality (e.g. subjects are asked to write about their own death - mortality salience) to test whether the intervention produces a greater adherence to cultural worldviews and self-esteem (e.g. change in subjects' attitudes toward an author who disagrees with their worldview). The analysis of 164 articles showed that priming with mortality salience has a robust overall effect in experimental settings [18].

Self-regulation theory analyses an individual's capacity to control their own behaviour over time and across changing circumstances, which is a key capability for maintaining performance. Four processes are involved in self-regulation: self-monitoring, self-evaluation, self-reactions, self-efficacy evaluations. An analysis of 102 articles showed that all the self-regulation variables have a positive association with mastery-approach orientations (i.e. motivation to succeed vs. motivation to avoid failure) defined as a stable personality trait, but the effects are weaker when actual performance is considered [21].

Self-determination theory (Edward L. Deci and Richard Ryan) is one of the most comprehensive general theories of motivation, integrating a total of six previous micro-theories/constructs into its framework. It explicates the psychological processes and external conditions for optimal performance and functioning. Individuals are assumed to possess innate tendencies for growth and development, but these tendencies require supporting environments. Satisfaction of needs for autonomy, competence and relatedness are prerequisites for a high level of performance and well-being. Intrinsic motivation is viewed as the most optimal type of motivation which is self-rewarding and associated with satisfaction of autonomy needs, intrinsic regulation and internal locus of causality. Extrinsic motivation or instrumental motivation is associated with varying levels of need satisfaction and various forms of regulation (integrated - satisfaction of needs, identified - personally held values, introjected - avoiding external disapproval or gaining external approval, external - gaining external rewards or avoiding punishment). At the lowest end

of the motivational spectrum is amotivation which is associated with lack of need satisfaction and lack of intentionality [139]. A meta-analysis from healthcare context used 184 datasets and analysed self-determination theory's constructs when applied to facilitate behaviour change in patients by fostering autonomy to achieve intrinsic motivation. The findings showed that the theory's constructs (personal and contextual) are related to each other and to health outcomes, and the direction of relations is generally in agreement with the theory's propositions [124].

Sensation-seeking (Marvin Zuckerman) is a narrow personality trait referring to the desire to engage in novel, stimulating, risky experiences. People high on the sensation-seeking trait are more likely to pursue dangerous hobbies, engage in gambling and risky sexual activities, etc., than individuals scoring-low on the trait. Men tend to have higher average scores than women across populations. Based on an analysis of 72 articles, using the same instrument for assessing the trait, it has been shown that overall sex differences in relation to sensation-seeking are stable across time, and effects are robust [30].

Social cognitive theory (Albert Bandura) is an explanatory framework which states that learning primarily takes place in a social context where person, environment and behaviour reciprocally determine each other. Self-efficacy (belief about self-competence to complete a certain action), outcome expectations (beliefs about the consequences of (not) performing an action with dimensions: physical, social, self-evaluative), socio-structural factors (facilitators, impediments) are key constructs of the theory, directly influencing the goals (distal or proximal) which are the immediate antecedents of behaviour. An analysis of 44 articles from a physical activity domain showed that models using the social cognitive theory for predicting physical activity accounted for 31% of the variance across studies, demonstrating the theory's validity and usefulness [193].

The **approach-avoidance achievement goal theory** of motivation (Carol Dweck, Ellen Leggett and Andrew Elliot) distinguishes between the valences attached to performance goals, where approach goals refer to attaining competence, and avoidance goals refer to avoiding incompetence. The dominant way an individual thinks about his/her own performance has implications for performance. The analysis of 17 articles from the field of sport psychology showed that among the multitude of variables, the distinction between approach and avoidance goals had a significant effect on performance, supporting the theory's predictions [108].

Self-affirmation theory (Claude Steele) investigates how individuals cope with information that is threatening to their self-concepts. The technique of self-affirmation aims to restore self-perceptions of adequacy to overcome resistance and defensive responding against threatening information and to facilitate necessary behaviour

changes (in the context of health and education). Based on the analysis of 41 articles (144 effects in total) from the domain of healthcare, it was shown that subjects receiving self-affirmation intervention showed greater message acceptance, stronger motivation for change and healthier behaviour compared to control groups. The small, but significant effects were observed across various settings, providing some support for the theory's adequacy [42].

Self-discrepancy theory investigates the associations between emotional states (positive and negative) and self-evaluations (i.e. subject's self-perception) and specifies three versions of the self: actual self (i.e. the attributes possessed by the person), ideal self (i.e. attributes the person would like to possess) and ought self (i.e. attributes deemed important by the person). The theory proposes that discrepancies between the three self-representations account for their affective and motivational significance (e.g. emotional vulnerability, psychopathology). Based on the analysis of 70 articles, small positive effect sizes were observed between self-discrepancy, various psychopathologies and negative emotions. Small negative associations were observed between self-discrepancy and self-esteem, in support of the theory [115].

The **ARCS model** of motivation (John M. Keller) was developed for educational settings and the name refers to the four variables (i.e. attention, relevance, confidence, satisfaction) in the model which specify factors relevant for maintaining student motivation for optimal learning. Teaching materials may be developed based on the recommendations of the model. In an analysis of 26 studies, it has been shown that following the model's recommendations for constructing teaching materials, results in positive changes in student motivation, and materials mainly influence motivation through attention [35].

2.3 Summary of chapter considering the requirements of the project

Based on the surveyed literature of behaviour prediction and motivational theories, the behaviour of interest (outcome) has a key role in selecting an appropriate theory or construct. Behaviour needs to be subject to definition and measurement, there is a need to measure the relevant attributes of the subjects and empirical relationships are necessary to generate valid predictions. Mechanical predictions are superior to expert judgments in a variety of contexts due to their better reliability. While experts often try to minimize all errors, mechanical prediction approaches make less error by accepting error as inherent in low validity environments. Therefore, a mechanical approach is preferred for the project to minimize analyst involvement.

Since no method is perfect, and even mechanical predictions generate a signific-

ant amount of unexplained variance, it is important to analyse the fundamental problems associated with inferences. This activity revealed several practical and fundamental challenges: reference class problem, different uncertainties associated with predictions and explanations, confusion about explanatory models and predictive models, ambiguity of constructs, discrepancy between constructs and measurable data introduced by operationalization, biases of researchers, availability of information about subject's or other's behaviour. A potentially useful approach for overcoming some of the challenges is to combine approaches by considering their strengths and the requirements of the project.

While motivational theories are rarely tested in a truly predictive fashion, they are often utilized for modifying behaviour. The seeming paradox (i.e. rarely used for predicting behaviour but often used for controlling behaviour) can be resolved by considering the reduction of uncertainties enabled by the control over some aspects of the environment. Control over the environment increases internal validity of the situation (similar to experimental procedures), enabling useful modifications of behaviour, without necessarily relying on predictions.

Nevertheless, the mechanical approach combined with a motivational theory could be a viable approach for overcoming the key limitations associated with inaccessibility of subjects and lack of historical record of previous behaviour in the context of CIRA's application (i.e. outcome to be predicted is one-time behavior). Doubt has been expressed whether predicting a specific person's behaviour would be of scientific interest: "...ask yourself what problem will be solved by studying person by situation interactions. It is an effort to predict what single individuals will do at single points in time. This is a point prediction. But science is rarely if ever about point predictions..." [79]. Whether such predictions are deemed scientific or not, correct predictions would be of great practical utility in many cases, for example predicting the behaviour of certain individuals who are in command of powerful military forces [52].

Furthermore, the practical advantages of predicting the behaviour of individuals is demonstrated by the success of world-leading organizations (e.g. Amazon, Netflix, YouTube, etc.) which rely on sophisticated recommender systems (predictive systems) for offering their services to users [134] (i.e. predicting/controlling what a specific person does at a specific point in time). The requirements for the prediction method in CIRA are summarized in Table 2.1 and compared to key features of a generic recommender system. A rough estimate about the valuation of improved human behaviour predictions can be gained from Netflix's famous competition in which the company offered a $ 1 million prize for an improvement of 10.06% prediction accuracy [121] over the baseline prediction accuracy (4.75%) in 2009 [122].

Table 2.1: Comparison between operational requirements of CIRA and a typical recommender system, both of which makes predictions about future behaviours.

	Requirements for CIRA	**Recommender system (e.g. collaborative filtering [134])**
Goal	Prediction of behaviour	Prediction of behaviour
Object of prediction	Utility of a choice (broad range of conscious behaviours representing threat and opportunity risks)	Utility of an item (user interest / likelihood of selecting an item)
Relationship between observed behaviour - predicted behaviour	Context mismatch	Within context of application
Profile	Psychological	Behavioural (past rating of items)
Purpose of profiling	Connect observed behaviour in context A with behaviour of interest in context B	Match user preferences with similar user profiles based on past behaviour
Assumption of profiling	Stability of psychological profile, similarity of people, transferability between contexts	Similarity between users
Type of information available for profiling	Public observables	Revealed preferences within the system with relevance for behaviour of interest
Availability of data for profiling	Restricted	Big Data
Method of data collection for profile	Unobtrusive, indirect inference	Direct
Assumption about humans	Non-cooperative, adversarial	Interested in getting relevant recommendations, acceptance of terms and conditions
Additional requirement	Transparency to human interpretation	Increase turnover

Chapter 3

Related Work

The purpose of this chapter is to provide an overview about the key research results relevant for the objectives of the thesis. The chapter is organized into seven sections. The overview starts off by presenting the Conflicting Incentives Risk Analysis (CIRA) method's key concepts and novel approach to risk analysis in Section 3.1. Section 3.2 presents the theory of Basic Human Values (BHV) which was used for modelling stakeholder utility factors in CIRA throughout the thesis. Next, the Smart Grid is introduced as a critical infrastructure enhanced by IoT technologies in need of novel risk analysis methods in Section 3.3. In Section 3.4 an overview about using psychology within the field of IS is provided. Section 3.5 briefly presents results related to the psychology of risk and situational aspects of decision-making. Section 3.6 presents several approaches for inferring psychological profiles unobtrusively, while Section 3.7 concludes the chapter by summarizing the connection among research results considering the objectives of the thesis.

3.1 Conflicting Incentives Risk Analysis method

The Conflicting Incentives Risk Analysis (CIRA) method developed by Rajbhandari and Snekkenes [130] combines ideas from game-theory, decision-theory and economics to overcome certain problems associated with several risk analysis methods which rely on the frequentist notion of probability to characterize risks (e.g. ISO 27005 [85]). Lack of historical data for reliable probability estimations in case of dynamic and emerging systems, insufficient methodology for addressing risks related to conscious human decisions and enormous complexity are three key challenges that may restrict the practical utility of traditional risks analysis methods.

CIRA focuses on key components of any complex system to construct a novel concept of risk: decision-makers (i.e. stakeholders); their potential actions; and the expected consequences of the actions. By focusing on conscious human decisions CIRA shifts the level of abstraction from the technical aspects, resulting in a different notion of risk, which does not rely on probability estimations from the system's past behaviour. CIRA uses the concept of utility (comprising of several utility factors) to model individual stakeholders (i.e. real persons) [131]. Two classes of stakeholders are distinguished: the **risk owner** and the **strategy owner**. The risk owner is the person exposed to the actions or inactions of the strategy owner (i.e. facing a risk). The strategy owner is capable of executing certain actions which have an impact on both stakeholders' overall utility. Risk is conceptualized as the extent to which stakeholder incentives are misaligned. Risk is subjective for the risk owner and is expressed as a pair of numbers (incentive, consequence) capturing the strategy owner's strength of motivation to implement an action and the consequences for the risk owner (in terms of change in overall utility). The conflicting incentives conceptualization of risk gives rise to two types of risks: **threat risk** and **opportunity risk**. Threat risk refers to situations where the strategy owner could benefit from an action which produces a loss for the risk owner, whereas opportunity risk refers to situations where the strategy owner would have to take a loss in overall utility to increase the risk owner's overall utility.

The method relies on predicting the strategy owner's intentional future behaviour (choices) to characterize risks. Predictions refer to assessing action desirability for the strategy owner by operationalizing relevant utility factors, assessing their weights and assessing how actions change the values of the utility factors. The procedure needs to be conducted without relying on direct interaction between the analyst and strategy owner (i.e. assuming inaccessible, adversarial subjects). Risk mitigation in CIRA is about seeking alignment between stakeholders (i.e. modification of the weights assigned to the utility factors or modification of the extent to which actions change the values of the utility factors [159]). Misalignment of incentives can encompass threats generally attributed to conscious human behaviour within established risk analysis methods (e.g. non-compliance, attackers with explicit intention to do harm, insiders, etc.), as well es externalities (i.e. side effects of conscious decisions) resulting from operating in a highly complex environment. CIRA assumes that individual decision-makers are responsible for the existence of risks and that the appropriate level of analysis is the individual decision-maker who benefits/suffers from decisions which have impact on other stakeholders. While several well-known risk analysis methods (e.g. ISO 27005 [85], FAIR [91], NIST 800-30 [90], etc.) acknowledge the importance of focusing on human threats, none of the methods place human behavior at the center of the entire risk analysis procedure. Thus, CIRA represents a unique and radical approach, requiring extensive work

to explore and extend its capabilities to maximize its potential benefits. Even though CIRA focuses on conscious human behaviour it lacks a foundation in psychology. Thus, the integration of suitable psychological theories with a proven track record of practical utility is needed to enhance the method's applicability to real-world cases involving real stakeholders.

3.2 Theory of basic human values

The theory of basic human values (BHV) integrates several previous theories of motivation (Hofstede, Rokeach, etc.) which identify values as key constructs for explaining social and personal organization and change [147]. Values are central concepts for characterizing societies, cultural groups, individuals and for explaining the motivational basis of behaviour [147]. The BHV theory was developed to identify a set of universally recognized values which can parsimoniously capture fundamental requirements of human existence: biological needs of individuals, requisites of coordinated social interaction, survival and welfare need of groups. Conceptual definitions were generated to capture six key features of the proposed values [146]: 1. Values are beliefs linked to affect. When values are activated, they are emotionally charged. 2. Values capture abstract, desirable end-goals that motivate action. 3. Values are trans-situational, unlike context-specific norms or attitudes. 4. Values are standards or criteria, which guide the selection and evaluation of actions, people and events largely unconsciously. 5. Values form a hierarchy within individuals. 6. Actions are guided by the trade-off between competing values, given that they are activated in a decision context, and important to the subject (i.e. central to the self-concept).

Thus, the theory of BHV specifies the motivational content of desirable end-goals and suggests the process (i.e. trade-off between values) which produces behaviour. The theory is in continuous development, but theoretical and empirical research generated a list of 10 basic values, which are universally recognized in all cultures (i.e. if a proposed value was not recognized in a specific region it was excluded from the final list) [146]. The abstract goals captured by the ten values can be briefly summarized as follows [146]:

1. **Power**: need for social status and prestige, control over people and resources.
2. **Achievement**: need for success and social recognition by demonstrating competence.
3. **Hedonism**: importance of pleasure experienced by the satisfaction of sensuous needs.
4. **Stimulation**: need for variety, novelty and excitement.

5. **Self-direction**: need for autonomy, independence in choosing, acting, exploring, etc.
6. **Universalism**: care for the welfare of all people and nature.
7. **Benevolence**: need to maintain and promote the welfare of others with whom one is in close contact (in-group).
8. **Conformity**: inhibition of actions that would violate social norms or expectations.
9. **Tradition**: acceptance and respect of cultural values, customs, religion.
10. **Security**: need for social order, safety of nation, relationships and self.

The ten values form a circular structure presented in Figure 3.1 grouped into four higher level dimensions. Adjacent values are motivationally more compatible, whereas values on the opposite sides of the circle are in conflict. Decisions are guided by the expected changes in the relevant value scores as a result of selecting an action.

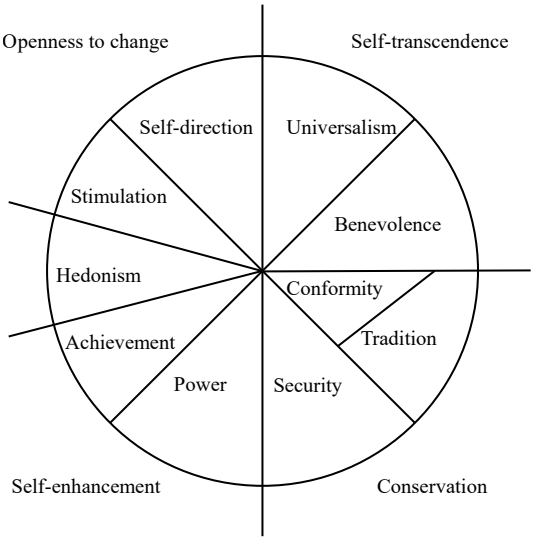


Figure 3.1: The theory of basic human values with the 10 values and four higher dimensions forming a circular structure, based on: [146].

A lot of empirical work used the theory of BHV covering a broad range of topics: theory-testing based on a meta-analysis of 88 studies [163] and theory valida-

tion [150], exploration of differences between groups (e.g. national value priorities [148], sex differences in value priorities from 70 countries [149], value differences among various occupations [101], etc.). The theory has also demonstrated some practical utility in applied contexts such as predicting unethical behaviour [45], explaining associations between leadership styles and organizational outcomes [14], commitment at workplace [25] etc. Since the theory defines abstract, broad concepts, it trades off potential accuracy of predicting context-specific one-time behaviour [41, 45] for applicability to a broad range of behaviors. Some key advantages of the BHV theory make it preferable to various other motivational theories include:

- Existence of a common vocabulary and terminology, minimizing ambiguity.
- Availability of validated, established instruments, which can increase validity, comparability and replicability of findings.
- The comprehensiveness and universality of motivational constructs makes the theory applicable to a broad range of contexts and behaviors.
- The theory proposes goal contents and a process of decision-making parsimoniously.

Additionally, the theory demonstrated better distinctive and predictive capabilities than one of the most widely used models of personality (i.e. Big Five) in an empirical investigation [168], corroborating its practical utility.

3.3 Internet of things and the Smart Grid

Gartner predicted (in 2014) that a total of around 25 billion internet of things (IoT) devices will be in operation by 2020 across industries worldwide and that utility companies will be the top IoT users due to Smart Meter installations [118]. A careful balancing was recommended between the new business opportunities enabled by IoT devices and risks that may arise from misuse or loss of valuable information. It was expected in 2018 that spending on IoT security will reach $1.5 billion in 2018 and $3.1 billion by 2021, since organizations often have no control over several aspects of the IoT value chain (e.g. software and hardware, technical standards, etc.) and regulations for IoT security are just beginning to be developed. It was predicted that regulatory compliance will be the main driver of security spending [11] and the latest prognosis estimated 1.37 billion IoT devices among utilities in 2020 [59].

By the beginning of 2019, around 98% of endpoints have been upgraded with Smart Meters in the Norwegian electricity distribution grid, totalling an investment cost at

NOK 9 billion, for approximately 3 million IoT devices [141]. This represents a first major step toward a SG, which refers to the concept of the traditional electricity grid enhanced by IoT devices at various points in the infrastructure enabling extensive monitoring and controlling capabilities [178]. The evolution of the traditional grid is motivated by several external and internal factors such as meeting electricity demands of the future (e.g. electrification of the transportation sector); compliance with EU and national directives [179] pushing toward more environmentally friendly energy sources (renewables); saving costs by utilizing already existing infrastructures to the maximum extent possible; need to accommodate distributed source of power generation which are often intermittent, etc. While the traditional electric grid was based on the centralized electricity generation paradigm the envisioned SG is characterized by a bidirectional flow of electricity and information. The SG is expected to operate more reliably with reduced operating and maintenance costs; be more flexible and highly autonomous; create new markets and business opportunities [58]. The SG will be a highly complex social-technical system with a critical role in maintaining and supporting other critical societal functions. Key domains of the SG are depicted in Figure 3.2 based on work by the U.S. National Institute of Standards and Technology [63].

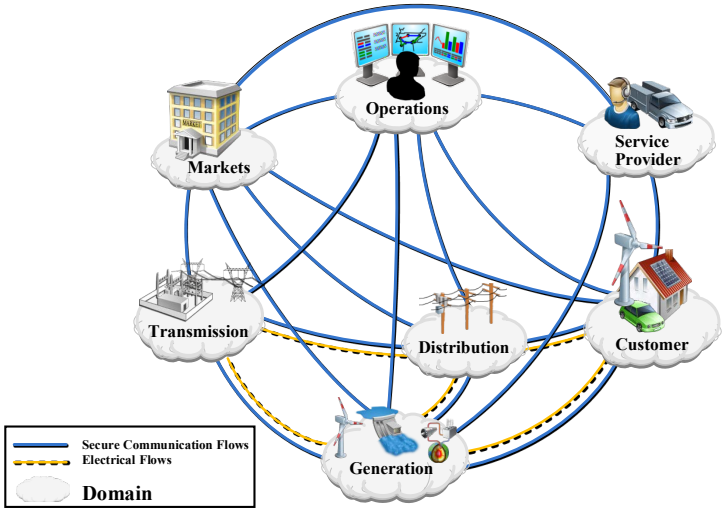


Figure 3.2: Conceptual model of major Smart Grid domains. Based on: NIST [63].

With the introduction of information communication technologies (ICT) at each level of the grid, the security of information (i.e. confidentiality, availability, integrity) will be of key importance to all stakeholders connected (physically or functionally) to the grid. The risks introduced by ICT to the grid, need to be assessed and kept under control to ensure national security. Cyber-attacks on the SG may

become more commonplace [23], and opportunities for illegal moves by individuals can increase [103]. However, due to the high number of stakeholders involved in the development and operation of SGs (e.g. politics, regulation authorities, standardization bodies, service and market providers, aggregators of flexibility, entities responsible for information and data exchange, generation operators, system operators, device manufacturers, software companies, consumers) a more subtle type of problem may become more widespread: risks due to the misalignment of stakeholder incentives. Network convergence [119] represents such a risk arising from the utility companies' motivation to save costs by moving operations to the internet whereas society at large gets exposed to novel threats (and pays the costs in case of an incident). A highly complex system as the SG has various interfaces with other social systems (e.g. regulations, markets, technologies, etc.) each of which may be used to game the system (as an individual trader [44]) or to create more tangible disruptions in societies (e.g. power grid hack in Ukraine [106]). Since the envisioned SG is a highly complex, emergent, dynamic system (lacking historical data about system behaviour) and strategic human decisions have far-reaching impacts on a large number of other stakeholders, it represents a system in need of novel risk analysis approaches. Therefore, the SG represents a particularly suitable case for the CIRA method. Additionally, the SG was the main focus of the IoTSec project which encompassed the research work presented in this thesis along with several other scientific activities [87]. Thus, the requirements of the project assisted in the selection of the use case for CIRA.

3.4 Information security and psychology

Information security is both a technical and a people problem [145]. Human interaction with IT systems can have negative impact on IS in a variety of ways, therefore, IS started to take a multidisciplinary approach and incorporate findings from psychology in order to secure critical information systems from undesirable consequences of human behaviour [5]. The concept of security rests on the assumption that there is at least one motivated individual who has a conscious intention to cause harm for another entity. Therefore, IS is primarily interested in the psychology of the attacker: the person or group intentionally trying to compromise the confidentiality and/or integrity and/or availability of information assets. However, human behaviour may negatively impact the objectives of security in a variety of ways: unintentional errors, non-compliance with security policies due to lack of awareness or goal-conflicts and intentional attackers (i.e. insiders, hackers, crackers, advanced persistent threats, etc.). Verizon's yearly data breach investigations report estimated that 46% of data breaches in 2009 were attributed to insiders [183], whereas in 2019 only 8% of breaches were conducted by insiders, while 45% involved hacking, 22% involved social engineering attacks and 22% of breaches were due to some

kind of errors [184].

According to Schneier "..security is only as good as its weakest link, and people are the weakest link in the chain" cited in [143]. The weakest link concept refers to the idea that users are security operators and are accountable for their actions but often fail to fulfil their role due to various reasons, increasing the organization's exposure to threats. Certain cognitive processes (e.g. heuristics and biases [129]), cognitive limitations and lack of awareness give rise to unintentional errors whereas conscious circumvention of security mechanisms happens when they are perceived as obstacles to the primary task at hand (i.e. goal-conflict) [1]. User-centered design is needed to reduce cognitive load on users (e.g. remembering many secure passwords) and an understanding of the users' perspective is needed during the design phase of security mechanisms to increase their adoption and acceptance among users [187]. A literature review of employee's IS awareness and behaviour showed that the topic has generated a significant research interest resulting in 113 research papers, utilizing a total of 54 different psychological theories (theory of planned behaviour being the most popular) [104]. Improving the organization's culture and safety climate is recommended to improve IS related behaviours within the organization [127] and a value-based compliance model is suggested which may replace the traditional control-based compliance model [74].

Almost all major IS risk analysis methods (e.g. ISO 27005 [85], Risk IT [88], NIST 800-30 [90]) enumerate several personal characteristics (e.g. motivation, traits) which have been historically observed in relation to various threats (insiders, outsiders). Malicious attackers have been descriptively classified based on their motivations (goals or expectations from compromising a system), capabilities (skills, resources), triggers for an attack, methods used, and trends associated with the attacker classes [71]. Since a wide range of activities conducted by insiders can be detrimental to organizations (e.g. espionage, sabotage, stealing of intellectual property etc.) several attempts aim at identifying personal characteristics or behavioural cues that are potentially indicative of an insider threat. Based on case studies of historical insider threats, a wide range of psychological attributes and traits (e.g. introversion, social and personal frustrations, computer dependency, ethical flexibility, entitlement, narcissism, lack of empathy) are listed in [153] to stimulate further research attempts in better understanding and mitigating insider threats. Various solutions have been proposed to detect and potentially prevent insider threats to IS: using methods from criminal profiling [126], extensive monitoring of employee behaviour in the system and combining the data with psychometric tests [97], focusing on observable behavioural cues (e.g. disgruntlement, anger management issues, lack of dependability, etc.) which may be indicating a potential insider threat [64].

3.5 Psychology of risk and situational aspects of decision-making

IS risk analysis methods taking a quantitative approach are interested in two probabilities: “the probability that an action will occur that has the potential to inflict harm on an asset, and the probable loss associated with the harmful event” [91]. The first probability refers to quantifying the probability of a threat exploiting a vulnerability (i.e. an event using the terminology of ISO 27005 [2]), the second refers to quantifying probable loss magnitudes (i.e. consequences in terms of monetary, technical or human impact [85]). When data is not available to derive these probabilities, qualitative methods are used which combine ordinal labels such as "high", "medium", "low" and rely on the subjective judgement of the analyst to calculate the risks. Despite their limitations like range compression, reversed rankings, uninformative ratings, suboptimal resource allocation, etc. qualitative methods are used frequently [28, 27]. It has been suggested that the optimal level of investment into IS controls should not exceed 37% of expected losses assuming a risk-neutral decision-maker and that the costs associated with protecting highly vulnerable information assets become extremely high as the vulnerability of the asset becomes very large [61].

However, humans are rarely risk neutral. Real-world decision-makers’ attitudes toward risk have been shown to follow a four-fold pattern, due to the differences of the value function for gains (concave) and losses (convex) relative to the active reference point of the decision-maker and due the unexpected characteristics of the weighing function which relates decision weights to stated probabilities (with certain biases) [94]. Since small probabilities are over-weighted and high probabilities are under-weighted, real-world decisions-makers often exhibit risk aversion for high probability gains and low probability losses, whereas risk-seeking is observed for low probability gains and high probability losses (in comparison to options with certainty using carefully manipulated payoffs) [94].

Since “security is a trade-off” [144], it is crucial to analyse how risk perception affects decision-making (i.e. what is deemed secure enough?). It has been shown that risk perceptions are influenced by certain characteristics of the hazards, which rarely match with risks based on objective calculations. People tend to overestimate risks which are rare, new and unfamiliar, sudden, affect them personally, spectacular, immediate, intentional or man-made, talked about, etc., whereas risks that are on the opposite ends of these dimensions tend to get underestimated [158, 144]. Rational decision-makers are insensitive to different formulations of the same problem; however, human decision-makers show sensitivity to situational manipulations. When a problem is framed as a gain from the reference point, risk aversion is

the dominant response but, when the identical problem is framed as a loss from the reference point, a preference reversal occurs (i.e. risk-seeking becomes the dominant response) [95].

The theory of bounded rationality [156] not only incorporates major cognitive limitations into its conceptualization of human behaviour, but highlights the importance of environmental/situational influences on decision-making as well: “a great deal can be learned about rational decision-making by taking into account, at the outset, the limitations upon the capacities and complexity of the organism, and by taking account of the fact that the environments to which it must adapt possess properties that permit further simplification of its choice mechanisms... what we call the "environment" will depend upon the "needs", "drives" or "goals" of the organism...” [156].

Traditionally, social psychology claims that situations are the primary determinants of behaviour with convincing demonstrations of situational impacts on behaviour, which induce uniform behaviours across people (irrespective of backgrounds, personalities, etc.). Famous examples are enumerated in [137]: Muzafer Sherif’s experiments regarding the formation and stability of group norms in face of ambiguous and uncertain information (i.e. judgements about perceptual illusions), resolution of inter-group conflict and hostility by introduction of a superordinate goal, Solomon Asch’s experiments about individual’s conformity with the obviously erroneous majority, studies by John Darley and Bibb Latané demonstrating that the presence of other people inhibits individual’s tendency to intervene in emergency situations (bystander effect), Stanley Milgram’ elaborate experiments about obedience to authority [137]. On the other hand, personality psychologists claim that the relative stability of behaviour over contexts and the observed dissimilarities between people are due to the existence of meaningful stable traits [41]. Several solutions have been proposed to resolve the long-standing person-situation debate and opposition such as: interaction [107, 19, 55], reciprocal interaction [40, 84], exploration of within-person variability [47, 48], characterizing situations as affordances by focusing on their objective features [132, 191]. The definition and measurement of situational attributes has important implications for predictions and it also gives rise to non-trivial problems (e.g. frame problem in artificial intelligence [34] concerned with how to determine which pieces of information are relevant/irrelevant in a given situation considering the potentially unlimited number of consequences/situational attributes).

3.6 Unobtrusive profiling

Unobtrusive measures refer to data collected by means which do not rely on direct interaction between analyst and subject. Since unobtrusive measures are non-

reactive, they potentially avoid problems arising from the presence of the researcher or analyst (e.g. socially desirable responses, demand characteristics, change in behaviour due to interaction or observation - Hawthorne effect, etc.), increasing the validity of data and findings [186]. Furthermore, in case of inaccessible, unwilling or adversarial subjects, unobtrusive measures could be the only feasible approach. Unobtrusive data collection methods may rely on found data (e.g. erosion measures), captured data (e.g. simple observation), retrieved data (e.g. running records or personal and episodic records), and data from computer mediated communications [105]. The subject's unawareness about the act of observation is assumed to be a key feature for generating valid data [105], however non-reactivity should also capture the analyst's undesirable effects on the measurement [38, p. 191]. Developments in information and communication technologies provide new opportunities and tools for data collection and at the same time online behaviour represents an emerging, novel domain where human behaviour can be studied. The extent to which online behaviour fulfils the assumptions of non-reactivity is debatable since privacy notifications are assumed to increase awareness about the various measurements taking place. On the other hand, stated (attitudes) and revealed preferences (behaviour) often show a mismatch in case of privacy [13]. Unobtrusive data collection can take several forms determined by the goals of the activity (e.g. theoretical or applied work), sources of information, methods utilized and choice of psychological constructs under investigation. A broad overview about existing combinations follows.

Observation of music preferences can reveal information related to personality traits, political orientation and cognitive ability [133], however analysis of handwriting (by human experts) does not produce valid personality assessments using the Big Five model of personality [33]. Classification of emotions and various behavioural predictions can be achieved by combining machine learning with real-time facial expression monitoring [4]. Psycholinguistic analysis of written texts on social media can be used to build Big Five psychological profiles [60]. Psycholinguistic text analysis on data from Twitter was used to construct psychological profiles (i.e. Big Five, basic human values, human needs) in [62]. A similar approach (i.e. personality profiling using text analysis) was followed to analyse brand preferences, which is relevant from a marketing perspective [190]. Depending on the type of data generated by a certain social media platform (e.g. likes on Facebook), various inferences can be made relating digital traces to personality (Big Five), sexual orientation, political preference [102]. Attitudes toward law enforcement can be constructed from a combination of features available on YouTube's platform for the detection of potential insiders [98], and trait narcissism can be potentially evaluated by identifying influential users from Twitter for the same purpose [99]. The behaviour of Smart Home occupants (i.e. electricity use) can be predicted

from past behavioural data [22]. Video games can be useful sources of information for assessing users' personality (Big Five) [176]. Monitoring of the internal written communications of employees can be used to construct personality profiles (Big Five) for the detection of potential insider threats [17]. A meta-analysis of 38 studies showed that a wide range of psychological or psychosocial constructs (Big Five traits, Dark-Triad traits, well-being, depression, emotional distress, satisfaction with life, intelligence, social satisfaction, personal values, coping style, self-monitoring skills, substance use) can be inferred from digital traces of online behaviour. However, the types of the digital traces used for inferring the psychological constructs moderate the accuracy of the resulting profiles [152]. Ownership of high-status cars showed significant associations with two of the Big Five traits (i.e. low agreeableness and high conscientiousness) showing that offline data sources can also be valuable sources of information [110]. Data collected frequently by modern mobile phones (i.e. logs of calls, messages and accelerometer data) has been also successfully used to derive Big Five personality traits [57].

While unobtrusive measures based on digital behavioural traces may be highly useful for various profiling purposes they rely on the presuppositions that a subject is present on such platforms (i.e. Facebook, Twitter, YouTube, etc.) or uses a certain device, and that the analyst somehow gains access to the public or private digital traces generated by a subject. Since these assumptions may not hold in operational risk analysis settings, unobtrusive measures which do not rely on specific services or technologies can be more feasible in highly constrained environments.

3.7 Summary of chapter

This chapter aimed at presenting the breadth of theories and approaches considering the main objectives of the present work. CIRA is a novel and unique risk analysis method which puts the greatest emphasis on human behavior among existing risk analysis methods. Therefore, in order to enhance its capabilities and to maximize its benefits extensive research is needed. Various stakeholders need to be characterised by their motivational hierarchy using practically useful theories supported by valid and reliable instruments to enable behavior prediction. The theory of BHV has been selected as such a theory demonstrating several desirable properties. Several emerging systems pose great challenges to traditional risk analysis methods and the SG represents one such system which could see significant benefits from a novel risk analysis method. The selection of the SG as an use case is motivated by the broader context in which the present research work was embedded (i.e. IoTSec project). Since CIRA's domain of application (i.e. information security) is often characterized by adversarial and unavailable subjects, the psychological assessment must rely on unobtrusive profiling methods, which are becoming more and more

prevalent in the digital age. A lot of work has been conducted about the perception of risks, demonstrating decision-makers' sensitivity to situational cues. Thus, the exploration of situational aspects of decision-making is motivated by the desire to go beyond and improve upon the existing approaches to behavior prediction.

Chapter 4

Methodology and methods

This chapter provides an overview about the methodological considerations throughout the research project and presents key methods used to address specific research problems. The chapter starts by presenting the key concepts associated with scientific knowledge in general in Section 4.1, then the Design Science Research (DSR) methodology is presented in Section 4.2, which guided the entire research project. The connection between various stages of the project are discussed through the lens of the DSR framework. Next, a more detailed discussion about the specific research methods (including theoretical and practical considerations, methods of data collection and analysis) is provided covering specific stages of the research project in Section 4.3. The chapter concludes with a summary of the methods used throughout the research project in Section 4.4.

4.1 Scientific inquiry

Science is a collective human activity which aims at generating knowledge about the world. The generated knowledge then may be utilized to modify certain aspects of the world. Scientific understanding is generally associated with a striving for objectivity and rationality, a critical stance, systematic application of methods to generate knowledge and to achieve precision and coherence [173]. "A research paradigm is a set of commonly held beliefs and assumptions within a research community about ontological, epistemological, and methodological concerns" [89, p. 167]. Ontology refers to nature of reality (i.e. what entities exist and what is the relation between them). Epistemology refers to ways of generating knowledge about the world and methodology deals with the assumptions and underlying considerations about specific methods which can generate valid knowledge about reality.

According to [89] positivism and interpretivism are the most established research paradigms in information system research. Positivism assumes that reality exists independent of human actions and similarly to natural sciences it strives to explain cause and effect relationships among entities (ontology). For epistemology, positivism assumes that objective knowledge is attainable about the social world, therefore, methodology emphasizes the use of quantitative methods during data collection and analysis. Interpretivism on the other hand, assumes that the social world is not independent of human actions, it is constructed and shaped by humans and is fundamentally subjective (ontology). Positivism assumes that knowledge can be obtained (epistemology) by studying the experiences of people and viewing them as subjects rather than objects. Therefore, the interpretivist methodology emphasizes methods in which researchers take part in a phenomenon and provide in-depth accounts about the phenomenon of interest.

In a somewhat similar fashion, the analysis of the prevalent assumptions in social sciences discussed in [111] provides a subjectivist-objectivist distinction between approaches for generating knowledge. The objectivist approach assumes realism for ontology, positivism for epistemology, a deterministic human nature and favours a nomothetic (suitable for identifying general laws) methodology. The subjectivist approach assumes nominalism (universals or general ideas are mere names without a corresponding reality) for ontology, anti-positivism for epistemology, voluntarism about human nature and idiographic (focusing on the individual and particular) methodology. Research strategies/approaches and studies can be categorized according to a qualitative-quantitative distinction. Table 4.1 provides a summary of the concepts most often associated with qualitative and quantitative approaches.

Table 4.1: Characteristics of qualitative and quantitative strategies to research [173, 82].

	Qualitative	**Quantitative**
Information	Subjective, rich	Objective, narrow
Internal validity	Low	High
Setting	Naturalistic	Artificial
Design	Unstructured	Structured
Realism	High	Low
Construct validity	High	High - Low
Reliability	Low	High
Control	Low	High
Goal	Exploratory	Confirmatory
Quantifiability of phenomenon	Low	High
Method of inquiry	Inductive	Deductive
Sample size	Low	High
Applicability when	Context is important, Uniqueness is relevant	Phenomena are quantifiable Goal is generalizability of results

4.2 Design Science Research

The overall research project was guided by the Design Science Research (DSR) methodology [89, p. 176], which concentrates on the creation and evaluation of purposeful artefacts through an iterative loop called the Design Cycle [78]. The iterative process focuses on the construction and evaluation of artefacts which may be theories, methods, models, instantiation, etc. The nature of the overall problem (i.e. prediction of human behaviour) requires incorporation of research strategies and methods from psychology which is characterized by a methodological plurality. Therefore, a pragmatic approach was taken by considering the requirements and objectives that a solution needs to fulfil and selecting the most appropriate methods for solving the task by considering the strengths and weaknesses associated with various methods.

The Design Cycle receives input from the Relevance Cycle which identifies problems and needs for new solutions in the environment and the Rigor Cycle which provides theories and methods from the scientific knowledge base [77]. Figure 4.1 provides and overview about the three DSR cycles, which establish a connection between the various research activities and the outputs (i.e. artefacts and articles). The Design Cycle is initiated by the Relevance Cycle by identifying problems as well as opportunities for improvements. An additional feature of the Rigor Cycle is that it feeds back its contribution to the knowledge base, adding scientific value to the theories in terms of new applications or new insights.

As the overall goal was to develop a mechanical prediction method by constructing the motivational profiles of inaccessible subjects who are not identical to the people from whom the information is collected, the following characteristics were prioritized: internal validity, reliability, generalizability of results, transferability of findings to other individuals, quantifiability of results and improvements. These objectives are in opposition with the subjectivist approach which focuses on understanding individuals in their particular contexts with the hope of achieving better predictions (similar to clinical approach to prediction). Since evidence shows that clinical approaches to behaviour prediction are prone to unreliability due to the inconsistencies of the expert or analyst, quantitative and more rigorous approaches were utilized where possible. The following section presents the methods used in each research article.

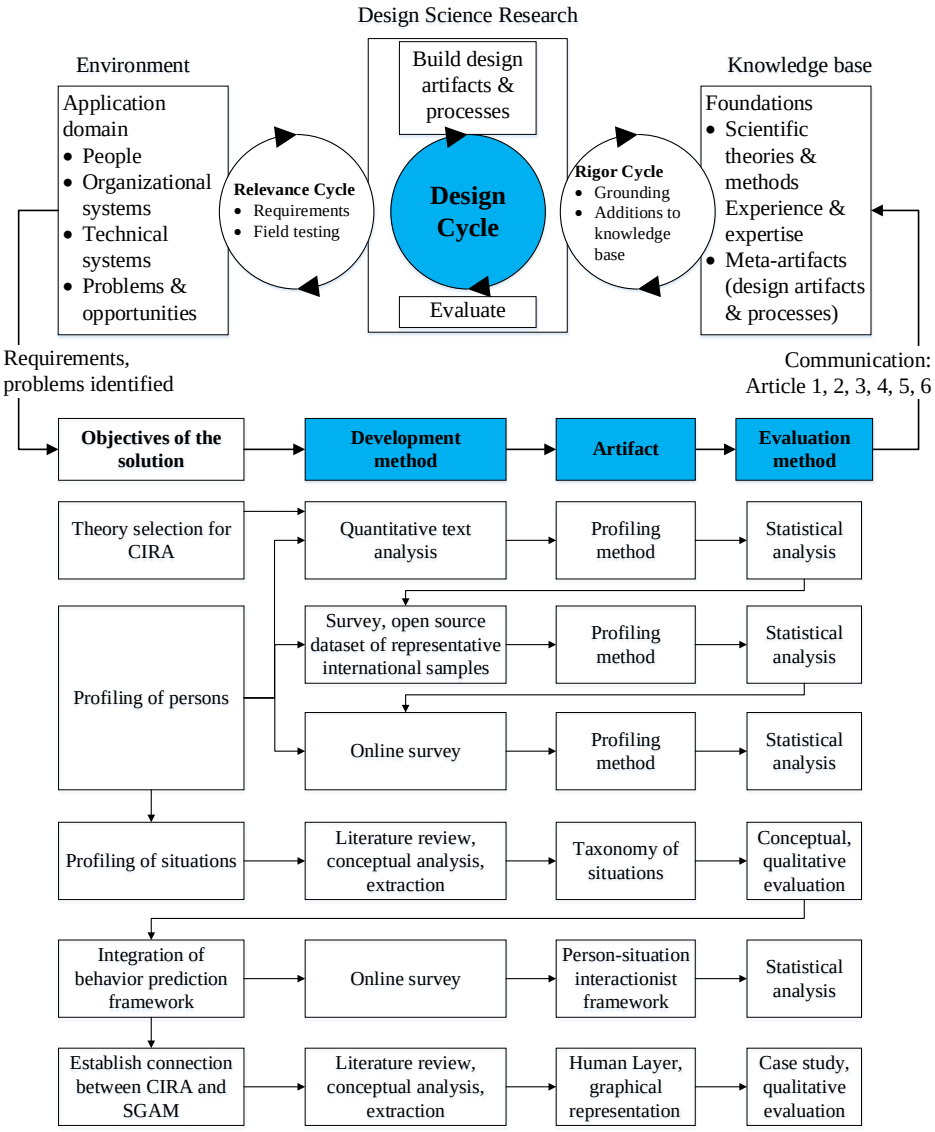


Figure 4.1: Connection between DSR Cycles, research activities and publications. Framework based on [77].

4.3 Applied methods

Article 1: Predicting CEO Misbehaviour from Observables: Comparative Evaluation of Two Major Personality Models [168].

The purpose of this stage was to compare two psychological models capturing individual differences and to assess the feasibility of an unobtrusive data collection method considering inaccessible subjects (for the risk analysis). Identification of the sample was guided by the following considerations: need for a sufficient amount of publicly available data from subjects required for reliable psycholinguistic analysis; utterances had to be spontaneous (i.e. not rehearsed or prepared by others); availability of track record of real-world behaviours with a negative consequence for another entity (e.g. organization) matching the concept of moral hazard, where attribution of responsibility can be assured. Therefore, organizational leaders were selected and a convenience sample (i.e. accessible to researchers) was identified by relying on publicly available datasets (i.e. Wikipedia). The first step of the data collection was done by identifying suitable archival data (i.e. publicly available video recordings available in English, with captions) and ensuring their validity (real-time capturing, monitoring and saving the data). The second step of the data collection used a quantitative psycholinguistic approach provided by the IBM Watson Personality Insights (PI) service to convert the texts produced by the subjects to a vector of psychometric features. Methods used for analysing the data and evaluating the psychological models included descriptive and inferential statistics to evaluate the two models' distinctive and predictive performance with respect to independent research results and real-world behaviors.

Article 2: Using Demographic Features for the Prediction of Basic Human Values Underlying Stakeholder Motivation [172].

The key to achieve generalizability of any research finding is to ensure probabilistic sampling, since "you cannot statistically generalize to a population from which you have not taken a random sample..." [37, p. 195]. Therefore, the European Social Survey (ESS) was identified as a suitable secondary data source which utilizes strict probability sampling in all participating European countries (N=23). A survey is a systematic method for gathering information from a sample of subjects to construct quantitative descriptors of the attributes of the population [67]. Individual-level ground truth profile information (collected by the PVQ-21 instrument) was extracted from the dataset along with predictors matching the criteria identified at the outset of the study (i.e. publicly observable pieces of basic information) for the relevant target population (e.g. members of the working-age population). Data analysis (i.e. evaluation of the proposed profiling method) was conducted by two different analytic techniques (i.e. traditional multiple linear regression and random

forest machine learning) to evaluate the proposed profiling method's performance.

Article 3: Construction of Human Motivational Profiles by Observation for Risk Analysis [167].

Key considerations driving the development of the profiling method was to achieve generalizability of results from sample to population by capturing associations between profile data and region-specific observables (e.g. ownership of items). Therefore, several approaches were explored to ensure that sampling is conducted as close to random sampling from the working age population as possible (e.g. contacting market research agencies, etc.). Due to various constraints, sampling was restricted to employees of the university (NTNU). Call for participation to an online survey was delivered by e-mail to all employees of the organization to ensure each potential respondent has an equal probability to participate. The online survey was available in English and Norwegian translated by a speaker of both languages and proofread by a professional proofreading service. The survey was hosted on internal servers provided by NTNU using the LimeSurvey tool. The survey was completely anonymous, potential subjects had to accept an informed consent in order to start the survey. Relevant ethical [96] and technical guidelines (Norwegian Centre for Research Data [53]) were consulted to ensure compliance with regulations. The data collection was open for a total of 114 days from the beginning of the activity. Several steps were taken to ensure the validity of data (e.g. removal of extreme outliers, removal of respondents with unreasonably fast completion times). Potential biases in the sample were investigated by identifying publicly available population-level statistics. Two instruments were included in the survey for data collection. The Portrait Value Questionnaire (PVQ-21) was used for gathering ground truth motivational profile information. PVQ-21 was designed to be applicable to literate people of most ages and is a 21-item validated psychometric instrument. The PVQ-21 was chosen because it is relatively short, thus reduces respondent burden, which comes at a cost (e.g. reduced reliability which was assessed by the Cronbach's alpha scored measuring internal consistency and reliability). Centered scores were computed to correct for subjective interpretation of the six-point response format. The second instrument was a questionnaire developed to collect information about habits and ownership of items that are observable in any public context. Questions related to habits asked for approximate frequencies of the activities over the past year using a 9-point response format with textual anchors for each option. Other questions used binary, numerical input or open-ended response formats [20]. Data collection produced a total of 225 independent variables which were used to investigate the predictive utility of publicly observable features for constructing motivational profiles. Evaluations used quantitative statistical methods and several experiments were conducted using the data: quantitative comparison to previous

internal results; comparison to independent results; cross-validation; calculation of prediction intervals for characterizing error in case of individual predictions.

Article 4: A Taxonomy of Situations within the Context of Risk Analysis [166].

Since conceptualization and measurement of attributes of situations is a much less mature area than that of persons, a qualitative approach was chosen, which starts with open questions (as opposed to the deductive, quantitative approach focusing on testing a hypothesis). The primary method of data collection was literature survey, which aimed at identifying existing solutions to the problem of conceptualizing and organizing situational aspects of the decision-making process. Existing taxonomies were reviewed and relevant concepts from CIRA and the literature were extracted. Extraction is a qualitative method for data collection which relies on documents, records, or other archival sources [73]. Extraction was followed by conceptual analysis which aims at clarifying the "meaning of concepts, expose conceptual problems in models, reveal unacknowledged assumptions and steps in arguments, and evaluate the consistency of theoretical accounts" [112]. The artefact was internally evaluated (by its developers) using the logical argument and illustrative scenario methods [174].

Article 5: Prediction of threat and opportunity risks: evaluation of a psychological approach using attributes of persons and situations [169].

The key objective of this stage of the research was to assess whether a combined behaviour prediction approach (using attributes of persons and situations together) could improve predictive capabilities. Since the research problem is concerned with a problem at a fundamental level of cognition (i.e. do different people perceive and evaluate the same situations in a similar way?) it was assumed that to address the research problem subjects are not required to possess specialized knowledge or skills. Therefore, the composition and size of the sample was guided by recommendations for the statistical analyses [181]. A caution for relying purely on sample sizes to test a theory is provided by Meehl: "if you have enough cases and your measures are not totally unreliable, the null hypothesis will always be falsified, regardless of the truth of the substantive theory" [117, p. 808]. The first phase of respondent recruitment was conducted online among a random sample of university students who received an invitation to participate in an online survey. Due to low response rates additional subjects were recruited on Amazon's Mechanical Turk online workplace. Several controls were implemented to maximize validity of the final dataset (N = 59). Anonymous data collection was conducted through an online survey hosted on servers of the university implemented in LimeSurvey. The survey contained the English version of the PVQ-21 instrument for collecting ground-truth motivational profile information, and two separate sections with various instructions

accompanying the dilemmas developed in Article 4 [166]. In the first section of the survey, respondents were requested to provide subjective evaluations on the dilemma-options by considering how the selection of each option would impact their overall utility. These scores were used to compute the utility of each dilemma-option. The third section asked respondents to make an explicit choice between the two options of each dilemma. These were used as outcome (dependent) variables in the analyses. The sections were separated by the PVQ-21 and basic demographic questions to maximize tasks between dilemma-related tasks. Methods used for data analysis and performance evaluation included descriptive and inferential statistical methods (logistic regression, intraclass correlation to measure interrater agreements to explore the extent of objectivity in situation assessments).

Article 6: Representing decision-makers in SGAM-H: the Smart Grid Architecture Model Extended with the Human Layer [170].

The objective of this stage of the research was to increase CIRA's applicability to SG use cases. A lack of common understanding about CIRA's usefulness and applicability to the SG eco-system among various stakeholders was identified through the Relevance Cycle in the broader context of the research (i.e. IoTSec project) within the DSR, which motivated the construction of the artefact (i.e. Human Layer, graphical representations) within the Design Cycle. A literature review was performed to identify existing solutions within the scientific literature. The process identified the Smart Grid Architecture Model (SGAM) as a widely used representation of the SG eco-system which has certain deficiencies from CIRA's perspective (e.g. lack of representation of human stakeholders). Thus, the method of data collection was literature review aided by conceptual analysis of existing solutions and extraction of concepts. Development of the artefact followed an iterative process and focused on developing graphical representations of key CIRA concepts compatible with the SGAM model. The evaluation of the artefact was conducted through a hypothetical case study inspired by real-world events which is a qualitative, internal, descriptive evaluation method.

4.4 Summary of chapter

In sum, throughout this research project a pragmatic, mixed method strategy [182] was followed which emphasizes complementarity between different paradigms (e.g. positivism and interpretivism [89]) and suggests that the selection of research methods should be evaluated on the basis of several factors including the goals of the research (e.g. generalizability, validity, reliability, etc. [173, 82]), maturity of the field, quantifiability of the phenomenon of interest. Overall, a quantitative strategy was prioritized, which is alignment with recommendations to improve IS risk analyses by focusing more on quantifiability and measurability of complex phenomena

to improve the quality of subsequent decision-making [80]. The interdisciplinary nature and the complexity of the research goal also required exploration of less well-established areas. Therefore, qualitative methods were used in cases where scientific understanding is far from definitive and in cases where novel artefacts had to be created to solve specific problems.

Chapter 5

Summary of Research Articles

This chapter provides a summary of the six research articles that constitute the thesis and address the main research problems. For each article, the problem statement, methods used, and key results are presented. The chapter concludes with a brief summary of the articles in Section 5.7.

5.1 Article 1: Predicting CEO Misbehaviour from Observables: Comparative Evaluation of Two Major Personality Models [168].

Since CIRA relies on the prediction of a variety of stakeholder behaviours to characterize risks and assumes that stakeholders are inaccessible/non-cooperative with the analyst for traditional psychological assessments, several challenges have to be solved for a behaviour prediction method to fulfil the requirements. This study aimed at achieving two main goals within the context of the research project. First, to compare the discriminative and predictive performance of two well-established, comprehensive psychological theories of personal attributes in order to guide the selection of a suitable theory for further work. Second, to evaluate the feasibility of an unobtrusive data collection method to be used in highly constrained environments with respect to availability of subjects. The main problem statement of this study is as follows: "Can publicly observable variables reflecting individual choice be used to construct psychological profiles suitable for predicting behaviour in the context of risk analysis?" [168]. In order to enable an unbiased comparison among theories, a common sample and a common method was used for creating motivational profiles. The psycholinguistic data analysis was conducted using the IBM Watson PI service. The main hypothesis of the study was that group membership is associated with a

selection bias manifested in the psychological profiles within a sample of CEOs. The sample (N=116) was chosen since these stakeholders operate at the strategic decision-making level within organizations, and have the highest potential impact on other stakeholders. Furthermore, other stakeholder groups may have legal constraints when interacting with the public on behalf of the organization.

A literature survey was conducted for identifying theoretical frameworks which investigate the potential mechanisms producing a selection bias among various groups of professionals. Furthermore, relevant findings were identified which investigate negative consequences for the organization due to the psychological characteristics of the leaders.

Two types of data were used in the study: data produced by subjects (verbal utterances, raw data) and data derived from the IBM Watson PI service by conducting a psycholinguistic analysis on the raw data. The output from Watson PI was subjected to statistical analyses. Two descriptive data analysis methods combined with the various outputs provided by the service were used to describe group-level differences among the sample of CEOs and the general population, and two methods (one descriptive and one predictive) were used to analyse between-group differences among two CEO groups when a historical track record of rule-breaking behaviour was introduced into the models. Both psychological theories were evaluated with respect to their performance for detecting differences among the same sample of subjects.

The most important finding is that the BHV model outperformed the Big Five model across all set of analyses, therefore has better distinctive capabilities to detect a selection bias. Furthermore, it has better predictive capabilities when the same real-world behaviours are used as grouping variables between the two CEO groups. Results from independent studies using representative or near-representative samples were analysed to assess the distinctive performance of both models. Figure 5.1 shows the group-level differences among the CEO sample and a representative sample from the general population based on the BHV profiles. Figure 5.2 shows the group-level profile differences according to the Big Five personality model for the sample of CEOs and the general public.

The predictive performance of both personality models was compared using two logistic regression models which utilized the personal attributes as predictors of behavioral outcomes (i.e. negative externalities inflicted on other stakeholders). Table 5.1 shows the performance of the model using the personal attributes of the BHV model and Table 5.2 shows the evaluation of the Big Five model's performance.

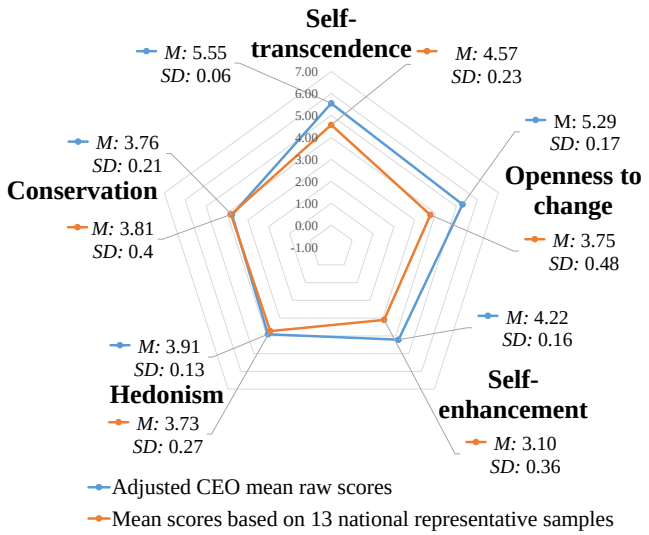


Figure 5.1: Comparison of CEO raw profile scores from the IBM Watson PI service to research results obtained from representative samples for the Basic Human Values profiles.

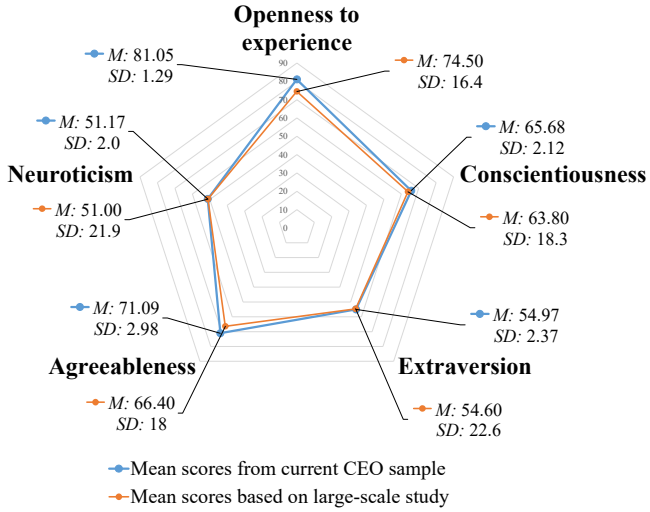


Figure 5.2: Comparison of CEO raw profile scores from the IBM Watson PI service to research results obtained from representative samples for the Big Five personality dimensions.

Based on these findings the theory of BHV was selected as a model of personal

Table 5.1: Logistic regression model using the Basic Human Values profiles.

Predictor	β	*SE* β	Wald's χ^2	*df*	*p*	Odds ratio
Constant	-1.15	0.24	23.68	1	0.00*	0.32
Conservation	-0.50	0.27	3.47	1	0.06	0.61
Openness to change	-0.74	0.29	6.38	1	0.01*	0.48
Hedonism	-0.05	0.29	0.03	1	0.87	0.87
Self-enhancement	0.22	0.32	0.47	1	0.49	1.24
Self-transcendence	-0.24	0.28	0.78	1	0.38	0.78
Test			χ^2	*df*	*p*	
Overall model evaluation			**12.82**	**5**	**0.02***	
Goodness-of-fit-test:						
Hosmer & Lemeshow			12.34	8	0.14	

Note. *p $<$ 0.05. Cox and Snell R^2 = .105. Nagelkerke R^2 = .152.

Table 5.2: Logistic regression model using the Big Five profiles.

Predictor	β	*SE* β	Wald's χ^2	*df*	*p*	Odds ratio
Constant	-1.11	0.23	23.66	1	0.00*	0.33
Openness to experience	-0.09	0.25	0.13	1	0.71	0.91
Conscientiousness	-0.40	0.30	1.75	1	0.19	0.67
Extraversion	-0.61	0.27	5.12	1	0.02*	0.54
Agreeableness	-0.08	0.26	0.09	1	0.77	0.93
Neuroticism	0.70	0.31	4.97	1	0.03*	2.01
Test			χ^2	*df*	*p*	
Overall model evaluation			**10.76**	**5**	**0.06**	
Goodness-of-fit-test:						
Hosmer & Lemeshow			13.65	8	0.09	

Note. *p $<$ 0.05 Cox and Snell R^2 = .089. Nagelkerke R^2 = .129.

attributes to be integrated in CIRA for operationalizing key utility factors. The following papers investigated various approaches for constructing motivational profiles by modifying the assumptions about the availability of relevant pieces of information linked to stakeholders.

5.2 Article 2: Using Demographic Features for the Prediction of Basic Human Values Underlying Stakeholder Motivation [172].

The purpose of this study was to analyse the extent to which publicly available pieces of information can be used for constructing the motivational profiles of inaccessible stakeholders. While the method investigated in the previous study relies on the availability of utterances produced by the subjects, in real risk analysis settings the availability of such information may be limited. Therefore, this study aimed at investigating the utility of the most basic pieces of information (i.e. demographics), which are available in most settings for the construction of the BHV profiles. Existing research work (theoretical an empirical) on the differences between various groups (age, life stages, gender, education, country of birth, occupation) was surveyed to establish the theoretical foundations of the study.

The study utilized the European Social Survey to answer the main research questions. The high-quality dataset was built by using strict probability sampling from 23 European countries. Representativeness ensures that conclusions can generalize to the populations. The dataset was screened for suitable variables, which met the requirements (i.e. ease of observation in any context and accurate assessment by any analyst). A total of 14 features were identified which were used as predictors within the predictive models, where the dependent variables were the ground truth BHV profiles. Two data analytic techniques were used in order to establish the findings: multiple linear regression and distributed random forest machine learning. Experiments with the machine learning approach established that the predictive models are better than random and educated guessing (i.e. guessing the mean) and the multiple linear regression method performed slightly better than the machine learning method. Comparison between the models' performance relied on the R^2 metric, since it was provided by both data analytic procedures. The linear regression method was evaluated as more suitable, since it achieved slightly better performance (possibly due to SPSS's automatic data preparation feature) and because the resulting models are more easily interpretable by humans. Figure 5.3 shows the contribution of each predictor to the final models and the overall variance explained in contrast to the unexplained variance using the linear regression method. Figure 5.4 shows the overall predictive accuracy achievable by the same set of features using the machine learning method.

The findings establish with high reliability the extent of maximal and realistic uncertainty reduction with respect to stakeholder motivational profiles when the strongest restrictions are assumed about the availability of subjects. The findings also establish a solid benchmarking baseline for further investigations.

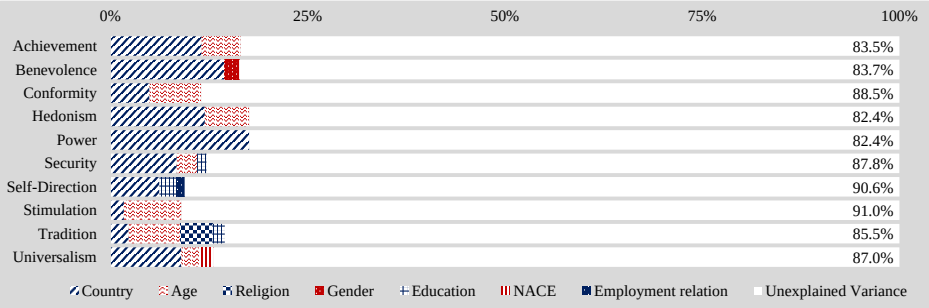


Figure 5.3: Feature importance for predicting the 10 basic human values from observable features by the LR approach relative to unexplained variance expressed in terms of R^2 scores.

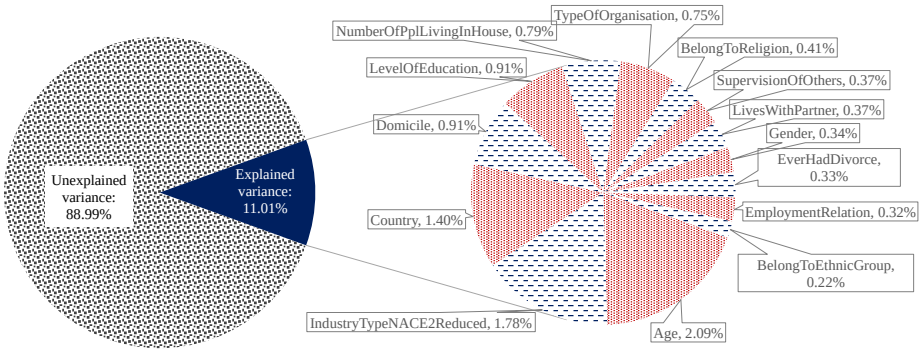


Figure 5.4: Mean feature importance for predicting the 10 basic human values from observable features by ML approach.

5.3 Article 3: Construction of Human Motivational Profiles by Observation for Risk Analysis [167].

While the previous study assumed a highly restricted operational environment for the risk analysis, this study assumes contexts where publicly observable pieces of information representing stakeholders' past and current choices are available for the construction of motivational profiles. Public observables refer to visible evidences about conscious choices from the subject's past available for the analyst (i.e. consumer choices, habits).

The study surveys the literature of the most widely used IS risk assessment methods and provides a discussion about each method's approach to the problem of assessing risks related to human behaviour. It is demonstrated that all methods recognize the importance of focusing on deliberate human actions, but most methods lack solid guidelines and methodology for the assessment of human-related risks. As public observables may have region-specific characteristics (i.e. availability of different products, different consumption habits) it was important to sample from the population which corresponds well with the CIRA method's potential future operational environment's population. Therefore, a call for participation in an online survey was distributed to all active employees of the organization (NTNU) by e-mail, ensuring equal probability for each potential respondent to be included in the study. The survey collected ground truth BHV profile information and information about a variety of habits and consumer choices which are potentially indicative of value trade-offs from the subjects' past.

The utility of public observables for constructing stakeholder motivational profiles was compared to the results obtained from the previous study and significant improvements were detected for all ten attributes. The models' performance was evaluated on unseen data to assess their performance in realistic settings by cross validation. Furthermore, the performance was evaluated by a comparison to the psychometric instrument's test-retest reliabilities obtained from independent studies. In order to represent the uncertainties in individual's predicted profile scores, a model was created to incorporate the prediction interval as an error term with a normal distribution. Figure 5.5 presents the maximal potential improvement achievable from the set of observables collected in the study compared to demographic attributes analysed in the previous study. Figure 5.6 shows the extent to which the current set of public observables can be used as surrogate predictors of the BHV structure of inaccessible stakeholders. No measurement is perfect even when the gold standard instrument is used, which is demonstrated by the length of the grey bars relative to the full scale (i.e. maximum accuracy). However, for all the ten basic human values, observables can reduce uncertainty to at least half of the

original instrument's accuracy.

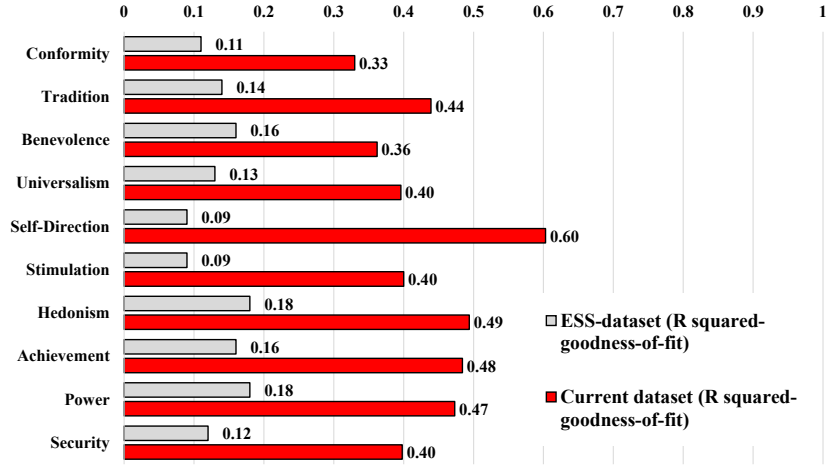


Figure 5.5: Prediction accuracy of Basic Human Values in terms of the R^2 metric. Red bars represent the maximum accuracy achieved after the models were built with the Stepwise feature selection algorithm in SPSS. Grey bars show the goodness of fit metrics for the same variables using demographic features from [172].

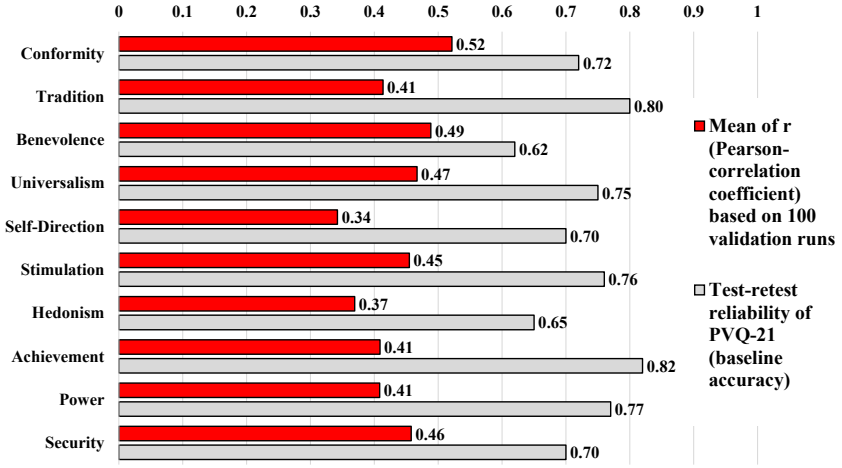


Figure 5.6: Prediction accuracy of Basic Human Values in terms of the Pearson correlation coefficients between predicted and ground-truth scores (red bars). The test-retest reliability is a measure of the correlation between the results of the PVQ-21 taken at different times by the same respondents (grey bars).

5.4 Article 4: A Taxonomy of Situations within the Context of Risk Analysis [166].

Previous studies within the research project focused entirely on individual characteristics of the stakeholder, however the context and immediate situation in which choices are made also need to be taken into account to reach more accurate predictions. Therefore, this study investigated approaches for conceptualizing and organizing situational attributes which have empirically demonstrated influence on decision-makers. First, this study provides an overview about existing and potential approaches to the prediction of human behaviour. It is identified that formalized approaches are more consistent, therefore more reliable than the clinical approach relying on expert judgment, however the clinical approach implicitly takes into account both person- as well as situation-related variables to generate a prediction. Therefore, a formal method which explicitly combines both pieces of information without the inconsistency of the individual would be desirable. To achieve this objective, there is a need to define relevant situational attributes which can be subjected to measurement. The main goal of this study is to identify and systematize such situational attributes.

A literature review is provided on existing attempts to define situational attributes with respect to the behaviour of interest. Existing approaches were categorized along two dimensions: breadth (comprehensive and domain-specific) and approach used for development (theoretical and empirical). The study conceptually analyses the risk concepts of CIRA and maps them to related concepts in psychological studies of decision-making. Dilemma types are mapped to distinct risks identified by CIRA in Figure 5.7. Additional dimensions were extracted from studies of moral decision-making and were used to define the levels of variables associated with each situational dimension. Thus, Dilemma Type, Context, Physical distance (between strategy owner and risk owner), Level of conflict were utilized to develop the classification scheme. The resulting taxonomy of situations is depicted in Figure 5.8. The utility of the taxonomy of situations is demonstrated using illustrative scenarios for classifying an existing dilemma according to the taxonomy's structure and by generating novel dilemmas based on the structure. The main purpose of the taxonomy was to enable the systematic and principled generation of dilemmas which can be used for operationalizing threat and opportunity risks and for testing the enhanced CIRA method's predictive capabilities. A total of 36 dilemmas were developed (i.e. 2 for each leaf node of the taxonomy).

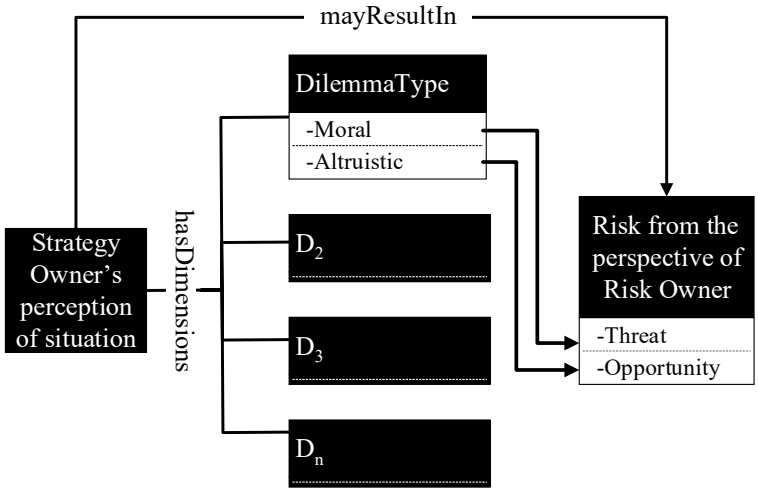


Figure 5.7: Initial conceptual model of the situation taxonomy with mapping of psychological constructs to risk types distinguished by CIRA.

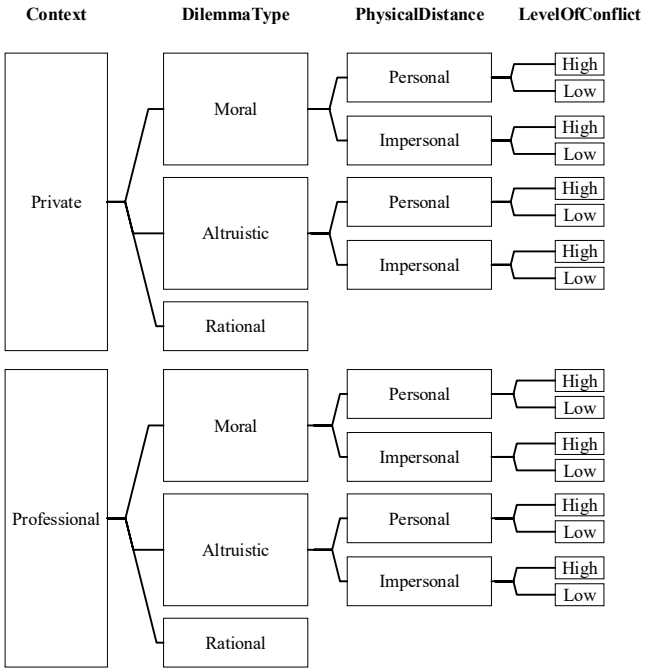


Figure 5.8: Structure of the proposed taxonomy of situations.

5.5 Article 5: Prediction of threat and opportunity risks: evaluation of a psychological approach using attributes of persons and situations [169].

This study aimed at testing the main hypothesis that predictive models, which combine personal and situational attributes can achieve higher prediction accuracy than traditional predictive models relying on personal attributes only. Furthermore, to assess the combined method's expected practical feasibility, potential analyst performance was evaluated by focusing on the analysts' capability of capturing subjects' value trade-offs. The paper surveys the literature of behaviour prediction within IS distinguishing between psychologically and motivationally distinct behavioural categories within scope of CIRA. The proposed approach utilizes the BHV model to capture how actions change the values of the utility factors and how such impacts can be assessed by analysts. The taxonomy of situations developed in the previous study was used to operationalize threat and opportunity risks identified in CIRA which were the key stimuli for research subjects in an online survey. The final sample size consisted of a total of 59 fully completed surveys, which were subjected to several statistical analyses. It was established that the person-situation interactionist predictive method outperforms models relying on personal attributes only. Table 5.3 presents the summary of the analyses comparing the two behaviour prediction approaches. The combined method's performance is superior to the "personal attributes-only" model across almost all dilemmas (with one exception).

Table 5.3: Comparison of the two approaches for predicting identical outcomes.

	% of overall correct classification		% of variance explained (Nagelkerke's R^2)	
	Personal attributes only	Person-situation attributes	Personal attributes only	Person-situation attributes
Dilemma20	86.4	**89.8***	30	**50***
Dilemma22	93.2	91.5	44	41
Dilemma23	86.4	**94.9***	36	**58***
Dilemma26	74.6	**81.4***	24	**50***
Dilemma28	76.3	**91.5***	11	**70***
Dilemma29	89.8	**93.2***	28	**31***
Dilemma32	71.2	**91.5***	29	**76***
Dilemma34	81.4	**86.4***	30	**45***
Dilemma36	91.5	**93.2***	35	**40***
Note. * improvement of predictive accuracy from personal attributes-only model				

However, the practical utility of the method may be negatively impacted by inconsistent human analysts. The relatively low interrater agreements about the value trade-offs demonstrate that situational influences are to a large extent perceived subjectively. Figure 5.9 demonstrates how the strategy owner's choices are determined by the expected changes to the relevant utility factors and how errors influence the analyst's predictions.

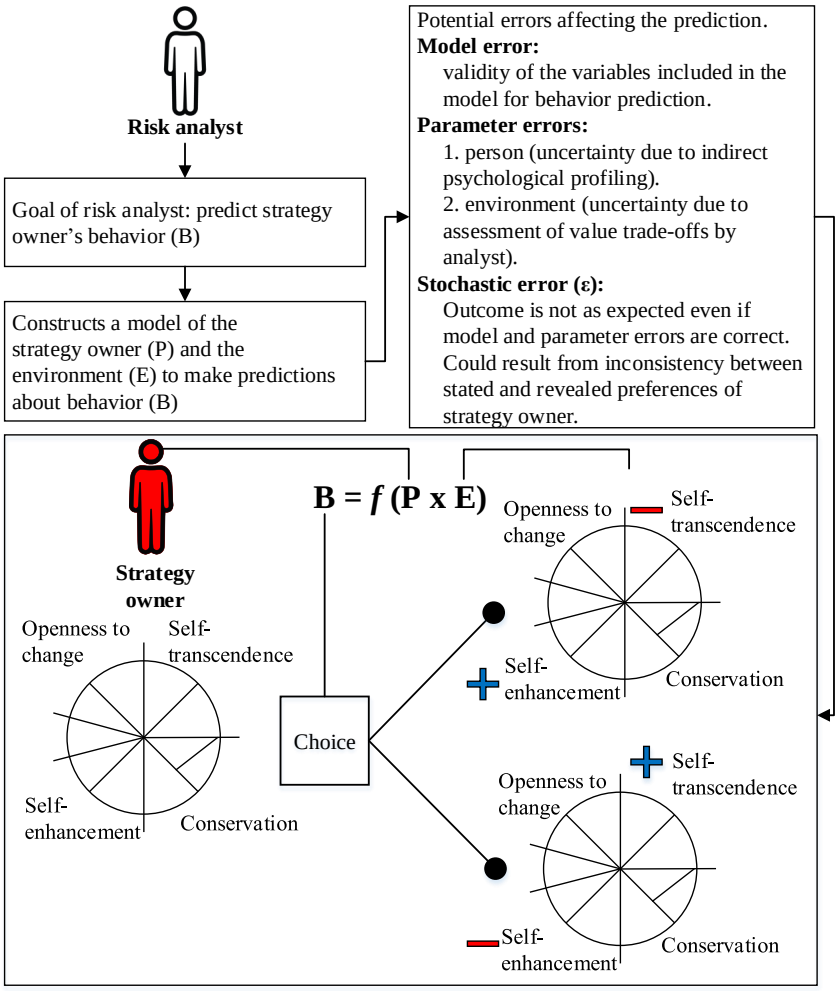


Figure 5.9: Abstraction of the Strategy owner's decision-making process by the risk analyst highlighting three main sources of potential errors (i.e. model error, parameter error and stochastic error according to [185]). Behaviour (B) is assumed to result from the interaction between attributes of the person (P) and attributes of the environment/situation (E) [107].

5.6 Article 6: Representing decision-makers in SGAM-H: the Smart Grid Architecture Model Extended with the Human Layer [170].

This study aimed at establishing a crucial connection between CIRA and the SG ecosystem in order to make the method applicable to SG use cases and to emphasize the importance of focusing on risks related to conscious human decisions. Furthermore, the study incorporates recent developments of CIRA, and demonstrates the details of the methodology to assess the risks attributed to strategy owners. Graphical representations for context establishment and risk communication are provided. The study overviews the literature related to the uses and modifications of the Smart Grid Architecture Model (SGAM), which is a widely utilized model of the SG ecosystem. Next, an overview is provided about approaches for modelling humans in various contexts, highlighting the importance of several design considerations. An artefact is proposed and constructed to establish the connection between CIRA and SGAM: the Human layer with its constituent elements. The construction of the artefact was achieved by the analysis of relevant scientific publications; extraction and visualization of key concepts from CIRA. The artefact is evaluated through a hypothetical case study (inspired by real-world historical incidents) demonstrating the entire risk analysis process at a Distribution System Operator (DSO). Risk treatment options are provided by considering alignment of stakeholder incentives with the societal/organizational goals. Figure 5.10 shows the SGAM extended with the Human layer and Figure 5.11 shows key concepts of CIRA modelled graphically.

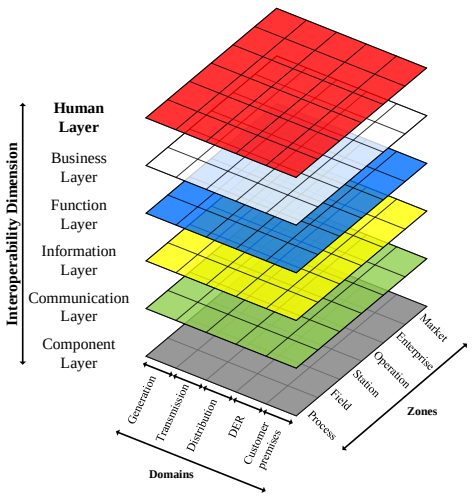


Figure 5.10: The extended SGAM including the Human Layer.

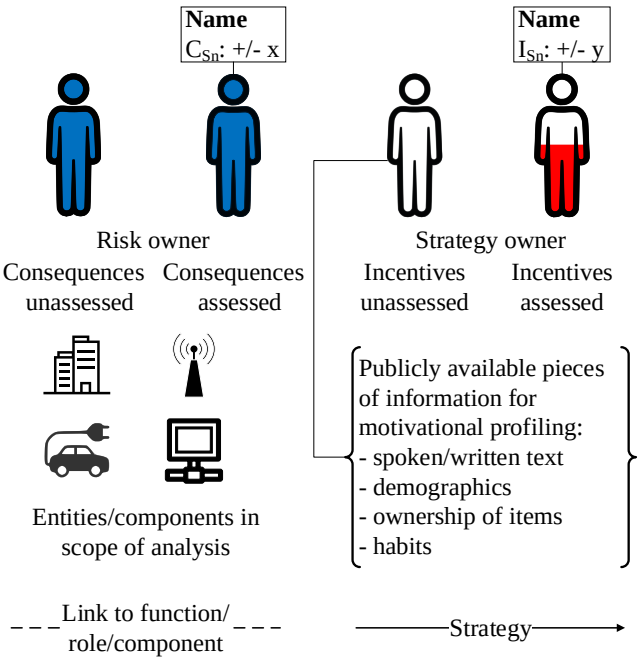


Figure 5.11: Components of the Human Layer.

5.7 Summary of chapter

In sum, this chapter presented the objectives of each research article, the key findings and their relation to the overall research goal of predicting the behaviour of inaccessible stakeholders for the purpose of risk analysis within CIRA. First, two major psychological theories capturing a set of personal attributes have been compared in terms of their performance for detecting a selection bias among a sample of CEOs and in terms of their predictive capabilities when negative organizational outcomes were considered. The theory of BHV outperformed the Big Five model, therefore it was selected for inclusion in CIRA.

Concurrently, an existing commercial service was evaluated for its suitability as an unobtrusive profiling method (i.e. IBM Watson PI). Two additional profiling methods have been developed in order to accommodate various assumptions about the availability of publicly available pieces of information linked to inaccessible subjects. Each method was evaluated thoroughly using several data analytic techniques and independent research results. Thus, a total of three different profiling methods have been explored for inclusion in CIRA to enable the construction of psychological profiles of inaccessible, adversarial subjects. However, several obser-

vations suggested that the psychological profiles in isolation may be insufficient to predict subjects' behavior accurately.

Therefore, the next activities concentrated on defining and conceptualizing attributes of situations, which are deemed as the immediate determinants of behaviors. Thus, a connection has been established between the fundamentally economic risk concepts identified in CIRA and results from the psychology of moral decision-making in order to develop a taxonomy of situations. The taxonomy defines situational attributes which have important implications for the decision-making process, demonstrated by previous empirical results. The taxonomy was used for the principled construction of dilemmas operationalizing threat and opportunity risks, which are central concepts in CIRA.

These dilemmas were then used in another empirical study to assess whether a novel approach to behaviour prediction (i.e. utilizing attributes of persons and situations in combination) can improve over the performance of traditional approaches (i.e. using personal attributes only) model. It has been demonstrated that the predictive capabilities do improve by the inclusion of situational attributes, but the practical utility largely depends on analysts' performance. The involvement of the analyst should be minimized to achieve consistent predictions in real-world settings.

Finally, the connection between CIRA and the SG eco-system was established by the development of the Human layer fulfilling the requirements from the broader scientific context of the research work. The Human layer was designed as an extension to the SGAM to increase CIRA's applicability to SG use cases. Recent developments of CIRA are encompassed in a case study demonstrating the entire risk analysis procedure at a DSO.

Chapter 6

Thesis Contributions

This chapter provides an overview about the contributions of the thesis focusing on the objective of enhancing the behaviour prediction capabilities of CIRA, thus improving its applicability to real-life risk analyses. It is important to distinguish between the concept, object and method of measurement. A measurement is defined as a "quantitatively expressed reduction of uncertainty based on one or more observations" [80, p. 20]. Key contributions are discussed based on the distinction between objects and methods of measurement with respect to people and situations following the main research questions.

6.1 Object of measurement - Selection of a suitable psychological theory for CIRA

Research question 1: Which psychological theory can be integrated into CIRA to enable practically useful characterization of individual stakeholders?

The main goal at this stage was to compare and evaluate candidate objects of measurement (i.e. theoretical constructs capturing features of persons) with respect to their predictive validity which is most relevant for practical purposes [29]. The evaluation guided the choice of psychological model to be integrated into CIRA. Key criteria for potential models included comprehensiveness (i.e. breadth of personal features covered by the models), enabling their applicability for diverse behaviours; stability of constructs within persons over time; demonstrated validity of the constructs as supported by evidence from existing research results; availability of validated instruments which unambiguously operationalize the constructs; availability of datasets to enhance compatibility and comparability of results from

independent sources. With these considerations **Article 1** [168] focused on the evaluation of two major models of personal attributes (BHV and Big Five) with respect to their compatibility with CIRA (i.e. operationalization of psychologically relevant and meaningful utility factors). The benefit of using the Watson PI service was that it enabled overcoming some of the challenges prevalent in the field with respect to evaluating and comparing models (e.g. using different instruments and operationalizations of the constructs; using different contexts and different samples). The common sample, method and identical behavioural outcomes enabled an objective performance comparison of the two models. While the PI service can generate two other models of personal attributes (needs and consumption preferences), they were excluded from evaluation due to non-fulfilment of several criteria. The key contributions related to this stage can be summarized as follows:

1. Narrowing down potential psychological theories to be integrated into CIRA and empirical comparison of two major psychological models of personal attributes for characterizing stakeholders. Overcoming challenges with respect to comparability of findings.

2. Analysis of personal attributes characterizing members of the CEO group, which on the one hand are desirable for the role, and on the other hand, may represent risks to the organization. Description of the psychological characteristics of a potential reference class (CEOs).

3. Selection of the BHV model to be integrated into CIRA, based on theoretical and empirical considerations: better performance compared to the Big Five model in terms of detecting a selection bias among the sample of CEOs of world-leading organizations using comparisons to the Watson PI service's validation dataset; comparisons to empirical results from previous studies. Establishing the BHV model's superior distinctive capabilities (compared to Big Five) for predicting real-world rule breaking behaviour in a retrospective analysis focusing on threat risks (i.e. negative outcomes for the organization).

6.2 Methods of measurement - Construction of stakeholder motivational profiles from publicly available pieces of information

Research question 2: Which unobtrusive data collection methods can be utilized for building stakeholder motivational profiles, taking into account the limited access to subjects during risk analysis?

"Although this may seem a paradox, all exact science is dominated by the idea of approximation..." - Bertrand Russel cited in [80, p. 21]. Most objects of measurement in psychology are theoretical constructs resulting from induction, abstraction, theoretical and empirical considerations, thus several concepts of validity have been developed: construct, content, predictive, concurrent, discriminant, convergent [29]. A consequence of semantically and theoretically meaningful constructs is that measurements are indirect and various inferences are introduced into the measurements. Operationalization (i.e. assigning observables to constructs) is a critical part of the chain of inference.

The objective of this stage of the research was to evaluate various methods of measurements with the goal of constructing the motivational profiles (using the BHV model) of inaccessible and potentially adversarial stakeholders by relying on various pieces of publicly available information. Thus, the key question can be formulated as: what is the validity of various pieces of information linked to stakeholders as surrogate profiling instruments? **Article 1** [168] was based on the assumption that relatively rich pieces of information are available (i.e. written or spoken texts produced by subjects) which can be subjected to psycho-linguistic techniques to construct motivational profiles. However, in operational settings this may be an inaccurate assumption, and in several cases, subjects intended to be included in the analysis had no publicly available textual or verbal utterances. Therefore, **Article 2** [172] was based on a more restricted assumption which corresponds to a scenario with minimal availability of data from subjects. Basic demographic features are available almost in all scenarios and can be easily assessed by analyst, therefore their utility was assessed by using one of the biggest dataset available containing demographic and motivational profile data: the ESS database which contains representative samples from 23 European countries collected by strict probability sampling. This analysis established a baseline for further studies and answered the research question with high reliability about the usefulness of demographic data. Relaxing the assumptions about the availability of publicly observable pieces of information, **Article 3** [167] assumed that evidence of conscious choices (i.e. habits and ownership of items) from the subjects' past and present can improve the accuracy associated with measuring motivational profiles. The key contributions of these articles can be summarized as follows:

1. Demonstration of the feasibility of relying on publicly available pieces of information for motivational profiling by utilizing a commercial service (IBM Watson PI) for the purpose of risk analysis. Analysis of real-world behaviour within a sample of CEOs, whose decisions have the highest organisation-wide impact.

2. Development and statistical evaluation of various profiling methods which minimize analyst subjectivity (i.e. increasing consistency and reliability) when constructing stakeholder motivational profiles using various assumptions about the availability of relevant input data.

3. Comprehensive investigation about the utility of demographic data for motivational profiling. Strong support for the potential maximal utility of demographics due to the high-quality representative international samples found in the publicly available ESS dataset. Converging results obtained from two data analytic techniques (i.e. Multiple Linear Regression and Machine Learning) evaluated by the R^2 performance metric strengthen the findings.

4. Detailed overview provided about the ways in which the most widely used IS risk assessment methods tackle human threats. Several forms of evaluations performed to assess the validity of publicly observable pieces of information representing conscious choices as surrogate predictors of motivational profiles. Illustrative example developed about how the profiles can be utilized to assess action desirability (i.e. prediction of behaviour) by using the utility calculations (i.e. multi-attribute utility theory) established in CIRA.

6.3 Object of measurement - Situational aspects of decision-making

Research question 3: What situational features need to be considered with respect to risk types identified in CIRA?

Considering that behaviour predictions in general have a high level of uncertainty (demonstrated in the literature), this stage of the research investigated the possibility of identifying relevant situational features which (when evaluated and included in predictive models) could potentially improve predictive capabilities of CIRA. Therefore, **Article 4** [166] surveyed the existing literature which aims at organizing situational attributes that exert influence on decision-makers. Furthermore, the study proposed a connection between the risk concepts distinguished in CIRA and existing research results from the field of the psychology of moral and altruistic decision-making. The key contributions related to this stage can be summarized as follows:

1. Mapping CIRA risk types (i.e. threat and opportunity risks) to established psychological constructs by conceptual analysis of similarities between relevant concepts. Specification of situational dimensions which are relevant for risk realization from the strategy owner's perspective.

2. Proposal and evaluation of a domain-specific taxonomy of situations focusing on situational attributes with proven influence on decision-makers. Dimensions and the corresponding levels within the proposed taxonomy are as follows: DilemmaTypes (Moral, Altruistic and Rational); Contexts (Private and Professional); PhysicalDistance (Personal and Impersonal); LevelOfConflict (High and Low).

3. Key utility of the taxonomy is the operationalization of threat and opportunity risks as moral and altruistic dilemmas for the purpose of testing and improving the predictive capabilities of CIRA. Development of 36 dilemmas (two for each leaf-node of the taxonomy) enabled by the systematic and principled manipulation of relevant situational attributes specified in the taxonomy.

6.4 Overall evaluation of predictive capabilities including method of measurement - analyst as instrument

Research question 4: To what extent does a person-situation interactionist framework improve predictive capabilities?

While the preceding stages mainly focused on the construction of motivational profiles of inaccessible subjects, the integration and evaluation of the P-S interactionist framework within CIRA was lacking. Therefore, **Article 5** [169] was dedicated to the task of filling this gap by utilizing the dilemmas developed in the previous stage which served as stimuli for respondents. Respondents were taking the role of the strategy owner, making choices as well as the role of potential analysts assessing relevant value trade-offs elicited by the situations. Previously unexplored issues were investigated to evaluate the predictive capabilities of the model assuming ideal settings (i.e. perceived value trade-offs assessed by subjects), as well as realistic settings (i.e. analyst inference needed to assess value trade-offs objectively). Key contributions of this stage are as follows:

1. A survey of the relevant literature identified that psychological and empirical approaches for behaviour prediction are widely used in the operational context of IS, specifically in connection with insider threats and end-user compliance. However, human threats at the strategic level of decision-making are rarely investigated by empirical methods using psychological attributes of individual decision-makers, who are ultimately responsible for the existence of negative externalities and moral hazard. In order to address this gap and to contribute to CIRA's improvement two approaches to behaviour prediction were compared and evaluated.

2. Evaluation of the P-S interactionist framework's usefulness for predicting choices in a setting which aimed at achieving a high ecological validity (i.e. similarity to real-world events). Evaluation of the P-S interactionist approach to behaviour prediction.

3. It has been established that predictive models which utilize personal and situational attributes simultaneously, provide significantly better predictive capabilities across almost all dilemmas compared to models which rely on personal attributes only.

4. There are clear benefits of using a P-S interactionist framework for behaviour prediction, however its practical utility is largely dependent on analyst performance. Intraclass correlation was used as a measure of inter-rater reliability to assess analyst performance with respect to producing objective value trade-offs. Value trade-offs demonstrate a low degree of agreement among raters, suggesting a low degree of objectivity inherent in situations. Error is introduced into the chain of inference due to inconsistent perceptions about value trade-offs elicited by situations across analysts. The chain of inference is analysed in terms of the other sources of error which are introduced at various stages of the inference: model error, parameter errors (i.e. profiling of people, value trade-off assessments by analyst) and stochastic error.

5. Potential solutions are proposed for overcoming the limitations introduced by analysts in order to make more accurate predictions in real-world risk analyses.

6.5 Enhancement of the Smart Grid Architecture Model

Research Question 5: How to increase CIRA's applicability to Smart Grid scenarios?

A final requirement from the broader context (i.e. IoTSec project [87]) of the research was to increase CIRA's applicability to SG scenarios. It has been observed that the most widely utilized model of the SG ecosystem (i.e. SGAM) lacks a representation of human decision-makers, which may lead to an unilateral focus on technical aspects at the expense of neglecting risks that are attributed to conscious, motivated human stakeholders. Therefore, **Article 6** [170] aimed at establishing the crucial connection between CIRA and the SG, thereby increasing CIRA's applicability to the domain of critical infrastructures. The major contributions of this stage are as follows:

1. Proposal and evaluation of the Human layer, giving rise to SGAM-H: an enhanced version of the SGAM, which maintains a high degree of compatibility with the original model's structure. Constituent elements of the Human layer were created by extracting key concepts from CIRA and by converting them to graphical representations.

2. Presentation of a fully worked out case study applying the CIRA method to analyse intra-organisational risks at a Distribution System Operator. Demonstration of recent developments of the method: key utility factors derived from the Balanced Scorecard (BSC) method; distinction between work-related utility factors derived from relevant key performance indicators (KPIs) and personal utility factors operationalized as basic human values; formula provided for error calculations.

3. The SGAM-H aims at facilitating the construction of a common understanding among professionals involved in the development and risk analysis of SGs about the relevance of motivated human decisions and the risks attributed to them, which is a key initial step towards forming a more complete picture about potential issues affecting emerging critical infrastructures. The SGAM-H model assists with context establishment and risk communication in the risk management procedures.

Chapter 7

Limitations and Future Work

The purpose of this chapter is to identify and discuss the limitations of the research project. Corrective actions are suggested, and some topics are outlined worth considering for future work.

7.1 Methodological limitations

The first limitation relates to sample sizes in Article 1, Article 3 and Article 5. While several steps were taken to increase the samples and each sample meets the minimum requirements for the statistical analyses used, the generalizability of findings from sample to a target population is always restricted, unless respondents are randomly drawn from the population. This limitation impacts many social science studies where convenience samples are used, which may introduce bias into the results. Replication studies with access to probability samples from the relevant populations (e.g. CEOs, operators of critical infrastructures, general population) may overcome this limitation. Article 2, however provides solid evidence about the utility of demographic features for building stakeholder profiles which can be generalized to 23 European countries. Replication studies may extend the analysis to non-European populations to establish more universal findings. The major limitations of Article 4 and Article 6 relate to the internal evaluation of the artefact, which is a weak form of qualitative evaluation. Future studies may use expert evaluations or field experiments to generate more quantitative assessments about the artefacts' real-world utility, which will require development of teaching materials and accessible, cooperative subjects.

The theory of BHV was chosen to be included into CIRA by several considerations: need to predict a variety of motivated behaviors which may result in threat or

opportunity risks; need to be suitable for unobtrusive data collection; need to enable comparisons with independent research results; availability of validated instruments; existing empirical results supporting the theory's usefulness; etc. However, the validated instrument accompanying the theory demonstrated lower reliabilities than expected, which is mainly due to its briefness. This represents a conscious methodological trade-off driven by the motivation to reduce respondent's burden and to increase completion rates. A more reliable instrument could be used in the future, if respondent effort is not a major concern or a proper compensation can be provided for subjects. Another limitation is related to the trade-off between the comprehensiveness of the BHV theory (defining abstract, broad motivational constructs) and its predictive capability for single instances of behaviour. Broad constructs can predict behavioural patterns over time, but their usefulness is more limited for predicting one-time behaviours [41]. Generally, the narrower a trait is, the better predictor it is (given that the trait and outcome are correlated). Furthermore, persons who are more extreme on a trait are generally more predictable using that trait. Finally, more specific situations improve predictive capabilities [41]. Thus, this research aimed at incorporating situational information into the predictive models to increase predictive performance. However, future studies could utilize narrower personality traits as well. This approach would require a more specific definition of behaviours within scope of the prediction. For example, if it hypothesized that the strategy owner's risk-taking behaviour has a direct impact on the risk owner's utility, a narrow psychological theory like sensation-seeking could be used (i.e. assessed without direct interaction and plugged into the predictive model) to improve predictive accuracy. The CIRA method is currently static (i.e. does not take into account passage of time); therefore, the inclusion of the temporal discounting construct [51] could be useful to model the strategy owner's sensitivity to delayed costs and benefits.

It should be noted that excessive attention has been paid to developing methods suitable for profiling individuals, which may be a sign of the fundamental attribution error (i.e. over-emphasising the importance of dispositions and under-emphasizing the role and impact of situational influences when observing other's behaviour, while opposite for explanations of own behaviour) [136]. However, the differences between people are relatively small (in terms of their value hierarchies), indicating that the model is more sensitive to differences between situations when characterizing action-desirability. This issue may be investigated by extensive historical case studies if a sufficiently large number of cases can be collected where the objective features of the situations can be identified, and the profiles of the individuals involved in the incident are available. Thus, a potential research question may be formulated as: did people with significantly different psychological profiles make similar choices in similar situations?

The extent of predictability of behaviours resulting in threat and opportunity risks has been quantitatively explored, but the practical utility of the method largely depends on the analyst's capability to reliably and accurately assess the value trade-offs of the strategy owner. This is a key limitation, which is expected to hamper the consistency and validity of the predictions. However, it is currently a hypothesis and more empirical research is needed. Furthermore, empirical evaluations of the entire CIRA method are lacking thus, future work could focus on addressing this limitation. To enable unbiased evaluations, it would be desirable to perform a large number of real-world case studies at various organizations. In order to quantify the performance improvement from the methods developed in this project, it would be desirable to use several versions of the CIRA method in a comparative fashion. Thus, two or three versions could be utilized in a large number of real-world settings: 1. theoretical framework relying completely on analyst intuition; 2. current version with the predictive upgrades described in this thesis; 3. another version with further improvements reducing analyst involvement in the entire chain of inference.

As it is noted in [56], conflict is a complex emergent phenomenon associated with adaptive and evolutionary mechanisms in which the whole is different than the sum of its parts. Therefore, the work presented in this thesis (using established scientific methods) could only address a small slice of the complex interactions captured by the concept of conflicting incentives. Further investigations are needed to explore how the dissected, isolated parts of the phenomenon interact when re-integrated.

Future work could increase the compatibility between the enhanced CIRA method and other IS risk assessment methods to complement each other in a mutually beneficial way. The methodology established in this thesis for assessing stakeholder motivation and predicting future behaviour could be a useful input to well-known risk assessment methods lacking exact methodology for the assessment of human threats.

7.2 Limitations of knowledge - uncertainty of environment

There are certain limitations which are related to the phenomenon of interest and predictions in general. The purpose of this project was to develop a behaviour prediction method, which can be used to predict a variety of motivationally different behaviours in highly constrained environments assuming adversarial and inaccessible stakeholders. Subject unavailability necessitates the use of unobtrusive profiling methods. To this end, the project aimed at creating a mechanical prediction method, which can minimize subjectivity introduced by the analyst. The mechanical prediction method can utilize information about the person whose behaviour is to be predicted, the situation in which the behaviour takes place or a combination of both pieces of information. The combined approach was chosen due to the

observation that models using personal attributes only, have low performance in general. Additionally, predictions are influenced by different uncertainties than explanations, which partially explains the mismatch between the performance of theoretical explanatory models and applied prediction models. The fundamental issue is related to the uncertainty of defining future states of the world. Even though the benefits of a P-S interactionist predictive model have been demonstrated in a best-case scenario, where situations are sufficiently well-defined, the uncertainties have key implications for the method's practical utility. First, the uncertainty of the environment limits the analyst's capability to define such situations in advance. Furthermore, the analyst operates as an instrument to assess how situations influence the strategy owners' utility factors and the analyst's inconsistency is detrimental to predictive validity. Future studies are needed to explore whether special training can improve this inconsistency.

Further investigations would have to establish the level of predictability using theory agnostic predictive models to distinguish between two types of uncertainties: those that could be reduced by additional knowledge and those that are irreducible even with increased knowledge. This could provide valuable information about the maximum predictability of human behaviour in the context of IS and could provide a rough estimate about the predictive performance which can be expected from traditional explanatory theories. Once certain problems plaguing machine learning methods (e.g. rare event prediction; predictions related to emerging events and dynamics of human behaviour; incomplete, inconsistent, imbalanced, and noisy data [164]) are addressed, more advanced theoretical and practical solutions can be developed. The practical application of the method is also subject to ethical considerations and relevant regulations. The GDPR in the EU requires that an explicit informed consent from subjects should be obtained before processing their personal data. Furthermore, it is required that decisions (whether or not by automated means) that significantly affect people should be explainable [43]. Thus, in order to provide explanations and to correct potential errors, profiling methods should be transparent for human interpretation which is a major challenge for several advanced machine learning methods.

In order to minimize the analyst's involvement in the inference process, (i.e. to increase the validity of predictions), a key assumption must be investigated further: that situations have objective attributes which can be defined and measured (similar to people's attributes). If this turns out to be true, automation of situation assessments could produce reliable results giving rise to more consistent predictions. In order to predict one-time behaviours, which were not observed previously in the subject's past, data would be required from two reference classes: one for persons and another one for situations. Thus, an extensive database of the behaviour of

similar people (i.e. reference class for person) in similar situations (i.e. reference class for situation) would have to be developed. The within-subject approach may be suitable for collecting data about subjects and situations concurrently using a modified (unobtrusive) version of the experience sampling method [31].

Chapter 8

Conclusions

In summary this research project investigated the question: how is it possible to predict human behaviour in an IS context, assuming no direct interaction between the subject and analyst? A potential answer to the question constitutes this thesis. The primary goal was to improve the CIRA method's real-world applicability by integrating theories from psychology to enable the prediction of stakeholders' future behaviour. Furthermore, the research work aimed at increasing the compatibility between CIRA and the SG eco-system. The predictability of human behaviour is a difficult problem which has generated a vast literature of theories and applications. The literature is rife with conflicting views about the topic of human behaviour prediction in terms of the basic assumptions, methodological considerations, and the goals of the endeavour. Predictions represent the toughest test for a theory from the perspective of basic sciences. On the other hand, predictions represent means to an end for applied sciences and for risk management. While opinions on what is considered scientific may differ, the practical utility of predicting individual's behaviour is undoubted. Anyone who is exposed to the decisions of others, would value accurate estimates about the potential consequences. A major lesson learned from the existing literature on approaches to behaviour prediction is that mechanical predictions (e.g. statistical, actuarial, algorithmic) are more accurate across a wide range of domains, than expert judgments. Once these methods are developed, their results are 100% reproducible, thus they can outperform human experts in noisy, low validity environments. Therefore, most research activities within this project aimed at developing a mechanical prediction method by utilizing quantitative approaches to fulfil the requirements of CIRA: produce generalizable results which can be used to predict a wide range of intentional behaviours in highly restricted environments.

Thus, a prediction method was envisioned which takes into account both personal

and situational attributes (i.e. using a P-S interactionist framework), which represents a strategy to improve upon the generally low predictive accuracies associated with traditional methods (i.e. using personal attributes only). The theory of BHVs, supported by several empirical investigations and validated psychometric instruments was integrated into CIRA to operationalize stakeholder utility factors, representing the person side of the P-S interactionist framework. Several unobtrusive profiling methods were developed to construct motivational profiles of inaccessible and potentially adversarial stakeholders. The inclusion of situational attributes (i.e. value trade-offs elicited by situations) showed that the P-S interactionist approach outperforms traditional methods in predicting behaviour. The predictive performance of the P-S interactionist method is 0.51 using Nagelkerke's R^2 and 0.30 using the more conservative Cox and Snell R^2 metrics averaged over all dilemmas. These results are comparable to the performance of other behaviour prediction methods reported in the literature which utilize gold standard psychometric instruments. While the method's performance is very close to what can be reasonably expected about the predictability of human behaviour in other contexts, these results represent best-case scenarios.

The integration of predictive capabilities into CIRA represents a qualitative step from a method which entirely relies on subjective analyst judgments. Even small improvements over the current baseline (i.e. guesswork) have important benefits when a mechanical method reliably outperforms an analyst on the long run in highly constrained and noisy environments. However, the real-world performance of the method could decrease for three reasons: uncertainties associated with unobtrusive profiling; the analyst's epistemic limitations for defining situations that will match with the actual situations faced by the strategy owner; and the analyst's inconsistency for accurately capturing the strategy owner's value trade-offs. To minimize these uncertainties more research effort is needed in the direction of defining and automating situation assessments.

The predictive model is more sensitive to characteristics of a situation (i.e. contribution of an action to the overall utility), than to personal characteristics of the decision-makers (i.e. small differences between subjects w.r.t. their value hierarchies). Thus, even if key stakeholders (e.g. CEOs, heads of state, presidents) share several peculiar psychological attributes, the main reason these individuals are of special interest is because they face special situations due to their positions. That is, the high-impact value trade-offs afforded by the situations. This is relevant both from the perspective of potential risk owners (who are exposed to the consequences of these decisions) and from the perspective of the decision-maker as well: unique positions increase the probability of encountering dilemmas which can significantly alter their overall utility compared to less special positions

(e.g. being approached by offers of $X million to pass a certain law; sacrificing significant resources to ensure re-election or face potential imprisonment; which military leader's friendship to seek, etc...). In sum, the model seems less sensitive to differences between decision-makers than to the differences between situations. Therefore, environmental and situational variables require more scientific attention, and even though highly accurate predictions may not be feasible, modifications of the environment can be a more promising approach to mitigating risks (or making predictions conditional on modifications). The vast number of motivational theories may be more useful for controlling and modifying behaviours in desirable ways than for predicting behaviour due to the different uncertainties associated with explanations and predictions. While logically, predictions would be prerequisites for behaviour modifications, it may be practically more feasible to control some aspects of the environment than it is to predict future behaviour.

To conclude, the research activities revealed that imperfect predictions may be primarily attributed to fundamental uncertainties of the environment. Therefore, the weakest link in security appears to be a lack of knowledge about the exact situations which may come up and require satisficing choices from humans.

Bibliography

[1] Anne Adams and Martina Angela Sasse. 'Users are not the enemy'. In: *Communications of the ACM* 42.12 (1999), pp. 40–46.

[2] Vivek Agrawal. 'Towards the Ontology of ISO/IEC 27005: 2011 Risk Management Standard'. In: *HAISA*. 2016, pp. 101–111.

[3] Vivek Agrawal and Adam Szekeres. 'CIRA Perspective on Risks Within UnRizkNow—A Case Study'. In: *2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*. IEEE. 2017, pp. 121–126.

[4] Sun Joo Ahn et al. 'Using automated facial expression analysis for emotion and behavior prediction'. In: *The Routledge Handbook of Emotions and Mass Media*. Ed. by K. Doveling, C. von Scheve and E.A. Konijn. Routledge international handbooks. Routledge Abingdon, 2010, pp. 349–369.

[5] Ross Anderson and Tyler Moore. 'Information security: where computer science, economics and psychology meet'. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 367.1898 (2009), pp. 2717–2727.

[6] Edmund L. Andrews. 'The Science Behind Cambridge Analytica: Does Psychological Profiling Work?' In: *Stanford Graduate School of Business* (Apr. 2018). [Online; accessed 14. Feb. 2020]. URL: https://www.gsb.stanford.edu/insights/science-behind-cambridge-analytica-does-psychological-profiling-work.

[7] Dan Ariely and George Loewenstein. 'The heat of the moment: The effect of sexual arousal on sexual decision making'. In: *Journal of Behavioral Decision Making* 19.2 (2006), pp. 87–98.

[8] Dan Ariely and Klaus Wertenbroch. 'Procrastination, deadlines, and performance: Self-control by precommitment'. In: *Psychological science* 13.3 (2002), pp. 219–224.

[9] Robert Aunger and Valerie Curtis. 'The anatomy of motivation: An evolutionary - ecological approach'. In: *Biological Theory* 8.1 (2013), pp. 49–63.

[10] James T Austin and Jeffrey B Vancouver. 'Goal constructs in psychology: Structure, process, and content'. In: *Psychological bulletin* 120.3 (1996), p. 338.

[11] Wunmi Bamiduro and van der Rob Meulen. 'Gartner Says Worldwide IoT Security Spending Will Reach $1.5 Billion in 2018'. In: *Gartner* (Mar. 2018). [Online; accessed 7. Jun. 2020]. URL: https://www.gartner.com/en/newsroom/press-releases/2018-03-21-gartner-says-worldwide-iot-security-spending-will-reach-1-point-5-billion-in-2018.

[12] Jonathan Bannet et al. 'Hack-a-vote: Security issues with electronic voting systems'. In: *IEEE Security & Privacy* 2.1 (2014), pp. 32–37.

[13] Bettina Berendt, Oliver Günther and Sarah Spiekermann. 'Privacy in e-commerce: stated preferences vs. actual behavior'. In: *Communications of the ACM* 48.4 (2005), pp. 101–106.

[14] Yair Berson, Shaul Oreg and Taly Dvir. 'CEO values, organizational culture and firm outcomes'. In: *Journal of Organizational Behavior: the International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 29.5 (2008), pp. 615–633.

[15] PJ Bickel and EL Lehmann. 'Frequentist interpretation of probability'. In: *Selected Works of EL Lehmann*. Springer, 2012, pp. 1083–1085.

[16] Cornelis J de Brabander and Rob L Martens. 'Towards a unified theory of task-specific motivation'. In: *Educational Research Review* 11 (2014), pp. 27–44.

[17] Christopher R Brown, Alison Watkins and Frank L Greitzer. 'Predicting insider threat risks through linguistic analysis of electronic communication'. In: *2013 46th Hawaii International Conference on System Sciences*. IEEE. 2013, pp. 1849–1858.

[18] Brian L Burke, Andy Martens and Erik H Faucher. 'Two decades of terror management theory: A meta-analysis of mortality salience research'. In: *Personality and Social Psychology Review* 14.2 (2010), pp. 155–195.

[19] Allan R Buss. 'The Trait-Situation Controversy and the Concept of Interaction'. In: *Personality and Social Psychology Bulletin* 3.2 (1977), pp. 196–201.

[20] James Carifio and Rocco J Perla. 'Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes'. In: *Journal of social sciences* 3.3 (2007), pp. 106–116.

[21] Douglas F Cellar et al. 'Trait goal orientation, self-regulation, and performance: A meta-analysis'. In: *Journal of Business and Psychology* 26.4 (2011), pp. 467–483.

[22] Sungjoon Choi, Eunwoo Kim and Songhwai Oh. 'Human behavior prediction for smart homes using deep learning'. In: *2013 IEEE RO-MAN*. IEEE. 2013, pp. 173–179.

[23] Kim-Kwang Raymond Choo. 'The cyber threat landscape: Challenges and future research directions'. In: *Computers & security* 30.8 (2011), pp. 719–731.

[24] Ada S Chulef, Stephen J Read and David A Walsh. 'A hierarchical taxonomy of human goals'. In: *Motivation and Emotion* 25.3 (2001), pp. 191–232.

[25] Patricia Cohen et al. 'The problem of units and the circumstance for POMP'. In: *Multivariate behavioral research* 34.3 (1999), pp. 315–346.

[26] Christopher J Collins, Paul J Hanges and Edwin A Locke. 'The relationship of achievement motivation to entrepreneurial behavior: A meta-analysis'. In: *Human performance* 17.1 (2004), pp. 95–117.

[27] Louis Anthony (Tony) Cox Jr. 'What's wrong with risk matrices?' In: *Risk Analysis: An International Journal* 28.2 (2008), pp. 497–512.

[28] Louis Anthony (Tony) Cox Jr, Djangir Babayev and William Huber. 'Some limitations of qualitative risk rating systems'. In: *Risk Analysis: An International Journal* 25.3 (2005), pp. 651–662.

[29] Lee J Cronbach and Paul E Meehl. 'Construct validity in psychological tests'. In: *Psychological bulletin* 52.4 (1955), p. 281.

[30] Catharine P Cross, De-Laine M Cyrenne and Gillian R Brown. 'Sex differences in sensation-seeking: A meta-analysis'. In: *Scientific reports* 3.1 (2013), pp. 1–5.

[31] Mihaly Csikszentmihalyi and Reed Larson. 'Validity and reliability of the experience-sampling method'. In: *Flow and the foundations of positive psychology*. Springer, 2014, pp. 35–54.

[32] Robyn M Dawes, David Faust and Paul E Meehl. 'Statistical prediction versus clinical prediction: Improving what works'. In: *A handbook for data analysis in the behavioral sciences: Methodological issues* (1993), pp. 351–367.

[33] Carla Dazzi and Luigi Pedrabissi. 'Graphology and personality: an empirical study on validity of handwriting analysis'. In: *Psychological reports* 105.3_suppl (2009), pp. 1255–1268.

[34] Daniel Dennett. 'Cognitive wheels: The frame problem of AI'. In: *Minds, machines and evolution: philosophical studies*. Ed. by Christopher Hookway. Cambridge University Press, 1984, pp. 129–150.

[35] Serkan Dinçer. 'The effects of materials based on ARCS Model on motivation: A meta-analysis'. In: *Ilkogretim Online - Elementary Education Online* 19.2 (2020), pp. 1016–1042.

[36] Renee DiResta et al. *The tactics & tropes of the Internet Research Agency*. [Online; accessed 8. Jul. 2020]. 2019. URL: https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1003&context=senatedocs.

[37] Eugene S Edgington. 'Statistical inference from N= 1 experiments'. In: *The journal of psychology* 65.2 (1967), pp. 195–199.

[38] Michael Ed Eid and Ed Ed Diener. *Handbook of multimethod measurement in psychology*. American Psychological Association, 2006.

[39] Hillel J Einhorn. 'Accepting error to make less error'. In: *Journal of personality assessment* 50.3 (1986), pp. 387–395.

[40] Robert A Emmons, ED Diener and Randy J Larsen. 'Choice of situations and congruence models of interactionism'. In: *Personality and individual differences* 6.6 (1985), pp. 693–702.

[41] Seymour Epstein. 'The stability of behavior: I. On predicting most of the people much of the time'. In: *Journal of personality and social psychology* 37.7 (1979), p. 1097.

[42] Tracy Epton et al. 'The impact of self-affirmation on health-behavior change: A meta-analysis'. In: *Health Psychology* 34.3 (2015), p. 187.

[43] European Parliament, Council of the European Union. 'Regulation (EU) 2016/679 of the European Parliament and of the Council (GDPR)'. In: *Official Journal of the European Union* (2016). [Online; accessed 15. Apr. 2020]. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679#d1e40-1-1.

[44] Christian-Oliver Ewald et al. 'Riding the Nordic German Power-Spread: The Einar Aas Experiment'. In: *SSRN 3557286* (2020).

[45] Gilad Feldman et al. 'The motivation and inhibition of breaking the rules: Personal values structures predict unethicality'. In: *Journal of Research in Personality* 59 (2015), pp. 69–80.

[46] Baruch Fischhoff and Don MacGregor. 'Subjective confidence in forecasts'. In: *Journal of Forecasting* 1.2 (1982), pp. 155–172.

[47] William Fleeson. 'Moving personality beyond the person-situation debate: The challenge and the opportunity of within-person variability'. In: *Current Directions in Psychological Science* 13.2 (2004), pp. 83–87.

[48] William Fleeson. 'Toward a structure-and process-integrated view of personality: Traits as density distributions of states'. In: *Journal of personality and social psychology* 80.6 (2001), p. 1011.

[49] Donna L Floyd, Steven Prentice-Dunn and Ronald W Rogers. 'A meta-analysis of research on protection motivation theory'. In: *Journal of applied social psychology* 30.2 (2000), pp. 407–429.

[50] David L Forbes. 'Toward a unified model of human motivation'. In: *Review of general psychology* 15.2 (2011), p. 85.

[51] Shane Frederick, George Loewenstein and Ted O'donoghue. 'Time discounting and time preference: A critical review'. In: *Journal of economic literature* 40.2 (2002), pp. 351–401.

[52] Norman Frederiksen. 'Toward a taxonomy of situations'. In: *American Psychologist* 27.2 (1972), p. 114.

[53] *Frequently asked questions*. [Online; accessed 14. Jun. 2020]. June 2020. URL: https://nsd.no/personvernombud/en/help/faq.html?id=2.

[54] Batya Friedman and Helen Nissenbaum. 'Bias in computer systems'. In: *ACM Transactions on Information Systems (TOIS)* 14.3 (1996), pp. 330–347.

[55] R Michael Furr and David C Funder. 'Persons, situations, and person-situation interactions'. In: *Handbook of personality: Theory and research* (2018), pp. 1–42.

[56] Giorgio Gallo. 'Conflict theory, complexity and systems approach'. In: *Systems Research and Behavioral Science* 30.2 (2013), pp. 156–175.

[57] Nan Gao, Wei Shao and Flora D Salim. 'Predicting Personality Traits From Physical Activity Intensity'. In: *Computer* 52.7 (2019), pp. 47–56.

[58] Vincenzo Giordano and Gianluca Fulli. 'A business case for Smart Grid technologies: A systemic perspective'. In: *Energy Policy* 40 (2012), pp. 252–259.

[59] Laurence Goasduff. *Gartner Says 5.8 Billion Enterprise and Automotive IoT Endpoints Will Be in Use in 2020*. [Online; accessed 7. Jun. 2020]. Aug. 2019. URL: https://www.gartner.com/en/newsroom/press-releases/2019-08-29-gartner-says-5-8-billion-enterprise-and-automotive-io.

[60] Jennifer Golbeck et al. 'Predicting personality from twitter'. In: *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE. 2011, pp. 149–156.

[61] Lawrence A Gordon and Martin P Loeb. 'The economics of information security investment'. In: *ACM Transactions on Information and System Security (TISSEC)* 5.4 (2002), pp. 438–457.

[62] Liang Gou, Michelle X Zhou and Huahai Yang. 'KnowMe and ShareMe: understanding automatically discovered personality traits from social media and user sharing preferences'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2014, pp. 955–964.

[63] Christopher Greer et al. *NIST framework and roadmap for smart grid interoperability standards, release 3.0*. Tech. rep. National Institute of Standards and Technology, 2014.

[64] Frank L Greitzer et al. *Identifying at-risk employees: A behavioral model for predicting potential insider threats*. Technical note PNNL-19665. Richland, WA: Pacific Northwest National Lab.(PNNL), 2010.

[65] William M Grove and Paul E Meehl. 'Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy'. In: *Psychology, public policy, and law* 2.2 (1996), p. 293.

[66] William M Grove et al. 'Clinical versus mechanical prediction: a meta-analysis'. In: *Psychological assessment* 12.1 (2000), p. 19.

[67] Robert M Groves et al. *Survey methodology*. Vol. 561. John Wiley & Sons, 2011.

[68] Terry M Gudaitis. 'The missing link in information security: Three dimensional profiling'. In: *CyberPsychology & Behavior* 1.4 (1998), pp. 321–340.

[69] Martin Hagger, Nikos Chatzisarantis and Stuart Biddle. ‘A meta-analytic review of the theories of reasoned action and planned behavior in physical activity: Predictive validity and the contribution of additional variables’. In: *Journal of sport & exercise psychology* (2002).

[70] Alan Hájek. ‘The reference class problem is your problem too’. In: *Synthese* 156.3 (2007), pp. 563–585.

[71] Sara LN Hald and Jens M Pedersen. ‘An updated taxonomy for characterizing hackers according to their threat properties’. In: *Advanced Communication Technology (ICACT), 2012 14th International Conference on*. IEEE. 2012, pp. 81–86.

[72] Ellen L Hamaker. ‘Why researchers should think "within-person": A paradigmatic rationale’. In: *Handbook of research methods for studying daily life*. Ed. by Matthias R. Mehl and Tamlin S. Conner. The Guilford Press, 2012, pp. 43–61.

[73] Margaret C Harrell and Melissa A Bradley. *Data collection methods. Semi-structured interviews and focus groups*. Tech. rep. Rand National Defense Research Inst santa monica ca, 2009.

[74] Karin Hedström et al. ‘Value conflicts for information security management’. In: *The Journal of Strategic Information Systems* 20.4 (2011), pp. 373–384.

[75] Olaf Helmer and Nicholas Rescher. ‘On the epistemology of the inexact sciences’. In: *Management science* 6.1 (1959), pp. 25–52.

[76] Richards J Heuer. *Psychology of intelligence analysis*. Center for the Study of Intelligence, 1999.

[77] Alan R Hevner. ‘A three cycle view of design science research’. In: *Scandinavian journal of information systems* 19.2 (2007), p. 4.

[78] Alan R Hevner et al. ‘Design science in information systems research’. In: *Management Information Systems Quarterly* 28.1 (2004), pp. 75–106.

[79] Robert Hogan. ‘Much ado about nothing: The person–situation debate’. In: *Journal of Research in Personality* 43.2 (2009), p. 249.

[80] Douglas W Hubbard and Richard Seiersen. *How to measure anything in cybersecurity risk*. John Wiley & Sons, 2016.

[81] Eva Hudlicka et al. ‘Predicting Group Behavior from Profiles and Stereotypes’. In: *Proceedings of the 13th BRIMS Conference*. 2004, pp. 362–373.

[82] Coolican Hugh. *Introduction to research methods and Statistics in Psychology*. 1995.

[83] ICA. *Assessing Russian Activities and Intentions in Recent US Elections*. [Online; accessed 8. Jul. 2020]. 2017. URL: https://www.dni.gov/files/documents/ICA_2017_01.pdf.

[84] William Ickes, Mark Snyder and Stella Garcia. 'Personality influences on the choice of situations'. In: *Handbook of personality psychology*. Elsevier, 1997, pp. 165–195.

[85] International Organization for Standardization. 'ISO 27005:2011'. In: *Information technology - Security techniques - Information security risk management* (2011).

[86] Interpol. 'Cybercrime - Future-oriented policing projects'. In: (2017). URL: https://www.interpol.int/content/download/5267/file/Cybercrime.pdf.

[87] IoTSec. *Security in IoT for Smart Grids (IoTSec)*. [Online; accessed 30. Oct. 2020]. Sept. 2020. URL: https://its-wiki.no/wiki/IoTSec:Home.

[88] ISACA. *The Risk IT Framework*. Risk IT. ISACA, 2009. URL: https://books.google.no/books?id=tG7VMihmwtsC.

[89] Paul Johannesson and Erik Perjons. *An introduction to design science*. Springer, 2014.

[90] Joint Task Force Transformation Initiative. *Guide for conducting risk assessments, NIST 800-30*. Tech. rep. National Institute of Standards and Technology, 2012.

[91] Jack Jones. *Factor analysis of information risk*. US Patent App. 10/912,863. Mar. 2005.

[92] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[93] Daniel Kahneman and Amos Tversky. 'On the psychology of prediction'. In: *Psychological review* 80.4 (1973), p. 237.

[94] Daniel Kahneman and Amos Tversky. 'Prospect theory: An analysis of decision under risk'. In: *Econometrica: Journal of the Econometric Society* (1979), pp. 263–291.

[95] Daniel Kahneman and Amos Tversky. 'Rational choice and the framing of decisions'. In: *Journal of business* 59.4 (1986), pp. 251–278.

[96] Ragnvald Kalleberg et al. *Guidelines for research ethics in the social sciences, law and the humanities*. 2006.

[97] Miltiadis Kandias et al. 'An insider threat prediction model'. In: *International Conference on Trust, Privacy and Security in Digital Business*. Springer. 2010, pp. 26–37.

[98] Miltiadis Kandias et al. 'Can we trust this user? Predicting insider's attitude via YouTube usage profiling'. In: *2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing*. IEEE. 2013, pp. 347–354.

[99] Miltiadis Kandias et al. 'Insiders trapped in the mirror reveal themselves in social media'. In: *International Conference on Network and System Security*. Springer. 2013, pp. 220–235.

[100] Rogier Kievit et al. 'Simpson's paradox in psychological science: a practical guide'. In: *Frontiers in psychology* 4 (2013), p. 513.

[101] Ariel Knafo and Lilach Sagiv. 'Values and work environment: Mapping 32 occupations'. In: *European Journal of Psychology of Education* 19.3 (2004), pp. 255–273.

[102] Michal Kosinski, David Stillwell and Thore Graepel. 'Private traits and attributes are predictable from digital records of human behavior'. In: *Proceedings of the National Academy of Sciences* 110.15 (2013), pp. 5802–5805.

[103] Brian Krebs. 'FBI: Smart Meter Hacks Likely to Spread'. In: (2012). [Online; accessed 26-June-2018]. URL: https://krebsonsecurity.com/2012/04/fbi-smart-meter-hacks-likely-to-spread.

[104] Benedikt Lebek et al. 'Information security awareness and behavior: a theory-based literature review'. In: *Management Research Review* (2014).

[105] Raymond M. Lee. *Unobtrusive Methods in Social Research*. Understanding social research. Open University Press, 2000.

[106] Robert Lee, Michael Assante and Tim Conway. 'Analysis of the cyber attack on the Ukrainian power grid, Defense Use Case'. In: *Electricity Information Sharing and Analysis Center (E-ISAC)* 388 (2016).

[107] Kurt Lewin. *Principles of topological psychology*. McGraw-Hill, 1936.

[108] Marc Lochbaum and Jarrett Gottardy. 'A meta-analytic review of the approach-avoidance achievement goals and performance relationships in the sport psychology literature'. In: *Journal of Sport and Health Science* 4.2 (2015), pp. 164–173.

[109] Edwin A Locke and Gary P Latham. 'New directions in goal-setting theory'. In: *Current directions in psychological science* 15.5 (2006), pp. 265–268.

[110] Jan Erik Lönnqvist, Ville-Juhani Ilmarinen and Sointu Leikas. 'Not only assholes drive Mercedes. Besides disagreeable men, also conscientious people drive high-status cars'. In: *International Journal of Psychology* (2019).

[111] Fred Luthans and Tim RV Davis. *Idiographic versus Nomothetic Approaches to Research in Organizations*. Tech. rep. University of Nebraska, 1981.

[112] Armando Machado and Francisco J Silva. 'Toward a richer view of the scientific method: The role of conceptual analysis'. In: *American Psychologist* 62.7 (2007), pp. 671–681.

[113] Charles F Manski. 'The use of intentions data to predict behavior: A best-case analysis'. In: *Journal of the American Statistical Association* 85.412 (1990), pp. 934–940.

[114] Simon J Marshall and Stuart JH Biddle. 'The transtheoretical model of behavior change: a meta-analysis of applications to physical activity and exercise'. In: *Annals of behavioral medicine* 23.4 (2001), pp. 229–246.

[115] Tyler B Mason et al. 'Self-discrepancy theory as a transdiagnostic framework: A meta-analysis of self-discrepancy and psychopathology'. In: *Psychological bulletin* 145.4 (2019), p. 372.

[116] Paul E Meehl. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press, 1954.

[117] Paul E Meehl. 'Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology'. In: *Journal of consulting and clinical Psychology* 46.4 (1978), pp. 806–834.

[118] van der Rob Meulen and Janessa Rivera. *Gartner Says 4.9 Billion Connected "Things" Will Be in Use in 2015*. [Online; accessed 7. Jun. 2020]. Nov. 2014. URL: https://www.gartner.com/en/newsroom/press-releases/2014-11-11-gartner-says-nearly-5-billion-connected-things-will-be-in-use-in-2015.

[119] Tyler Moore. 'The economics of cybersecurity: Principles and policy options'. In: *International Journal of Critical Infrastructure Protection* 3.3-4 (2010), pp. 103–117.

[120] *motivation - Wiktionary*. [Online; accessed 15. Jun. 2020]. June 2020. URL: https://en.wiktionary.org/wiki/motivation.

[121] *Netflix Prize: Forum / Grand Prize awarded to team BellKor's Pragmatic Chaos*. [Online; accessed 9. Jun. 2020]. Sept. 2009. URL: https://web.archive.org/web/20090924184639/http://www.netflixprize.com/community/viewtopic.php?id=1537.

[122] *Netflix Prize: Review Rules*. [Online; accessed 9. Jun. 2020]. Oct. 2006. URL: https://web.archive.org/web/20100106185508/http://www.netflixprize.com//rules.

[123] Mitchell J Neubert. 'The value of feedback and goal setting over goal setting alone and potential moderators of this effect: A meta-analysis'. In: *Human Performance* 11.4 (1998), pp. 321–335.

[124] Johan YY Ng et al. 'Self-determination theory applied to health contexts: A meta-analysis'. In: *Perspectives on Psychological Science* 7.4 (2012), pp. 325–340.

[125] Thomas WH Ng, Kelly L Sorensen and Lillian T Eby. 'Locus of control at work: a meta-analysis'. In: *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 27.8 (2006), pp. 1057–1087.

[126] Nick Nykodym, Robert Taylor and Julia Vilela. 'Criminal profiling and insider cyber crime'. In: *Computer Law & Security Review* 21.5 (2005), pp. 408–414.

[127] Kathryn Parsons et al. *Human factors and information security: individual, culture and security environment*. Tech. rep. Defence Science and technology organisation Edinburgh (Australia), 2010.

[128] Bryan Pfaffenberger. 'The rhetoric of dread: Fear, uncertainty, and doubt (FUD) in information technology marketing'. In: *Knowledge, Technology & Policy* 13.3 (2000), pp. 78–92.

[129] Shari Lawrence Pfleeger and Deanna D Caputo. 'Leveraging behavioral science to mitigate cyber security risk'. In: *Computers & security* 31.4 (2012), pp. 597–611.

[130] Lisa Rajbhandari. 'Risk analysis using 'conflicting incentives' as an alternative notion of risk'. PhD thesis. Gjøvik, Norway: Gjøvik University College, 2013.

[131] Lisa Rajbhandari and Einar Snekkenes. 'Using the Conflicting Incentives Risk Analysis Method'. In: *Security and Privacy Protection in Information Processing Systems*. Ed. by Lech J. Janczewski, Henry B. Wolfe and Sujeet Shenoi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 315–329.

[132] Harry T Reis. 'Reinvigorating the concept of situation in social psychology'. In: *Personality and Social Psychology Review* 12.4 (2008), pp. 311–329.

[133] Peter J Rentfrow and Samuel D Gosling. 'The do re mi's of everyday life: The structure and personality correlates of music preferences'. In: *Journal of personality and social psychology* 84.6 (2003), p. 1236.

[134] Francesco Ricci, Lior Rokach and Bracha Shapira. 'Introduction to recommender systems handbook'. In: *Recommender systems handbook*. Springer, 2011, pp. 1–35.

[135] William S Robinson. 'Ecological correlations and the behavior of individuals'. In: *American Sociological Review* 15.3 (1950), pp. 351–357.

[136] Lee Ross. 'The intuitive psychologist and his shortcomings: Distortions in the attribution process'. In: *Advances in experimental social psychology*. Vol. 10. Elsevier, 1977, pp. 173–220.

[137] Lee Ross and Richard E Nisbett. *The person and the situation: Perspectives of social psychology*. McGraw-Hill, 1991.

[138] Amy Rummel and Richard Feinberg. 'Cognitive evaluation theory: A meta-analytic review of the literature'. In: *Social Behavior and Personality: an international journal* 16.2 (1988), pp. 147–164.

[139] Richard M Ryan and Edward L Deci. 'Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being'. In: *American psychologist* 55.1 (2000), p. 68.

[140] Golnaz Sadri and Ivan T Robertson. 'Self-efficacy and work-related behaviour: A review and meta-analysis'. In: *Applied Psychology: An International Review* (1993).

[141] Brigt Olav Samdal, Hanne Bakke and Sissel Nygård. *Årsrapport 2019*. [Online; accessed 7. Jun. 2020]. May 2020. URL: http://publikasjoner.nve.no/rapport/2020/rapport2020_11.pdf.

[142] Theodore R Sarbin. 'A contribution to the study of actuarial and individual methods of prediction'. In: *American Journal of Sociology* 48.5 (1943), pp. 593–602.

[143] Martina Angela Sasse, Sacha Brostoff and Dirk Weirich. 'Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security'. In: *BT technology journal* 19.3 (2001), pp. 122–131.

[144] Bruce Schneier. 'The psychology of security'. In: *International conference on cryptology in Africa*. Springer. 2008, pp. 50–79.

[145] Eugene Schultz. 'The human factor in security'. In: *Computers & Security* 24.6 (2005), pp. 425–426.

[146] Shalom H Schwartz. 'An overview of the Schwartz theory of basic values'. In: *Online readings in Psychology and Culture* 2.1 (2012), pp. 2307–0919.

[147] Shalom H Schwartz. 'Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries'. In: *Advances in experimental social psychology*. Vol. 25. Elsevier, 1992, pp. 1–65.

[148] Shalom H Schwartz and Anat Bardi. 'Value hierarchies across cultures: Taking a similarities perspective'. In: *Journal of cross-cultural psychology* 32.3 (2001), pp. 268–290.

[149] Shalom H Schwartz and Tammy Rubel. 'Sex differences in value priorities: Cross-cultural and multimethod studies'. In: *Journal of personality and social psychology* 89.6 (2005), p. 1010.

[150] Shalom H Schwartz et al. 'Value tradeoffs propel and inhibit behavior: Validating the 19 refined values in four countries'. In: *European Journal of Social Psychology* 47.3 (2017), pp. 241–258.

[151] *Science: Conjectures and Refutations*. [Online; accessed 10. Jun. 2020]. July 1953. URL: https://nemenmanlab.org/~ilya/images/0/07/Popper-1953.pdf.

[152] Michele Settanni, Danny Azucar and Davide Marengo. 'Predicting individual characteristics from digital traces on social media: A meta-analysis'. In: *Cyberpsychology, Behavior, and Social Networking* 21.4 (2018), pp. 217–228.

[153] Eric Shaw, Keven G Ruby and Jerrold M Post. 'The insider threat to information systems'. In: *Security Awareness Bulletin* 2.98 (1998), pp. 1–10.

[154] Galit Shmueli et al. 'To explain or to predict?' In: *Statistical science* 25.3 (2010), pp. 289–310.

[155] Herbert Simon. *The Sciences of the Artificial*. Tech. rep. The MIT Press, 1996.

[156] Herbert A Simon. 'Rational choice and the structure of the environment'. In: *Psychological review* 63.2 (1956), p. 129.

[157] Herbert A Simon. 'Science seeks parsimony, not simplicity: Searching for pattern in phenomena'. In: *Simplicity, inference and modelling: Keeping it sophisticatedly simple* (2001), pp. 32–72.

[158] Paul Slovic. 'Perception of risk'. In: *Science* 236.4799 (1987), pp. 280–285.

[159] Einar Snekkenes. 'Position paper: Privacy risk analysis is about understanding conflicting incentives'. In: *IFIP Working Conference on Policies and Research in Identity Management*. Springer. 2013, pp. 100–103.

[160] Chaoming Song et al. 'Limits of predictability in human mobility'. In: *Science* 327.5968 (2010), pp. 1018–1021.

[161] Craig P Speelman and Marek McGann. 'Challenges to mean-based analysis in psychology: The contrast between individual people and general science'. In: *Frontiers in psychology* 7 (2016), p. 1234.

[162] Paul M Spengler. 'Clinical versus mechanical prediction'. In: *Handbook of Psychology, Second Edition* 10 (2012).

[163] Holger Steinmetz, Rodrigo Isidor and Naissa Baeuerle. 'Testing the circular structure of human values: A meta-analytical structural equation modelling approach'. In: *Survey Research Methods*. Vol. 6. 1. 2012, pp. 61–75.

[164] VS Subrahmanian and Srijan Kumar. 'Predicting human behavior: The next frontiers'. In: *Science* 355.6324 (2017), pp. 489–489.

[165] Stephen Sutton. 'Predicting and explaining intentions and behavior: How well are we doing?' In: *Journal of applied social psychology* 28.15 (1998), pp. 1317–1338.

[166] Adam Szekeres and Einar Arthur Snekkenes. 'A Taxonomy of Situations within the Context of Risk Analysis'. In: *Proceedings of the 25th Conference of Open Innovations Association FRUCT*. FRUCT Oy Helsinki, Finland. 2019, pp. 306–316.

[167] Adam Szekeres and Einar Arthur Snekkenes. 'Construction of Human Motivational Profiles by Observation for Risk Analysis'. In: *IEEE Access* 8 (2020), pp. 45096–45107.

[168] Adam Szekeres and Einar Arthur Snekkenes. 'Predicting CEO Misbehavior from Observables: Comparative Evaluation of Two Major Personality Models'. In: *E-Business and Telecommunications. ICETE 2018. Communications in Computer and Information Science*. Vol. 1118. Springer, Cham, 2019, pp. 135–158.

[169] Adam Szekeres and Einar Arthur Snekkenes. 'Prediction of threat and opportunity risks: evaluation of a psychological approach using attributes of persons and situations'. In: *Risk Analysis: An International Journal* 0.0 (2020). Under review, pp. 0–0.

[170] Adam Szekeres and Einar Arthur Snekkenes. 'Representing decision-makers in SGAM-H: the Smart Grid Architecture Model Extended with the Human Layer'. In: *Graphical Models for Security*. Accepted for publication. Cham: Springer International Publishing, 2020, pp. 0–0.

[171] Adam Szekeres and Einar Arthur Snekkenes. 'Unobtrusive Psychological Profiling for Risk Analysis'. In: *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications - Volume 1: SECRYPT*. INSTICC. SciTePress, 2018, pp. 210–220.

[172] Adam Szekeres, Pankaj Shivdayal Wasnik and Einar Arthur Snekkenes. 'Using Demographic Features for the Prediction of Basic Human Values Underlying Stakeholder Motivation'. In: *Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 2: ICEIS*. SciTePress, 2019, pp. 377–389.

[173] Ágnes Szokolszky. *Kutatómunka a pszichológiában [Research in Psychology]*. Osiris Kiadó, Budapest, 2004.

[174] Daniel Szopinski, Thorsten Schoormann and Dennis Kundisch. 'Because Your Taxonomy is Worth IT: towards a Framework for Taxonomy Evaluation'. In: *27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS 2019, Stockholm and Uppsala, Sweden, June 8-14, 2019*. 2019. URL: https://aisel.aisnet.org/ecis2019%5C_rp/104.

[175] John R Taylor. *An introduction to error analysis: The study of uncertainties in physical measurements*. Sausalito, California: University Science Books, 1997.

[176] Shoshannah Tekofsky et al. 'Psyops: Personality assessment through gaming behavior'. In: *In Proceedings of the International Conference on the Foundations of Digital Games*. Citeseer. 2013.

[177] Maferima Touré-Tillery and Ayelet Fishbach. 'How to measure motivation: A guide for the experimental social psychologist'. In: *Social and Personality Psychology Compass* 8.7 (2014), pp. 328–341.

[178] Østrem Trond et al. *Norwegian Smart Grid Research Strategy*. 2015.

[179] EUROPEAN UNION. *DIRECTIVE 2009/28/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC*. 2009.

[180] Wendelien Van Eerde and Henk Thierry. 'Vroom's expectancy models and work-related criteria: A meta-analysis'. In: *Journal of applied psychology* 81.5 (1996), p. 575.

[181] CR Wilson VanVoorhis and Betsy L Morgan. 'Understanding power and rules of thumb for determining sample sizes'. In: *Tutorials in quantitative methods for psychology* 3.2 (2007), pp. 43–50.

[182] Viswanath Venkatesh, Susan A Brown and Hillol Bala. 'Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems'. In: *MIS quarterly* (2013), pp. 21–54.

[183] INC Verizon. *2010 Data Breach Investigations Report*. 2010. URL: https://www.wired.com/images_blogs/threatlevel/2010/07/2010-Verizon-Data-Breach-Investigations-Report.pdf.

[184] INC Verizon. *2020 Data Breach Investigations Report*. [Online; accessed 25. Jun. 2020]. June 2020. URL: https://enterprise.verizon.com/resources/reports/2020-data-breach-investigations-report.pdf.

[185] Stephen Walker. 'Workshop on stochastic error, parameter error and model error'. In: *1996 General insurance convention: 2-5 October 1996*. Springer, 1996, pp. 461–469.

[186] Eugene J Webb et al. *Unobtrusive measures - Nonreactive Research in the Social Sciences*. Rand McNally, Chicago, 1966.

[187] Ryan West. 'The psychology of security'. In: *Communications of the ACM* 51.4 (2008), pp. 34–40.

[188] Wikipedia. *Watergate scandal*. URL: https://en.wikipedia.org/wiki/Watergate%5C_scandal.

[189] JC Wofford and Laurie Z Liska. 'Path-goal theories of leadership: A meta-analysis'. In: *Journal of management* 19.4 (1993), pp. 857–876.

[190] Yu Yang, Stephen J Read and Lynn C Miller. 'A taxonomy of situations from Chinese idioms'. In: *Journal of Research in Personality* 40.5 (2006), pp. 750–778.

[191] Yu Yang, Stephen J Read and Lynn C Miller. 'The concept of situations'. In: *Social and Personality Psychology Compass* 3.6 (2009), pp. 1018–1037.

[192] Tal Yarkoni and Jacob Westfall. 'Choosing prediction over explanation in psychology: Lessons from machine learning'. In: *Perspectives on Psychological Science* 12.6 (2017), pp. 1100–1122.

[193] MD Young et al. 'Social cognitive theory and physical activity: a systematic review and meta-analysis'. In: *Obesity Reviews* 15.12 (2014), pp. 983–995.

Part II

Research Articles

Chapter 9

Article 1: Predicting CEO misbehavior from observables: comparative evaluation of two major personality models

Adam Szekeres & Einar Arthur Snekkenes. Predicting CEO Misbehaviour from Observables: Comparative Evaluation of Two Major Personality Models. In: *E-Business and Telecommunications. ICETE 2018. Communications in Computer and Information Science. Vol. 1118*. Springer, Cham. 2019, pp. 135–158.

Abstract

The primary purpose of this study is to demonstrate how publicly observable pieces of information can be used to build various psychological profiles that can be utilized for the prediction of behavior within a risk analysis framework[1]. In order to evaluate the feasibility of the proposed method, publicly available interview data is processed from a sample of chief executive officers (CEOs) using the IBM Watson Personality Insights service. The hypothesis-that group membership gives rise to a specific selection bias-is investigated by analyzing the IBM Watson-derived personality profiles at the aggregate level. The profiles are represented by two major theories of motivation and personality: the Basic Human Values and the Big Five models. Both theories are evaluated in terms of their utility for predicting

[1]The final authenticated version is available online at `https://doi.org/10.1007/978-3-030-34866-3_7`.

adverse behavioral outcomes. The results show that both models are useful for identifying group-level differences between (1) the sample of CEOs and the general population, and (2) between two groups of CEOs, when a history of rule-breaking behavior is considered. The predictive performance evaluation conducted on the current sample shows that the binary logistic regression model built from the Basic Human Values outperforms the Big Five model, and that it provides a practically more useful measurement of individual differences. These results contribute to the development of a risk analysis method within the domain of information security, which addresses human-related risks.

9.1 Introduction

Strategic decisions are long-term plans produced by a small number of senior managers aimed at achieving well-defined organizational objectives, with significant impact (positive or negative) on the safety and security of organizations and information systems spanning across the entire range of the corporate hierarchy. Such decisions affect a wide range of stakeholders, thus a certain level of friction is unavoidable [4, 37, 32]. The principal-agent problem within the economics and management literature addresses the tension between management interests and governance objectives. The principal-agent problem arises in agency theory and describes a situation in which one party (principal) delegates work to another party (agent) who is responsible for performing that work on behalf of the principal. The theory is concerned with resolving two problems that may arise in any agency relationship [8]. The first problem relates to the possibility that the agent's and the principal's desires or goals are in conflict, and it is difficult or expensive for the principal to verify what the agent is actually doing (i.e. hidden actions). The second problem arises from the difference between the parties' attitude towards risk, where the principal and the agent might prefer different actions due to different risk preferences and due to information asymmetry (i.e. hidden information).

Information security is a domain where negative externalities (e.g. principal-agent problem) may be present at various levels of abstraction. The highly complex threat landscape is characterized by misaligned stakeholder incentives (e.g. cost of developing sufficiently secure hardware and software vs. being first on the market, etc.), asymmetric knowledge about vulnerabilities (hidden information) and various other factors [1]. Internet of Things (IoT)-enabled critical infrastructures are becoming more and more prevalent due to their economic benefits. While they offer increased levels of automation, crucial strategic decisions are still the responsibility of people in leading positions. This may lead to situations in which the safety, security and stability of societies is increasingly dependent on the motivation of fewer and fewer key decision-makers [9].

Most information security risk analysis frameworks focus on the technological aspects and neglect the strategic decision-making perspective. The Conflicting Incentives Risk Analysis (CIRA) method developed by Rajbandhari and Snekkenes aims to bridge this gap by focusing on human motivation when addressing information security risks [24]. The method's applicability to real-world cases is limited by the lack of psychological theories that would enable the prediction of stakeholder behavior. Therefore, this study aims at evaluating two major psychological models of personality in terms of their performance for predicting undesirable stakeholder actions without direct access to subjects. The necessity for using unobtrusive profiling methods arises from the assumption that real-world stakeholders would be reluctant to explicitly reveal their motivations and they would be inclined to provide socially desirable answers when traditional assessment methods (i.e. questionnaire, interview, etc.) are utilized, which would confound the validity of the whole risk analysis process. While this study focuses on the misconduct of CEOs, the analysis is applicable to any other class of stakeholders.

Problem Statement

The CIRA method focuses on the misalignment between stakeholder motivations for risk identification [25]. To improve the method, it is necessary to incorporate psychological theories that enable the characterization of individual stakeholders and the prediction of their future behavior without requiring direct interaction between the analyst and the subjects. Based on these requirements, the objectives of this study are as follows:

- compare two personality models that can be used to characterize individual stakeholders,
- assess an unobtrusive profiling method's suitability for the purpose of risk analysis,
- analyze how a specific group membership gives rise to a selection bias, manifested in the psychological profiles,
- compare the predictive performance of the personality models with regard to undesirable behavioral outcomes.

Research Questions

Based on the aforementioned requirements and goals, the primary research question is as follows: *can publicly observable variables reflecting individual choice be used to construct psychological profiles suitable for predicting behavior in the context of risk analysis [35]?*

The following sub-questions were constructed in order to answer the main research question:

- **RQ 1:** To what extent is it feasible to use an unobtrusive profiling method to derive stakeholder characteristics?
- **RQ 2:** Is it feasible to detect a potential selection bias by analyzing personality profiles at the group-level?
- **RQ 3:** How does the *Basic Human Values* model compare to the *Big Five* model in terms of predicting stakeholder misbehavior?

This work contributes to the literature of information security risk analysis by presenting how publicly observable stakeholder data (i.e. recorded interviews) can be utilized for the purpose of risk analysis. The method relies on an existing application (IBM Watson Personality Insights), while the purpose of the analysis differs significantly from its established use cases. To assess the method's feasibility this study focuses on organizational leaders due to the fact that other classes of stakeholders might not be allowed to interact officially with the public, however the approach can be applicable to any other classes of stakeholders (e.g. CFO, COO, CIO, CISO). This study extends on previous work [35], by including an additional psychological model, and by comparatively evaluating the two personality models in terms of their capabilities for predicting real-world behavior. The paper is structured as follows: Section 9.2 introduces relevant theories and the IBM Watson application, Section 9.3 provides an overview about the methods used in the study. Results of the conducted analyses are presented in Section 9.4. Section 9.5 provides an overview about the results and their relevance, including limitations and plans for further work. Section 9.6 summarizes and concludes the present study.

9.2 Related work

This section provides an overview about the psychological theories, constructs and the application that served as the foundations of this study.

Sources of Bias

There are several research perspectives that aim to provide an explanation about the processes that guide people with certain traits or characteristics into various work positions. Extensive research investigates how different characteristics are desirable on one hand, and how they might have a negative impact on organizational or societal objectives. Several disastrous outcomes have been linked to the decision-maker's psychological attributes, which explains the increased research interest into

the ethical aspects of high-impact decision-making [34, 38]. This section introduces two main mechanisms that contribute to a selection bias in executive roles (i.e. personal attraction to a specific role and selection of candidates by the board of directors).

Selection Bias by Personal Motivations

Need for power, prestige and money are assumed to be key motivators that draw individuals to the highly competitive corporate world. Various decisions which contribute to undesirable social outcomes (e.g. exploiting sweatshop labor, environmental pollution, etc.) have been attributed to key decision-maker's psychological features. Furthermore, several organizational risks (e.g. embezzlement, bribery, etc.) can be enumerated which represent a conflict between the self-interested individual and the overall organizational objectives. One explanation for such incidents is proposed by Boddy, who discusses the over-representation of corporate psychopaths in key decision-maker positions. According to his definition corporate psychopaths are "people working in corporations who are self-serving, opportunistic, ego-centric, ruthless and shameless who can be charming, manipulative and ambitious" who are drawn to corporations since they can provide individuals with highly valued resources [3]. Corporate psychopaths are outwardly charming, and engaging, skillful at manipulating others to their own advantage, with a lack of concern for the consequences of their actions, and give a high priority for their own goals and ambitions. Their ability to demonstrate desirable traits that the organization values for a certain position is easily exploited by such individuals when presenting a charming facade, which distinguishes them from the commonly held perception of the insane psychopath.

The authors of [2] set out to investigate the prevalence and consequences of psychopathic tendencies in a sample of 203 corporate professionals taking part in a management development program. The study was motivated by the "growing public and media interest in learning more about the types of person who violate their positions of influence and trust, defraud customers, investors, friends, and family, successfully elude regulators, and appear indifferent to the financial chaos and personal suffering they create" [2]. The findings revealed the complex association between situation-congruent self-presentation and how psychopathic traits (although not classified as Antisocial Personality Disorder) can be beneficial in corporate environments. The results showed that the highest psychopathy scores were obtained from high-potential candidates in senior management positions. A noteworthy finding of the study is how the corporation evaluated individuals with several psychopathic traits. High psychopathy scores were associated with perceptions of good communication skills, strategic thinking, and creative/innovative abilities and simultaneously, with poor management style, failure to act as a team

player, and poor performance appraisals (as rated by immediate bosses).

Another empirical study investigated the association between the Dark Triad personality traits and the basic human values structure [13]. The Dark Triad (Machiavellianism, Narcissism, and Psychopathy) is a popular grouping of individual differences that represent antisocial personality traits below clinical threshold. The antisocial aspect of the triad comes from the shared underlying attitudes and modes of behavior that characterize these traits. Entitlement, superiority, dominance, manipulativeness, lack of remorse, impulsivity are the common features of the Triad [13]. The study found in two different cultures (i.e. Swedish and American) that Hedonism, Stimulation, Achievement and Power values were the highest ranking values for individuals high on Dark Triad traits. The authors claim that those characterized by high scores on the Dark Triad traits, hold values that promote Self-enhancement at the expense of others, thus treating other people as means toward their gains. The association between Self-enhancement values and the Dark Triad traits is referred to as dark value system which has further moral implications.

Selection Bias by Role Requirements

The match between certain personality features and various organizational settings is investigated by the Person-Organization (P-O) fit theories. Morley [20] discusses a shift in recent recruitment practices in which the traditional focus on knowledge, skills and abilities (KSAs), has moved toward seeking an optimal fit between the candidate's personality, beliefs and values and the organization's espoused culture, norms and values. In a similar vein, the Attraction-Selection-Attrition (ASA) framework seeks a fit at the personal level between the candidate and the organization's work values. According to the ASA model, candidates are attracted to organizations that exhibit characteristics similar to their own, and organizations tend to select employees who are similar to the organization in key aspects [28]. Value congruence has become a widely accepted operationalization of P-O fit [16].

Role requirements vary a lot even within the same organization (e.g. managerial role requirements are different from the requirements of a production line worker). A large-sample study aimed at identifying a distinctive managerial profile in terms of the Big Five model of personality. Managers reached significantly higher scores on the following nine personality traits and facet when compared to members of other occupations: Extraversion, Assertiveness, Conscientiousness, Emotional Stability, Agreeableness, Optimism, Work Drive, Customer Service Orientation, Openness. The results can be practically useful during the personnel selection process to increase the P-O fit required for specific job types [18].

Another investigation was conducted to test the hypothesis that different work

environments can be differentiated by analyzing the value structures of the workers [14]. The enterprising environment (e.g. manager, banker, financial advisor) is characterized by material and concrete goals, and requires one to lead, convince or manipulate others in order to achieve desired organizational and financial goals. According to the hypothesis Power and Achievement values are most compatible with these requirements, while the enterprising environment would inhibit the expression of Benevolence and Universalism values. The results revealed a strong positive correlation between the enterprising occupations and Power and Achievement values, while a negative correlation was observed in relation to Universalism values. This study successfully differentiated occupations based on the dominant human values that are present in each particular field, providing further evidence about a selection bias in action.

The surveyed research results highlight some of the ways through which selection bias is introduced to different work roles and occupations. First, individuals with certain traits or characteristics are attracted to specific jobs, then the active selection process by the recruiters produces the final set of employees. Analyzing the risks to an organization largely depends on understanding the nature of these biases.

Conflicting Incentives Risk Analysis

The relevance of focusing on the stakeholder motivation is recognized in the Conflicting Incentives Risk Analysis (CIRA) method [25]. It identifies stakeholders (i.e. individuals), the actions that can be taken by these stakeholders and the consequences of the actions. A stakeholder is an individual who has interest in taking a certain action within the scope of the analysis. The procedure distinguishes between two types of stakeholders: *Strategy owner* (the person who is capable of executing an action) and the *Risk owner* (whose perspective is taken - the person at risk). At the core of the method is the economic concept of utility, which captures the benefit of implementing a strategy for each stakeholder. The cumulative utility encompasses several utility factors, each representing valuable aspects for the corresponding stakeholders, thus modelling an individual's motivation. Two types of risks are identified in the method: *Threat risk* relates to the perceived decrease in the total utility for the risk owner, and *Opportunity risk* relates to missed utility gains due to the strategy owner's lack of motivation (i.e. costs associated with a beneficial action). Thus, risk is conceptualized as a misalignment of incentives between these two classes of stakeholders, and risk identification focuses on uncovering activities that would be beneficial for the Strategy owner while potentially harmful for the Risk owner [31]. Threat risk closely resembles the concept of moral hazard as it captures a wide range of behaviors that are beneficial for one party and detrimental for the other who has to suffer the consequences [6]. This study focuses on Threat risks that can be attributed to the motivation of organizational leaders.

Theory of Basic Human Values

The theory of *Basic Human Values* by Shalom Schwartz [29] identifies 10 distinct values that are universally recognized across various cultures and provides a unified and comprehensive view on the motivation of individuals. Values both represent desirable end goals and prescribe desirable ways of acting. Six key features characterize all values:

- "Values are beliefs linked to affect.
- Values refer to desirable goals that motivate action.
- Values transcend specific actions and situations.
- Values serve as standards or criteria.
- Values are ordered by importance.
- The relative importance of multiple values guide action." [29]

Furthermore, all 10 values capture one of the three key motivational aspects that are grounded in universal requirements of human existence: needs of individuals as biological organisms, requisites of coordinated social interaction, and survival and welfare needs of groups. Values guide behavior, given that the context or situation activates the relevant values. The values form a circular structure which represents a motivational continuum, where adjacent values are compatible with each other and opposing values are in conflict. The ten values are grouped under 4 higher dimensions as represented by Fig 9.1. The theory acknowledges that most actions are expressive of more than one value, and that a person's specific value-hierarchy modifies his/her perceptions about the relevant aspects of a situation. This may give rise to different interpretations of the same situation across individuals.

Big Five Personality Traits

The five factor model of personality or the *Big Five* defines five broad, distinct dimensions, that capture individual differences in terms of emotional, interpersonal experiences, recurring ways of behavior, and motivational styles [19]. The model is the result of several decades of extensive research in the domain of personality psychology, and represents one of the most widely accepted and utilized conceptualizations of personality. The five factors emerged from lexicographic investigations and are regarded as fundamental and stable dimensions of human personality, recognized across cultures. The large-scale acceptance of the model, and the consensus in relation to the utility of the Big Five provided researchers with

Figure 9.1: Circular value structure, with 4 higher dimensions. Source: [29]

a common framework from different traditions, which enabled productive investigations in a wide range of domains. It's practical applicability has been demonstrated extensively in industrial/organizational, educational, clinical and other (e.g. [11]) settings. According to trait theory, individuals can be placed on a continuum along the five main dimensions, which comprise of six facets (narrower, more specific aspects of personality [19]) as shown in Table 9.1.

Table 9.1: The Big Five dimensions and narrow facets of personality, based on [19].

Openness to experience	Conscientiousness	Extraversion	Agreeableness	Neuroticism
fantasy	competence	warmth	trust	anxiety
aesthetics	order	gregariousness	straightforwardness	hostility
feelings	dutifulness	assertiveness	altruism	depression
actions	achievement striving	activity	compliance	self-consciousness
ideas	self-discipline	excitement-seeking	modesty	impulsiveness
values	deliberation	positive emotions	tender-mindedness	vulnerability

IBM Watson Personality Insights

Personality Insights (PI) is part of IBM's artificial intelligence platform called Watson. Previously known for defeating the top human players in Jeopardy, the service these days is a comprehensive set of artificial intelligence solutions available for the consumer market. The service is utilized in a wide range of fields including health care, weather forecast, electric load optimization, etc. The PI utilizes machine learning solutions to uncover an individual's psychological characteristics based on texts produced by the person. The PI service's main use cases involve targeted marketing, customer care services, automated personalized interactions, among

several others. The service produces profiles based on four different models of individual differences [35]:

1. Big Five personality model - these characteristics describe relatively stable behavioral tendencies and modes of experiences.
2. Needs - based on the earliest investigations into human motivation capturing an individual's high-level desires.
3. Basic Human Values - values capture both desirable goals that people pursue and standards of acting, thus providing a summary about the underlying motivations behind one's actions.
4. Consumption preferences - optimized for predicting the user's likelihood for buying a certain product or engaging in different activities.

In terms of the Basic Human Values, the service calculates scores for five high-level dimensions: Conservation, Openness to change, Self-enhancement, Self-transcendence and Hedonism separately, whereas the original formulation identifies only four dimensions, and places Hedonism in either Openness to change or Self-enhancement. The service provides scores on all the Big Five dimensions as well as scores for each facet. For each personality model the PI computes two scores: percentile scores and raw scores. "To compute the percentile scores, IBM collected a very large data set of Twitter users (one million users for English, ...) and computed their personality portraits. IBM then compared the raw scores of each computed profile to the distribution of profiles from those data sets to determine the percentiles. The service computes normalized scores by comparing the raw score for the author's text with results from a sample population" [12]. While the percentile scores can provide insights about an individual's position on a trait compared to PI's original sample, it is not well-suited to characterize an individual's profile for the purpose of choice predictions, since the value structure relative to a sample population does not necessarily correspond to the individual's own value priorities. To allow comparison between different populations and scenarios the service also provides raw scores which resemble scores the person would get when completing a corresponding personality inventory. Thus raw scores are more useful for making comparisons to results derived from other studies.

9.3 Methods

Participants

The convenience sampling method produced a sample which consisted of 116 CEOs (105 male, 11 female), aged between 34-95 years (M = 59.41, SD = 9.23) with

sufficient amount of texts for running accurate analysis by the IBM Watson service. The amount of text available for the individuals ranged between 264-11384 words (M = 3830.98, SD = 1672.28). The majority of the subjects were born in the USA (N = 52.6%), followed by India (N = 12.9%), United Kingdom (N = 6.9%) and 21 other countries (N = 27.6%). 84.4% of the sample had at least bachelor or equivalent level degrees. The total compensation for the CEOs in year 2016 ranged between $45,936 - $46,968,924 (M = $15,988,276.78, SD = $10,600,982.56) according to publicly available sources [27].

Data Collection

The data collection and production activities (i.e. interview source identification, preprocessing, Watson analysis) are identical to those explained in [35]. In order to answer the Research Questions it was necessary to run an initial pilot study to assess the feasibility of the data collection activity. During the pilot study the first step involved the identification of relevant sources of data. To this end the Wikipedia article on the List of chief executive officers of notable companies was used that contains CEOs with diverse national and industrial backgrounds [39]. At the time of the start of the data collection the list consisted of 174 subjects. The second step involved the identification of suitable sources of information that could be linked to the individual and provided sufficient input to the Watson service for achieving it's maximum precision (3000 words/subject is recommended by the service description). In this phase we relied on video interviews, interviews published in online newspapers, news articles, company communications and social media profiles. Although it was possible to collect the necessary amount of data from the individuals, the procedure was not feasible due to high diversity of contexts, the uncertainty related to the actual author of the texts and the time needed to collect the data, so in the final data collection phase this procedure was modified in the following way:

- The search was restricted to videos published on YouTube that (a) were in English, (b) the subject could be clearly identified while providing his thoughts, and (c) were supplemented with captions.
- The search then was executed by using the subject's name with the following additional terms (in the same order): - interview, talk, presentation. In case the first search term did not provide sufficient amount of text the next one was used.
- In order to achieve as high validity as possible for the analysis we aimed at collecting mainly interviews and discussions that are more spontaneous

and reflective in content (thus we aimed at minimizing the reliance on well-rehearsed communications or texts written by other parties for presentation purposes).

- Each video was carefully observed in real time to check the accuracy of the captions and to ensure that only the subject's utterances are extracted for analysis, while omitting any noise (interviewer/audience questions, false transcriptions, etc.)
- A fresh install of Google Chrome was utilized in incognito mode, to keep personalized search results to a minimum and to maximize the reproducibility of the search results.

After a sufficient amount of text was collected from the subjects, the texts were submitted to the Watson PI service producing the psychological profiles for each individual [35].

For the purpose of a more fine grained analysis, CEOs that have been associated with various forms of rule breaking behavior leading to moral hazard have been identified in the current sample. To this end extensive web searches were conducted with the name of the individual and the additional search term (e.g. fraud, scandal, corruption). The first 20 search results were screened for each subject in order to identify possible associations with moral hazard. Using a broad sense of the moral hazard concept, any behavior was eligible for inclusion which had a negative effect on the reputation of the organization by drawing public attention to the underlying misconduct (irrespective of the nature of the misconduct) and the actions were conducted under the administration of the CEO in focus. The activities included: bribery of public officials, tax evasion, accounting fraud, insider deals, ethical misconduct, etc. The procedure resulted in the identification of 31 CEOs (26.7% of the sample) associated with undesirable behavior, and enabled profile comparisons between the two CEO groups [35].

The Concept of Difference

To characterize group differences several approaches were considered. In the first approach the percentile scores derived from the Watson PI service were used, that inherently contain a comparison between the subject's results and the original sample's distribution, on which the service was validated ($N \sim 1$ million users) [12]. This approach provides an understanding about the CEO sample's overall position across each personality dimension. Since the parameters are not publicly available for the original sample, a reference distribution was used to test differences between the current and the hypothesized original sample.

The second approach utilizes the raw scores derived from the PI service, which are equivalent to the scores one would get when completing an actual psychometric test (as suggested by the Watson manual [12]). These scores can be compared to results obtained from different populations, therefore are more suitable for validation. The second procedure followed this line of reasoning, and aimed at identifying differences between the profiles of CEOs and the general population.

However, rank orders in isolation do not provide all the necessary information about and individual's trade-off decisions, since a preference reversal (i.e. choosing different strategies with the same value orders among individuals) is possible. Considering this fact and in accordance with the theory's formulation, the relative importance of values should be analyzed when certain strategies are evaluated. Furthermore, since several studies use different instruments and methodologies for assessing the personality models or use different levels of analysis, it was necessary to enhance the compatibility and comparability of research findings [17]. To this end, in the third procedure the raw scores were summed across all dimensions, and each score was multiplied by the Sum^{-1}, to quantify each value's contribution to the overall utility (=1). The same procedure was carried out for research results that served as reference for the comparisons. This approach provides an assessment of an individual's personality profile independent of the instrument used for conducting the profiling. All analyses were conducted with SPSS 25 by IBM.

9.4 Results

Percentile score comparisons with Watson PI Sample

The first procedure aimed at detecting the existence of a selection bias using the percentile scores of each personality model. Percentile scores from the Basic Human Values and the Big Five scores were transformed by mapping them to a standard normal distribution, then for each dimension One-Sample t-tests were conducted with a reference standard normal distribution (M = 0) to assess whether the scores were drawn from the specific hypothesized distribution.

Basic Human Values

The results indicate that the group means for Conservation (M = -1.57), t(115) = -29.30, Hedonism (M = -1.95), t(115) = -81.24, Self-enhancement (M = -1.24), t(115) = -30.06, and Self-transcendence (M = -0.84), t(115) = -21.19, were significantly different from the reference distribution's mean scores, $p \leq 0.001$ for each. The group mean score of Openness to change (M = 0.06), t(115) = 1.14, p = 0.25 was not significantly different from the hypothesized population mean. Fig 9.2 shows the distribution of all the values based on the transformed percentile scores.

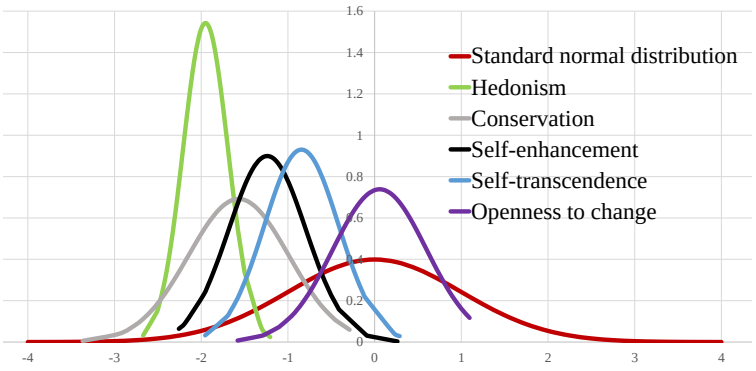


Figure 9.2: Basic Human Values percentile score distributions. [35]

Big Five

The same procedure was conducted for the Big Five dimensions and the results indicate that mean scores for the Big Five dimensions Openness to experience (M = 1.94), t(115) = 51.80, Conscientiousness (M = 0.62), t(115) = 14.58, Agreeableness (M = -0.79), t(115) = -11.33, and Neuroticism (M = 0.79), t(115) = 23.90 were significantly different from the reference distribution's mean scores, $p \leq 0.001$ for each. The mean score for Extraversion (M = -0.03), t(115) = -0.62, p = 0.54 was not significantly different from the hypothesized population mean. Fig 9.3 shows the distribution of scores on all the dimensions of the Big Five personality model using the transformed percentile scores.

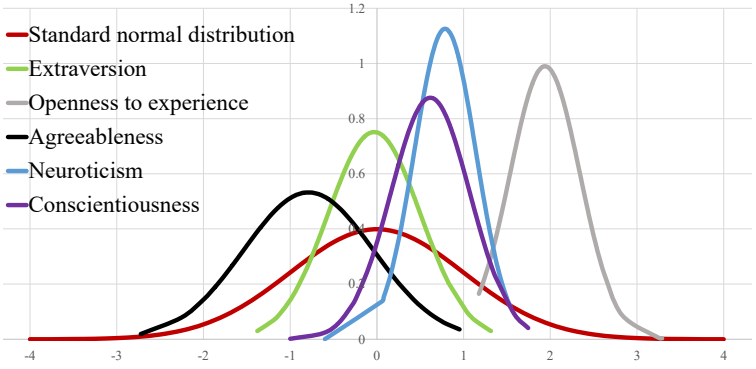


Figure 9.3: Big Five percentile score distributions.

Raw score comparison with samples from other studies

Raw scores provide information on how an individual would be scored when providing answers on the related personality inventory. Therefore, raw scores are more suitable for performing comparisons with results obtained from other published research studies.

Basic Human Values

In the following procedure the raw scores have been transformed to match with the original scale's scoring system used in the study by Schwartz and Bardi [30]. The representative or near-representative samples provide the necessary comparison that allows for a more detailed description of the value profiles. Fig 9.4a shows the general population's value priorities compared with the CEO value priorities based on the raw scores.

Big Five

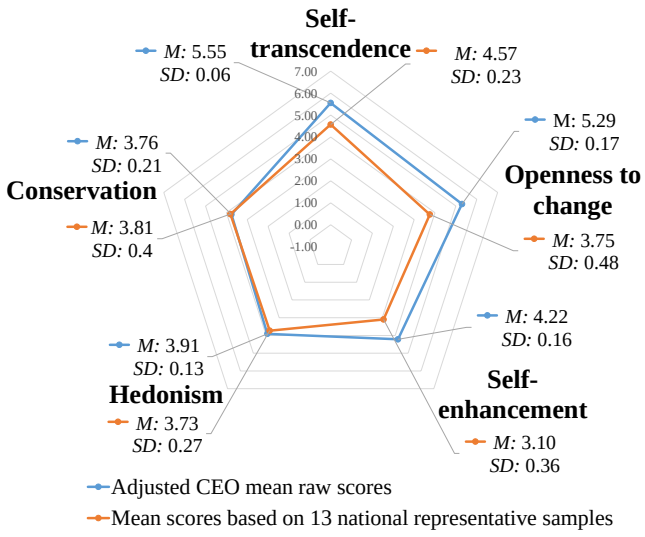
The Big Five profile scores were compared to a large-scale study, which gathered personality profiles from a sample of 132,515 American and Canadian internet users aged between 21-60 years [33]. The scores are reported using the percentage of maximum possible (POMP) scoring method, which is a metric constructed by a linear transformation of raw metric scores into a 0 to 100 scale, where 0 represents the minimum possible score and 100 represents the maximum possible score [5]. Therefore these scores are directly comparable to the raw scores derived from the IBM PI service (range 0-1). Fig 9.4b shows the mean score comparison between the large scale sample and the current CEO sample.

Comparison between CEO sub-groups

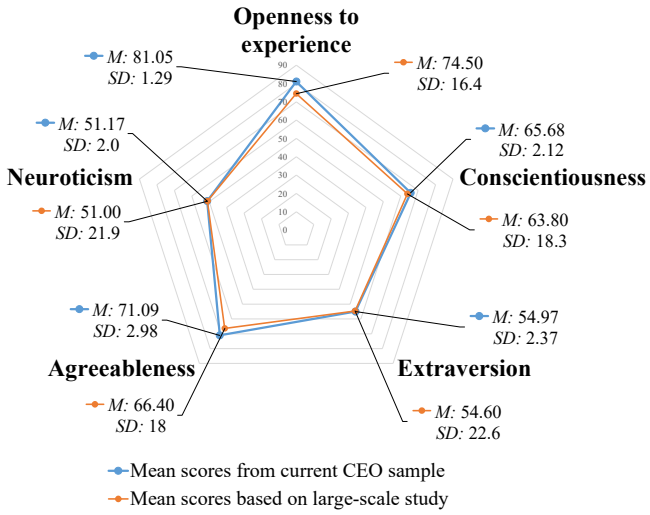
The following procedures aimed at analyzing differences among the two groups in the present CEO sample, based on a classification that identified a track record of rule-breaking behavior.

Basic Human Values

For the purpose of individual level choice prediction, the relative importance among the values has to be considered according to the original formulation of the theory. To this end, the profiles from the two CEO groups were converted to reflect relative importance among the Basic Human Values as described in 9.3, and five independent samples t-tests were performed on the raw scores to compare each value's importance across the two classes of CEOs to detect differences in the value profiles. Fig 9.5 illustrates the relative importance of values among the two CEO groups and the general population. Rank order of the values is marked above the



a Basic Human Values profile.



b Big Five personality dimensions.

Figure 9.4: Comparison of CEO raw profile scores from the IBM Watson PI service to research results obtained from representative samples.

bars where the CEO sample's ranking is followed by the general population's rank on each value. Table 9.2 shows the results of the performed t-tests.

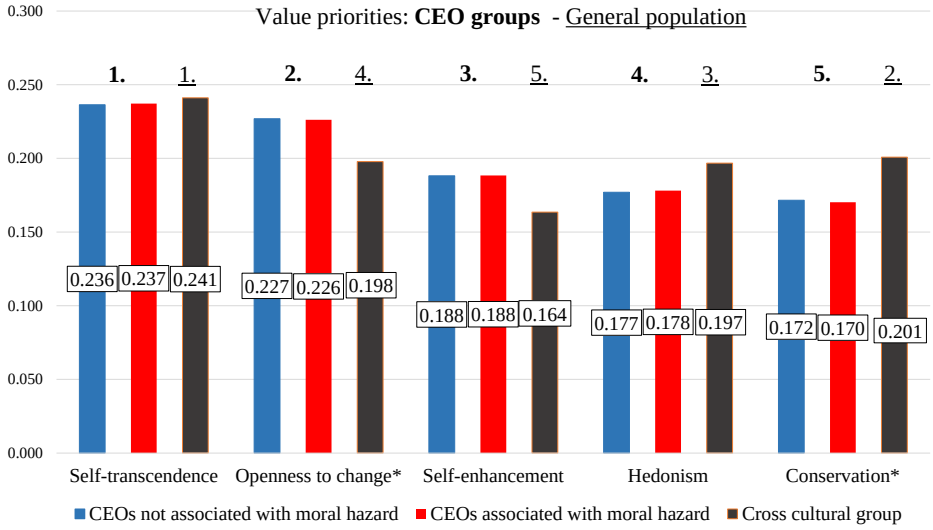


Figure 9.5: Comparison between the relative importance of the Basic Human Values among two groups of CEOs and general population. * marks a significant difference between the two CEO groups in terms of the importance of corresponding values [35].

Table 9.2: Results of the independent samples t-tests among two CEO groups using the Basic Human Values model [35].

	CEO raw scores associated with moral hazard (n = 31)		CEO raw scores not associated with moral hazard (n = 85)		
Values	M	SD	M	SD	t-test
Self-transcendence	0.82	0.01	0.82	0.01	n.s.
Openness to change	0.78	0.02	0.79	0.02	2.20*
Self-enhancement	0.65	0.02	0.65	0.02	n.s.
Hedonism	0.61	0.01	0.61	0.02	n.s.
Conservation	0.59	0.02	0.60	0.03	2.07*

Note. *p < .05; two-tailed.
M = Mean. SD = Standard Deviation

Big Five

The same grouping was used when running five independent samples t-tests to analyze which dimensions of the Big Five personality model indicate group-level differences among the two classes of CEOs. Table 9.3 presents results of the tests. Extraversion was the only dimension with significant difference between CEOs who have been linked to moral hazard, and those who have not, while the

other dimensions are statically indistinguishable from each other between these sub-groups.

Table 9.3: Independent samples t-tests among two CEO groups with the Big Five model.

	CEO raw scores associated with moral hazard (n = 31)		CEO raw scores not associated with moral hazard (n = 85)		
Big Five dimensions	M	SD	M	SD	t-test
Openness to experience	0.81	0.01	0.82	0.01	n.s.
Conscientiousness	0.65	0.02	0.66	0.02	n.s.
Extraversion	0.54	0.02	0.55	0.02	1.98*
Agreeableness	0.71	0.03	0.71	0.03	n.s.
Neuroticism	0.51	0.02	0.51	0.02	n.s

Note. *p = .05; two-tailed.
M = Mean. SD = Standard Deviation

Predictive performance comparison of the Basic Human Values and Big Five models

The final set of analyses focused on comparing the predictive capabilities of the two different personality models. Raw scores were transformed to z-scores and the guidelines provided by [23] were followed when conducting the analyses and presenting the results. Binary logistic regression models were built separately and the variables were entered in a single step in order to assess the overall predictive performance of the two theories. The dependent variable had two levels (i.e. clean track record vs evidence of rule-breaking, coded as 0 and 1). In case of the Basic Human Values model, the overall model evaluation proved that the model provided a significant improvement over the intercept only model, and the inferential goodness-of-fit test (Hosmer–Lemeshow) was insignificant ($p > .05$), suggesting that the model was fit to the data well. In case of the Big Five model, the overall model evaluation was not significantly better than the null-model.

Table 9.4 presents the overall model using the Basic Human values as predictors and Table 9.5 shows the details of the predictive performance evaluation of the model. For the Big Five personality dimensions, Table 9.6 shows the overall model and Table 9.7 shows the performance metrics related to this conceptualization of personality. Sensitivity and specificity were computed according to the guidelines provided by [10].

A final model was built, to test whether a combination of predictors from the two different theories could yield improved predictive performance. Predictors were

entered by using the conditional forward stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on conditional parameter estimates. The first block contained all Basic Human Values as predictors, and the next block contained all the Big Five dimensions. The resulting final model is shown in Table 9.8.

Table 9.4: Logistic regression model using the Basic Human Values profiles.

Predictor	β	*SE* β	Wald's χ^2	*df*	*p*	Odds ratio
Constant	-1.15	0.24	23.68	1	0.00*	0.32
Conservation	-0.50	0.27	3.47	1	0.06	0.61
Openness to change	-0.74	0.29	6.38	1	0.01*	0.48
Hedonism	-0.05	0.29	0.03	1	0.87	0.87
Self-enhancement	0.22	0.32	0.47	1	0.49	1.24
Self-transcendence	-0.24	0.28	0.78	1	0.38	0.78
Test			χ^2	*df*	*p*	
Overall model evaluation			**12.82**	**5**	**0.02***	
Goodness-of-fit-test:						
Hosmer & Lemeshow			12.34	8	0.14	

Note. *p < 0.05. Cox and Snell R^2 = .105. Nagelkerke R^2 = .152.

Table 9.5: Predictive performance evaluation of the Basic Human Values model.

	Predicted		
Observed	Yes	No	% Correct
Yes	7	24	22.6
No	4	81	95.3
Overall %			75.9

Note. TP: True Positive, TN: True Negative,
FP: False Positive, FN: False Negative,
Sensitivity = TP / (TP + FN) = 22.6%.
Specificity = TN / (TN + FP) = 95.3%.

Table 9.6: Logistic regression model using the Big Five profiles.

Predictor	β	*SE* β	Wald's χ^2	*df*	*p*	Odds ratio
Constant	-1.11	0.23	23.66	1	0.00*	0.33
Openness to experience	-0.09	0.25	0.13	1	0.71	0.91
Conscientiousness	-0.40	0.30	1.75	1	0.19	0.67
Extraversion	-0.61	0.27	5.12	1	0.02*	0.54
Agreeableness	-0.08	0.26	0.09	1	0.77	0.93
Neuroticism	0.70	0.31	4.97	1	0.03*	2.01
Test			χ^2	*df*	*p*	
Overall model evaluation			**10.76**	**5**	**0.06**	
Goodness-of-fit-test:						
Hosmer & Lemeshow			13.65	8	0.09	

Note. *p $<$ 0.05 Cox and Snell R^2 = .089. Nagelkerke R^2 = .129.

Table 9.7: Predictive performance evaluation of the Big Five Model.

	Predicted		
Observed	Yes	No	% Correct
Yes	4	27	12.9
No	2	83	97.6
Overall %			75

Note. TP: True Positive, TN: True Negative,
FP: False Positive, FN: False Negative
Sensitivity = TP / (TP + FN) = 12.9%.
Specificity = TN / (TN + FP) = 97.6%.

Table 9.8: Results of the logistic regression model by combining predictors from both theories.

Predictor	β	*SE* β	Wald's χ^2	*df*	*p*	Odds ratio
Constant	-1.13	0.23	23.73	1	0.00**	0.32
Openness to change	-0.61	0.23	6.93	1	0.01**	0.54
Conservation	-0.59	0.23	6.34	1	0.01**	0.56
Test			χ^2	*df*	*p*	
Overall model evaluation			**11.57**	**2**	**0.00****	
Goodness-of-fit-test:						
Hosmer & Lemeshow			9.36	8	0.31	

Note. **p $\leq$ 0.01 Cox and Snell R^2 = .095. Nagelkerke R^2 = .138.

9.5 Discussion

This study aimed at analyzing two different models of personality to detect a selection bias among chief executive officers by using text-based personality inferences provided by the IBM Watson PI service. Our results suggest that a selection bias can be detected by the Basic Human Values and the Big Five models as well. According to the results there are clearly identifiable differences among the universally established value structures in the general population and the sample of CEOs. Furthermore, differences can be identified in the Big Five profiles between these groups. This marked difference is interpreted as an evidence of a selection bias among organizational leaders. The importance of these differences in the motivational and personality structures is discussed in this section with directions for further work.

The analyses based on percentile scores revealed that both the Basic Human Value structure and the Big Five profile of the current sample of CEOs shows significant differences from the Watson Personality Insight service's hypothesized sample. With the exception of Openness to change (Basic Human Values) and Extraversion (Big Five), all other dimensions of the corresponding models showed differences from the original sample's hypothesized distributions. Due to the large sample size used during the validation of the service, it can be regarded as an indicator of valid differences between these samples, however due to the lack of detailed information about the original sample it is not possible to draw further conclusions based on percentile scores.

The second set of analyses focused on the utility of raw scores and the comparisons relied on established results from other large-scale studies. In terms of the Basic Human Values, the investigations revealed that there are important differences between the rank order of values among CEOs and the general population. While Self-transcendence values (i.e. care for the welfare of closely related others, as well as care for all the people and for nature) are most important for both groups the similarities between CEOs and non-CEOs end at this point. Openness to change (i.e. self-direction, independence, creating, stimulation and seeking out challenges) ranks as the second most important value in case of corporate leaders, while it is the second least important motivational factor for the general population. Openness to change and Conservation values can be found at opposing sides of the motivational circumplex, which reflects that decisions that promote the obtaining of a particular goal inhibit the simultaneous fulfillment of the competing need. Therefore a high priority given to Openness to change values would result in choices increasing novelty and chances for expressions of independent action at the expense of maintaining stability and stability. Self-enhancement values (i.e. expression of

competence, achievement of status and control over others) rank at the third position for CEOs, while it is the least important motivational value in the general population. Although one might expect that leaders of world-leading organizations (expressing power and achievement values) would be mainly motivated by Self-enhancement values at the expense of Self-transcendence values, these results contradict this expectation. The rank order difference of Self-enhancement values between non-CEOs (5.) and CEOs (3.) however clearly expresses their preference for high social status and prestige. While for non-CEOs, the second most important motivational tendencies relate to Conservation values (i.e. security, safety of self and of society, restraint of actions likely to harm others, respect for customs), these goals are less important to leaders, as it ranks the lowest on the their motivational hierarchy, indicating that actions promoting Conservation values have a much lower intrinsic motivational effect (e.g. in order to make an action appear at least as rewarding as an action expressing Openness to change values it has to be incentivized much more externally). The relative importance of values matches closely with the various Enterprising value profiles as discussed in [14], placing CEOs close to other occupations characterized by material and concrete goals.

In terms of the Big Five model, raw scores are more closely matched with those of the general population. A higher mean score on Openness to experience indicates elevated preference for adventure, novel experiences, curiosity and intellectual challenges, which can be seen as a desirable attribute for organizational leaders promoting growth, and motivating employees. On the other hand, it is also related to risk-taking behavior. Higher scores on Agreeableness is surprising, since lower scores are associated with competitiveness and self-direction, which are considered important leader characteristics. A more detailed analysis of the facet scores on this dimension could reveal which aspects contribute to the elevated score.

The third set of analyses aimed at identifying between-group differences within the current CEO sample, when previous history of misbehavior is taken into account. Based on the Basic Human Values model a slight, but significantly lower relative importance attributed to Openness to change and Conservation values was associated with various undesirable behaviors that can be detrimental to the reputation of the organization lead by the particular CEOs. Out of the Big Five dimensions only Extraversion showed a significant difference between groups, where lower Extraversion scores were associated with undesirable actions. This finding is similar to the results obtained by [26] which showed that self-reported computer criminal behavior was associated with higher levels of Introversion (i.e. lower levels on Extraversion) and similarly, no other significant differences were found between the two student groups in terms of the Big Five profiles.

The final evaluations were conducted to test the utility of the two major theories

for the prediction of behavioral outcomes. Since both theories aim at providing a comprehensive view on the organization of the human psyche by identifying basic and necessary structures that are pervasive and relatively stable within individuals [19], they were used in two separate logistic regression models as a single unit. A third model was built to investigate whether a combination of the two theories could achieve improvements over any of the models in isolation. The model built from the Basic Human Values represented a significant improvement from the null-model, and achieved the highest score on the R^2 metric ($R^2 = 0.152$) out of the three models. The logistic regression analysis including all the Big Five dimensions resulted in a model that was not significantly better than a null-model, which purely guesses the majority class. This finding is surprising considering that the Big Five is the most widely accepted and utilized model of personality, and several studies claim that it has substantial predictive utility in a wide range of domains [22, 21]. The final combined model contained no predictors from the Big Five (none of them reached the inclusion criteria), thus all variance explained by the model is attributed to Basic Human Values. The overall model reached a higher significance level (i.e. lower p value) at the expense of some explained variance (change from the model with all Basic Human Values in terms of R^2 is: -0.014). This results suggests that the two models are to a great extent overlapping, but the Basic Human Values model might be more comprehensive.

A limitation of the present study is the relatively small sample size, which can be extended in future studies, since the method of analyzing personality profiles by using the Watson PI service is a feasible method for gathering information about the motivation of decision makers for the purpose of risk analysis. Sample size limitations may potentially hamper the performance of the binary logistic regression models, therefore it would be necessary to increase the number of observations for events and non-events for improved models. It would potentially lead to better sensitivity and specificity scores, and in order to compute positive and negative predictive values, the prevalence rates of offending behavior could be investigated in future work [10]. Furthermore, a more detailed description and classification of the various forms of rule-breaking behavior could clarify the connection between the particular strategy owner's profile and the nature of negative impact inflicted upon the organization, to achieve a better assessment of the risks relating to individuals.

In a risk analysis setting direct access to subjects is a major limitation. Since previous work has established the extent to which the most easily available pieces of information (i.e. demographic features) are useful for constructing stakeholder profiles [36], future work will focus on other classes of observable features (e.g. ownership of items [7], or various forms of online behavior with digital traces [15], etc.) for the construction of psychological profiles.

9.6 Conclusion

This exploratory study aimed at analyzing how publicly observable pieces of information (i.e. spoken texts, group membership) associated with individuals can be utilized to detect a selection bias among groups of people working in similar roles. A set of chief executive officers were selected for the purpose of testing the methods' usefulness, for two main reasons: the availability of relevant and necessary data, and due to the significance of the role they play in organizations. However, the principles presented in this study are applicable to other classes of stakeholders as well, and are not limited to the CEO role. The selection bias is revealed by patterns of specific psychological characteristics that distinguish CEOs from the general population. Furthermore, within the analyzed CEO sample, additional differences could be detected among two groups that were generated by considering available evidence about rule-breaking behavior (i.e. association with moral hazard).

The specific psychological differences were investigated through two major theories that account for stable individual differences among people. The Big Five personality model is evaluated against the Basic Human Values model in terms of group-level differences, and in terms of predictive capabilities. The results show that both models are useful in detecting a hypothesized selection bias, but the Basic Human Values model performs better in terms of predictive utility as a comprehensive model of individual differences and motivation. The unobtrusive nature of the text analysis combined with the procedures described in this study enables risk analysts to study human-related risks in various environments where adversarial stakeholder behavior is assumed and it is crucial to be prepared against undesirable consequences of those actions (e.g. information security).

ACKNOWLEDGEMENTS

This work was partially supported by the project IoTSec – Security in IoT for Smart Grids, with number 248113/O70 part of the IKTPLUSS program funded by the Norwegian Research Council.

References

[1] Ross Anderson and Tyler Moore. 'Information security: where computer science, economics and psychology meet'. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 367.1898 (2009), pp. 2717–2727.

[2] Paul Babiak, Craig S Neumann and Robert D Hare. 'Corporate psychopathy: Talking the walk'. In: *Behavioral sciences & the law* 28.2 (2010), pp. 174–193.

[3] Clive R Boddy. ‘The implications of corporate psychopaths for business and society: An initial examination and a call to arms’. In: *Australasian Journal of Business and Behavioural Sciences* 1.2 (2005), pp. 30–40.

[4] John Alan Cohan. ‘" I didn’t know" and" I was only doing my job": Has corporate governance careened out of control? A case study of Enron’s information myopia’. In: *Journal of Business Ethics* 40.3 (2002), pp. 275–299.

[5] Patricia Cohen et al. ‘The problem of units and the circumstance for POMP’. In: *Multivariate behavioral research* 34.3 (1999), pp. 315–346.

[6] Allard E Dembe and Leslie I Boden. ‘Moral hazard: a question of morality?’ In: *New Solutions: A Journal of Environmental and Occupational Health Policy* 10.3 (2000), pp. 257–279.

[7] Thomas Dohmen et al. ‘Individual risk attitudes: Measurement, determinants, and behavioral consequences’. In: *Journal of the European Economic Association* 9.3 (2011), pp. 522–550.

[8] Kathleen M Eisenhardt. ‘Agency theory: An assessment and review’. In: *Academy of management review* 14.1 (1989), pp. 57–74.

[9] Olav B Fosso et al. ‘Moving towards the smart grid: The Norwegian case’. In: *Power Electronics Conference (IPEC-Hiroshima 2014-ECCE-ASIA), 2014 International*. IEEE. 2014, pp. 1861–1867.

[10] Alan G Glaros and Rex B Kline. ‘Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model’. In: *Journal of clinical psychology* 44.6 (1988).

[11] Hannes Grassegger and Mikael Krogerus. *The Data That Turned the World Upside Down*. Jan. 2017. URL: https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win.

[12] IBM. *The science behind the service*. [Online; accessed 14-February-2018]. 2017. URL: https://console.bluemix.net/docs/services/personality-insights/science.html#science.

[13] Petri J Kajonius, Bjorn N Persson and Peter K Jonason. ‘Hedonism, achievement, and power: Universal values that characterize the Dark Triad’. In: *Personality and Individual Differences* 77 (2015), pp. 173–178.

[14] Ariel Knafo and Lilach Sagiv. ‘Values and work environment: Mapping 32 occupations’. In: *European Journal of Psychology of Education* 19.3 (2004), pp. 255–273.

[15] Michal Kosinski, David Stillwell and Thore Graepel. 'Private traits and attributes are predictable from digital records of human behavior'. In: *Proceedings of the National Academy of Sciences* 110.15 (2013), pp. 5802–5805.

[16] Amy L Kristof-Brown, Ryan D Zimmerman and Erin C Johnson. 'Consequences of individual's fit at work: a meta-analysisof person–job, person–organization, person–group, and person–supervisor fit'. In: *Personnel psychology* 58.2 (2005), pp. 281–342.

[17] Marjaana Lindeman and Markku Verkasalo. 'Measuring values with the short Schwartz's value survey'. In: *Journal of personality assessment* 85.2 (2005), pp. 170–178.

[18] John W Lounsbury et al. 'Core personality traits of managers'. In: *Journal of Managerial Psychology* 31.2 (2016), pp. 434–450.

[19] Robert R McCrae and Oliver P John. 'An introduction to the five-factor model and its applications'. In: *Journal of personality* 60.2 (1992), pp. 175–215.

[20] Michael J Morley. 'Person-organization fit'. In: *Journal of Managerial Psychology* 22.2 (2007), pp. 109–117.

[21] Sampo V Paunonen. 'Big Five factors of personality and replicated predictions of behavior'. In: *Journal of personality and social psychology* 84.2 (2003), p. 411.

[22] Sampo V Paunonen and Michael C Ashton. 'Big five factors and facets and the prediction of behavior'. In: *Journal of personality and social psychology* 81.3 (2001), p. 524.

[23] Chao-Ying Joanne Peng, Kuk Lida Lee and Gary M Ingersoll. 'An introduction to logistic regression analysis and reporting'. In: *The journal of educational research* 96.1 (2002), pp. 3–14.

[24] Lisa Rajbhandari and Einar Snekkenes. 'Intended actions: Risk is conflicting incentives'. In: *Information Security* (2012), pp. 370–386.

[25] Lisa Rajbhandari and Einar Snekkenes. 'Using the Conflicting Incentives Risk Analysis Method'. In: *Security and Privacy Protection in Information Processing Systems*. Ed. by Lech J. Janczewski, Henry B. Wolfe and Sujeet Shenoi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 315–329.

[26] Marcus K Rogers, Kathryn Seigfried and Kirti Tidke. 'Self-reported computer criminal behavior: A psychological analysis'. In: *digital investigation* 3 (2006), pp. 116–120.

[27] Salary.com. *Executive Compensation It Starts with the CEO*. [Online; accessed 14-February-2018]. 2004. URL: https://www.salary.com/executive-compensation-it-starts-with-the-ceo/.

[28] Benjamin Schneider, W Goldstein Harold and D Brent Smith. 'The ASA framework: An update'. In: *Personnel psychology* 48.4 (1995), pp. 747–773.

[29] Shalom H Schwartz. 'An overview of the Schwartz theory of basic values'. In: *Online readings in Psychology and Culture* 2.1 (2012), pp. 2307–0919.

[30] Shalom H Schwartz and Anat Bardi. 'Value hierarchies across cultures: Taking a similarities perspective'. In: *Journal of cross-cultural psychology* 32.3 (2001), pp. 268–290.

[31] Einar Snekkenes. 'Position paper: Privacy risk analysis is about understanding conflicting incentives'. In: *IFIP Working Conference on Policies and Research in Identity Management*. Springer. 2013, pp. 100–103.

[32] Bahram Soltani. 'The anatomy of corporate fraud: A comparative analysis of high profile American and European corporate scandals'. In: *Journal of business ethics* 120.2 (2014), pp. 251–274.

[33] Sanjay Srivastava et al. 'Development of personality in early and middle adulthood: Set like plaster or persistent change?' In: *Journal of personality and social psychology* 84.5 (2003), p. 1041.

[34] Cheryl K Stenmark and Michael D Mumford. 'Situational impacts on leader ethical decision-making'. In: *The Leadership Quarterly* 22.5 (2011), pp. 942–955.

[35] Adam Szekeres and Einar Arthur Snekkenes. 'Unobtrusive Psychological Profiling for Risk Analysis'. In: *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications - Volume 1: SECRYPT*. INSTICC. SciTePress, 2018, pp. 210–220.

[36] Adam Szekeres, Pankaj Shivdayal Wasnik and Einar Arthur Snekkenes. 'Using Demographic Features for the Prediction of Basic Human Values Underlying Stakeholder Motivation'. In: *Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 2: ICEIS*. SciTePress, 2019, pp. 377–389.

[37] Karen A Van Peursem et al. *Three cases of corporate fraud: an audit perspective*. 2007.

[38] Glen Whyte. 'Decision failures: Why they occur and how to prevent them'. In: *Academy of Management Perspectives* 5.3 (1991), pp. 23–31.

[39] Wikipedia. *List of chief executive officers Wikipedia, The Free Encyclopedia*. [Online; accessed 06-December-2017]. 2004. URL: https://en.wikipedia.org/wiki/List_of_chief_executive_officers.

Chapter 10

Article 2: Using Demographic Features for the Prediction of Basic Human Values Underlying Stakeholder Motivation

Adam Szekeres & Pankaj Shivdayal Wasnik & Einar Arthur Snekkenes. Using Demographic Features for the Prediction of Basic Human Values Underlying Stakeholder Motivation. In: *Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 2: ICEIS*. SciTePress. 2019, pp. 377–389.

Abstract

Human behavior plays a significant role within the domain of information security. The Conflicting Incentives Risk Analysis (CIRA) method focuses on stakeholder motivation to analyze risks resulting from the actions of key decision makers. In order to enhance the real-world applicability of the method, it is necessary to characterize relevant stakeholders by their motivational profile, without relying on direct psychological assessment methods. Thus, the main objective of this study was to assess the utility of demographic features-that are observable in any context-for deriving stakeholder motivational profiles. To this end, this study utilized the European Social Survey, which is a high-quality international database, and is comprised of representative samples from 23 European countries. The predictive performances of a pattern-matching algorithm and a machine-learning method are compared to establish the findings. Our results show that demographic features are

marginally useful for predicting stakeholder motivational profiles. These findings can be utilized in settings where interaction between a stakeholder and an analyst is limited, and the results provide a solid benchmark baseline for other methods, which focus on different classes of observable features for predicting stakeholder motivational profiles.

10.1 Introduction

Information security is considered to be a highly technical domain, where research on the human element gets relatively low attention, given the involvement and impact of individuals on the system's safety and security. However, "...people are responsible for stealing passwords, committing intellectual property crimes, skimming financial accounts, selling information to competitors, breaking into databases, cyber-snooping, and committing a host of other offenses against organizations and their systems. Ironically, the disciplines that assess, evaluate, and solve human based problems have not been an integral part of the information security measures used to protect data..." [13]. It is suggested that there is a need for synthesis between various disciplines in order to improve on the attempts that aim to protect against threats to information systems. More than a decade later, Greitzer and Hohimer [12] concluded that insider threats ranked among the most problematic cyber-security challenges that threaten government and industry information infrastructures. Furthermore, they identified that there were no systematic methods that provided a complete and effective approach to preventing undesirable actions (e.g. data leakage, espionage, and sabotage).

More recent incidents (e.g. using technical expertise and insider privileges to reprogram Smart Meters [19], cheating with emission rates [2], financial misreporting [20], creating abusive websites [9], etc.) also call for methods that incorporate intentional, deliberate human behavior into risk assessments. While the specific details of the enumerated incidents vary greatly, they are still united by some common features:

- It is possible to identify a person or a group who had a strong motivation to take certain actions.
- It is possible to identify a person or a group who suffered the consequences of those actions but who were unintentionally exposed to those transactions.

Such situations are recognized in the economic literature as negative externalities [21] and the concept has been applied within the domain of information security, where motivated actors have the potential to exert a negative influence on a large

number of other stakeholders who have little influence on the outcome of those actions [1].

Assessing stakeholder motivation could be the key to preparing against such events, since motivation is a central concept in understanding human behavior; it aims to answer the question concerning why people do the things they do [8]. During the past centuries, researchers have generated a vast number of theoretical constructs and systems which vary in the level of the analysis (e.g. instincts, biologically determined drives, needs, social and cognitive motivations), the scope (e.g. general principles vs. task-specific motivations), and the terminology. Through describing stakeholder motivation we can enable the prediction of future behaviors and check whether the likely behavior is in alignment with the goals of other affected stakeholders. However, people are not expected to cooperate in any analysis that aims to assess their motivations for risk-analysis purposes. Therefore, the main goal of the present study is to contribute to the information security risk management literature by investigating the utility of demographic features for deriving stakeholder motivational profiles in contexts where no direct interaction between the subject and analyst is assumed.

Following the Problem Statement and Research Questions, Section 10.2 describes the risk analysis method under development, and it's connection to the theory of basic human values. Section 10.3 explains how a publicly available high-quality dataset was utilized in the study, which is followed by describing the results in Section 10.4. Section 10.5 provides an overview of the conducted work, and Section 10.6 concludes with directions for future work.

Problem Statement

The main objective of this work is to investigate how stakeholder motivation can be predicted by utilizing publicly observable individual characteristics (e.g. demographic variables). The end goal is the development of a predictive model that can be utilized by an observer to derive the motivational profile of a previously unknown subject by collecting and aggregating various forms of publicly observable features connected to the subject.

Research Questions

To address the problem statement, the following research questions have been formulated:

1. To what extent can demographic features be utilized to construct stakeholder motivational profiles?
2. How well do different predictive models perform in terms of inferring stake-

holder motivational profiles?

10.2 Related work

This section provides an overview of the risk-analysis method under development, the motivational theory, and the related constructs that were included in the study.

Conflicting Incentives Risk Analysis

The importance of understanding stakeholder motivation is emphasized within the Conflicting Incentives Risk Analysis (CIRA) method [25]. This method identifies the stakeholders (i.e. individuals), the actions that can be taken by the stakeholders, as well as the consequences of these actions. A stakeholder is a physical person who has some interest in the outcomes of his actions. The procedure identifies two types of stakeholders: the *Strategy owner* (the person who is capable of executing an action) and the *Risk owner* (whose perspective is taken-the person at risk). Each stakeholder's motivation is modeled on the concept of utility, which entails the consideration of the benefit of the action performed from the perspective of the stakeholder. This cumulative utility encompasses several utility factors, each representing aspects of life considered important by the corresponding stakeholders. Two types of risks are identified in the method: Threat risk refers to the perceived decrease in the total utility of the risk owner and Opportunity Risk refers to the lack of potential increase in utility because the strategy owner is not motivated enough to take actions that would be beneficial for the Risk owner. Therefore, risk is conceptualized as a misalignment of incentives between these two classes of stakeholders, and risk identification is about uncovering activities that would be beneficial for the Strategy owner, and potentially harmful for the Risk owner, or vice versa [33]. Therefore, Threat risk closely resembles the concept of moral hazard; it captures a wide range of behaviors that are beneficial for one party and detrimental for another (i.e. the strategy owner inflicting negative externalities on the risk owner) [5]. Previous work explored the feasibility of inferring key stakeholders' motivational profiles based on the linguistic analysis of interviews given by inaccessible subjects [34].

Theory of Basic Human Values

The theory of basic human values, developed by Schwartz, [28] identifies ten distinct values that are universally recognized across various cultures, and it provides a unified and comprehensive view on human motivation. The theory incorporates several previous approaches that emphasized the centrality of values in human behavior (e.g. Hofstede and Rokeach on cultural differences [31]). Values both represent desirable end-goals and prescribe desirable ways of acting. Schwartz

summarizes the six core features that characterize values:

- "Values are beliefs linked to affect.
- Values refer to desirable goals that motivate actions.
- Values transcend specific actions and situations.
- Values serve as standards or criteria.
- Values are ordered by importance.
- The relative importance of multiple values guide actions."

Furthermore, all of the ten distinct values in the theory encapsulate one of the three key motivational aspects that are grounded in the universal requirements of human existence: the needs of individuals as biological organisms, the requisites of coordinated social interaction, and the survival and welfare needs of groups. Values guide behavior, given that the decision context, or situation activates the relevant values. The ten values form a circular structure that captures a motivational continuum, where adjacent values are compatible with each other, while opposing values are in conflict. The ten values are grouped under four higher dimensions, as represented by Figure 10.1 [27].

Figure 10.1: Circular value structure, with 4 higher dimensions comprising of the 10 basic human values.

Goldberg, Sweeney, Merenda, and Hughes [10] describe how one of the most enduring topics in the history of psychometrics is the strength of association between

group and individual differences, and the many controversies centered around the issue of how various demographically defined groups differ in terms of important human attributes. In their study, they investigated the differences between the Big Five personality traits and four demographic variables (i.e. gender, age, education, and ethnic status). The study concluded that most demographic-personality associations are of trivial size, with an average correlation of 0.08 (across the four demographic variables and the five personality dimensions included in the study). However, these results are not directly comparable to the value-demographic association yet, they nevertheless provided some initial insights into the strength of associations between demographic features and psychological variables. Schwartz [29] discusses the reciprocal relationship between value priorities and life circumstances and provides empirical evidence on the hypothetical relationships. Choices guided by values influence the life circumstances, but certain life circumstances (e.g. the type of profession, raising children, etc.) also affect the possibility of, and constraints placed upon, enacting particular choices. People tend to adapt their values to fit into their life circumstances by upgrading the importance of values that are readily attainable, while downgrading the importance of values of which the pursuit is blocked. As people's demographic variables (e.g. age, gender, education, income level, etc.) largely impact the circumstances to which they are exposed, these differences are expected to have a direct effect on the value priorities. Based on the value system's structure, the following subsections present validated and hypothesized relationships between demographic variables and value priorities based on [29].

Age

Due to the general decline of physical strength and cognitive abilities, aging is expected to increase the importance of Security values, as the capacity to deal with change declines. Therefore, the opposing Stimulation value might decrease in importance as novelty and risk is viewed as increasingly threatening. Conformity and Tradition values might increase in importance, while Hedonism could potentially decrease due to the dulling of the senses. Achievement and Power values may also decrease in importance since older people become less able to perform demanding tasks and obtain social approval.

Life stages

In early adulthood people are primarily concerned with establishing themselves within the domains of work and family. The pursuit of Achievement and Stimulation values comes at the expense of the Security, Conformity, and Tradition values. Later, the motivation shifts to preserving the status already attained, both in the professional and in the family domains. The possibility of radical change narrows

and responsibilities constrain the opportunities for risk-taking. Taking these factors into consideration, it is expected that people in their middle adulthood express a stronger preference for values encompassed in the Conservation category. At later stages, close to retirement, the opportunities for expressing Achievement, Power, Stimulation, and Hedonism values further decrease.

Gender

In a cross-cultural, large scale study, Schwartz and Rubel investigated gender differences in value priorities [32]. The findings suggest that men attribute more importance to Self-enhancement and Openness to change values than women do, while for Self-transcendence values, the reverse is true. The differences are generally small, and account for less variance than age and culture do, for example.

Education

An explanation for the association between the level of education and the values is offered in [29]. According to the hypothesis education requires intellectual openness, and flexibility that is associated with Self-direction values. Challenging existing views and norms can be linked to a lower importance assigned to Conservation values, as they promote conformity and tradition. Furthermore, there might be a positive correlation with Achievement values as performance and meeting external standards is increasingly important as the level of education rises.

Country

The challenges faced by nations in organizing human activities are similar, but nations differ in the importance they attribute to certain values [30]. When values are analyzed at the societal level, three bipolar dimensions can be identified based on the alternative resolutions to each of the problems affecting all societies: Embeddedness vs. Autonomy (affective and intellectual), Hierarchy vs. Egalitarianism, and Mastery vs. Harmony. The importance assigned by various countries to the previous dimensions gives rise to eight distinct cultural regions, representing vague differences among cultures: Western Europe, East-Central Europe, Eastern Europe, Latin America, English-Speaking, Confucian, South-East Asia, and Africa-Middle East.

Occupation

Another study by Knafo and Sagiv [17] investigated the relationship between values and occupational choices. The survey-based study showed that the 32 occupations under investigation clustered according to the motivational profiles of the individuals within the profession, and that these clusters fit well into Holland's work typology. Universalism values negatively correlated with the Enterprising work en-

vironment, while Social environments correlated positively with both Universalism and Benevolence values, and correlated negatively with power and Achievement values. Artistic work environments correlated negatively with Conformity values while the Investigative environments correlated positively with Openness to change values.

These results suggest that there are meaningful and detectable differences among various groups of people. However, to our knowledge, there is no existing study that investigates how well the motivational profile can be predicted when solely based upon demographic features. Therefore, this study aims to establish predictive models from a high-quality database that contains representative samples from 23 European countries.

10.3 Materials and Methods

Sample and Procedure

The European Social Survey (ESS), round 8, edition 2.0, [22] served as the main source of answers to the research questions. The high-quality cumulative dataset contains individual-level data from 23 countries (Austria, Belgium, the Czech Republic, Estonia, Finland, France, Germany, Hungary, Iceland, Ireland, Israel, Italy, Lithuania, the Netherlands, Norway, Poland, Portugal, the Russian Federation, Slovenia, Spain, Sweden, Switzerland, and the United Kingdom), gathered using strict probability sampling methods. The survey's main objectives are to monitor and interpret changing public attitudes in Europe, to investigate relevant societal issues, and to establish social indicators across Europe. The original dataset contains a total of (n = 44 387) individual respondents with 536 variables. The ESS has been conducted every two years since 2001 across European many countries. The survey consists of two main parts:

- The core module - covers a wide range of topics (e.g. politics, social trust, household, socio-demographics, human values, etc.) that largely remain the same in each round to allow for longitudinal observations.
- The rotating module - increases the scope of the survey by focusing on specific topics between different times of administration (e.g. immigration, economic morality, justice, democracy, climate change, etc.)

Measures

In order to address the research questions, the following preparation procedures were conducted on the original cumulative dataset. In the first step, the complete

list of variables (N_{vars} = 536) was screened and then it was sorted into four main categories (demographics, attitudes, behaviors, and others). The next step focused on identifying the demographic attributes that met the inclusion criteria (i.e. the predictor variables should be publicly observable and easily identifiable by an observer). This resulted in a list of demographic variables being included in the present analysis (N_{vars} = 14), accompanying the basic human values. Table 10.1 contains the list of independent variables selected for the analysis. We aimed at maximizing the number of subjects with valid responses, therefore, the next step was to investigate the number of missing values in the sample. Since our objective was to analyze the predictability of the motivational profiles of individuals who are actively employed we used a listwise deletion of subjects with missing values on any of the remaining variables. The listwise removal of data is justified by the fact that most of the missing data was attributed to four variables associated with employment relations (the last four variables in Table 10.1), with a not-applicable label (e.g. the not actively working age-group) which contributed to a total of 7255 subjects with missing data, while the remaining missing data (n = 385) was distributed among the ten other independent variables (with the labels: refusal, do not know, no answer, not available). While it was not possible to determine whether the data was missing at random, completely at random, or not at random for the remaining small number of cases, the relatively small number enabled deletion without introducing a bias into the models. Additionally, the 89 levels of variable "Type of industry working for" were grouped according to the NACE rev. 2. section codes, resulting in 21 higher level groups [7] providing larger groups within occupational categories. The ESS dataset contains raw responses for the Human Values Scale, which is a 21-item survey instrument designed for self-assessment. In order to compute ground-truth scores from the raw item-level responses, we followed the procedures described in the accompanying manual [26]. Finally, all dependent variables (the ten basic values) were normalized to a range of [0-1] through the following method: $X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$, since it is provides a linear transformation and keeps the relationships among the original data [24].

Table 10.1: List of observable features used as predictors.

	Categorical variable (Yes/No)	Number of categories
Country	Y	23
Gender	Y	2
Age	N	-
Domicile	Y	5
Belonging to religion	Y	2
Belonging to a minority ethnic group	Y	2
Number of people living in the same household	N	-
Living with partner	Y	2
Ever had a divorce	Y	2
Highest level of education	N	-
Employment relation	Y	3
Supervising others at work	Y	2
Type of industry working in (NACE rev.2)	Y	21
Type of organization working for	Y	6

10.4 Results

This section describes the experiments conducted on the ESS dataset and the results obtained from two different types of analytic techniques. All subjects with valid responses on the 14 features were included in the final analyses (n = 36 747): 48.5% of the subjects were males and the mean age of all respondents was 50.41 years (SD = 17.55). Furthermore, the database was randomized and divided into three sets:

- Training set: 60%
- Development set: 20%
- Testing set: 20%

Multiple Linear Regression Approach

Several multiple linear regressions (LRs) were conducted to identify the most suitable set of features that can be utilized for predicting the human value scores based on the observable features presented in Table 10.1. This part of the analysis was conducted using IBM SPSS 25's automatic linear modeling module, which includes supervised merging of the categories, outlier detection, and several feature-selection methods [35]. For each of the ten basic values, the first step involved the assessment of the maximum possible predictive accuracy by using all the features, which aided us in providing an estimate of the highest potential accuracy achievable. Next, predictors were entered into the models using the forward stepwise selection algorithm. At each step, variables not yet included in the model were tested for inclusion until no variables met the inclusion criteria, using a limit of 4 as the maximum number of effects in the final model. This reflects a decision to trade-off a marginal improvement in accuracy for a simpler model with lower costs in terms of data collection. The procedure resulted in two models for each of the ten values, as shown in Table 10.2. Performance was measured by the R^2 (coefficient of determination), ranging between 0-1, which is a well-established, common measure of the success of predicting the dependent variable from the independent variables [23]. Formula: $R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$, where SS_{res} is the sum of the residual squares and SS_{tot} is the total sum of squares. This procedure enabled us to assess the observable feature's utility in terms of predicting the ten basic values, and to identify an optimal set of features that can sufficiently cover all the basic human values considering the added utility of each feature relative to what is already included in the model.

Table 10.2: Statistics of R^2 values for the Linear Regression approach. In the last column, values in parentheses represent the number of features used in the final model.

	Max possible R^2	Final R^2
Achievement	0.23	0.16 (2)
Benevolence	0.22	0.16 (2)
Conformity	0.17	0.11 (2)
Hedonism	0.22	0.18 (2)
Power	0.24	0.18 (1)
Security	0.20	0.12 (3)
Self-Direction	0.16	0.09 (3)
Stimulation	0.16	0.09 (2)
Tradition	0.24	0.14 (4)
Universalism	0.18	0.13 (3)

Figure 10.2 presents each dependent variable with the best set of demographic variables, that account for the largest amount of explained variance (see the 'Final R^2' column from Table 10.2 for the corresponding models). The colored bars represent demographic features that were included in the final models and their length represents the amount of variance explained by the corresponding variable. The white bars represent the amount of unexplained variance for each value, and as such, they express the amount of remaining uncertainty regarding a subject's motivational profile. Figure 10.4 and Figure 10.5 in the Appendix provides the details of all the final regression models for each of the ten values.

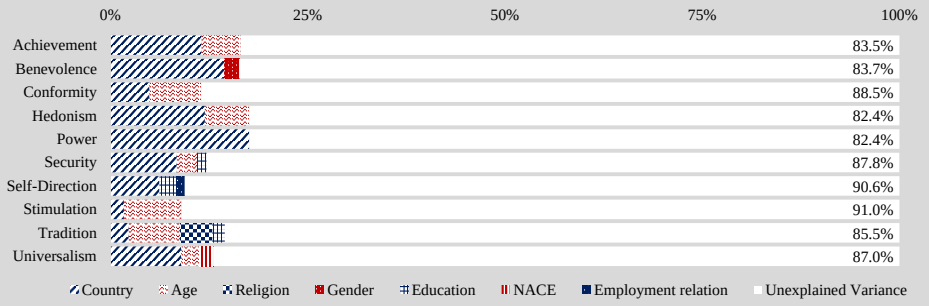


Figure 10.2: Feature importance for predicting the 10 basic human values from observable features by the LR approach relative to unexplained variance expressed in terms of R^2 scores.

Machine Learning Approach

This experiment utilized a machine learning (ML) approach for the prediction of the same set of basic human values. The regression models were trained using the H_2O.ai API, which is an open-source ML platform [15]. The Distributed Random Forest (DRF) regression algorithm was chosen for building models for each of the ten values separately, since the algorithm can properly handle categorical variables with several levels [14], and also provides useful internal estimates of error, correlation, and variable importance metrics [3]. Furthermore, when given a training dataset, the DRF creates a forest of classification (or regression trees) instead of a single tree.

DRF Training

During the training stage, the models were trained using a 5-fold cross validation procedure to obtain the final model of the training set. Table 10.3 presents the mean and the standard deviation of the root-mean square error (RMSE) scores for all of the five folds.

Table 10.3: Mean and SD of RMSE and R^2 for 5 fold cross validation training.

	RMSE		R^2	
Dependent Variable	Mean	SD	Mean	SD
Achievement	0.128	0.0002	0.141	0.0090
Benevolence	0.098	0.0009	0.126	0.0096
Conformity	0.127	0.0005	0.097	0.0041
Hedonism	0.106	0.0005	0.139	0.0134
Power	0.120	0.0004	0.159	0.0033
Security	0.112	0.0004	0.109	0.0095
Self-Direction	0.113	0.0013	0.072	0.0034
Stimulation	0.114	0.0008	0.074	0.0064
Tradition	0.104	0.0006	0.122	0.0092
Universalism	0.102	0.0008	0.106	0.0070

The RMSE scores indicate the absolute fit of the model as it is the square root of the variance of the residuals in the prediction model. As such it is a good measure of the model's predictive accuracy. The RMSE can be interpreted as the standard deviation of the unexplained variance and it has the same unit as the dependent variable [11]. The models were tuned on the hyperparameter 'number of trees' using the development set. The hyperparameter tuning favoured a higher number of trees. However, increasing the number of trees beyond 50 did not result in a significant improvement in terms of the RMSE. Therefore, for all of the ten models, 50 tree-solutions were selected.

DRF Testing

In the testing phase, the accuracy of the trained models was verified using the testing set. Table 10.4 reports the RMSE and R^2 performance metrics for each variable with additional comparisons between random guessing and specifically guessing the mean values for each of the dependent variables. This part of the experiment enabled an assessment of the model's superiority over various types of educated guesses.

Furthermore, Figure 10.3 reports the mean importance of the features across all of the ten basic human values based on the average contribution of each feature to the overall explained variance. Since these scores represent the average contributions across all of the values, it should be noted that certain values can be predicted with higher and lower accuracy, and the cost of obtaining certain demographic features should be considered during data collection.

Table 10.4: RMSE score comparison for each variable between Machine Learning model (ML), Mean Guessing (MG), and random guessing (RG).

Dependent Variable	ML	MG	RG
Achievement	0.1282	0.1376	0.1393
Benevolence	0.0974	0.1046	0.1485
Conformity	0.1267	0.1328	0.1454
Hedonism	0.1056	0.1133	0.1134
Power	0.1195	0.1293	0.1293
Security	0.1134	0.1195	0.1515
Self-Direction	0.1146	0.1180	0.1303
Stimulation	0.1144	0.1182	0.1244
Tradition	0.1031	0.1100	0.1445
Universalism	0.1017	0.1081	0.1086

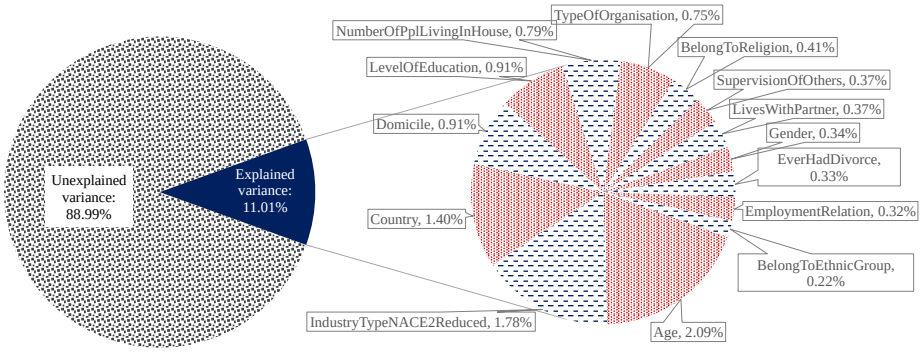


Figure 10.3: Mean feature importance for predicting the 10 basic human values from observable features by ML approach.

Comparison of Approaches

Finally, a comparison between the predictive performance of the two approaches is presented in Table 10.5, across all of the dependent variables in terms of both the R^2 and RMSE scores. Since the interpretation of R^2 scores is relatively straightforward as the percentage of variability explained in the dependent variable by the independent variables, for the purpose of comparison, this measure of goodness of fit is used. In the case of both approaches, the predictability of Power is the highest, implying that Power can be predicted with the highest accuracy from the available set of demographic variables. On the other hand, Self-direction and Stimulation values are at the lowest end of predictability, which indicates that demographic features

are less useful for inferring these particular values. While the LR approach shows slightly better performance than the ML approach in terms of R^2 scores across all of the dependent variables, both data-analytic approaches converge on similar overall results in terms of predictive performance, which further consolidates the findings.

Table 10.5: Predictive performance comparison of machine learning (ML) and linear regression (LR) approaches in terms of R^2 and RMSE scores.

	ML approach		LR approach	
Dependent Variable	R^2	RMSE	R^2	RMSE
Achievement	0.13	0.128	0.16	0.127
Benevolence	0.14	0.097	0.16	0.095
Conformity	0.09	0.127	0.11	0.126
Hedonism	0.12	0.106	0.18	0.104
Power	0.15	0.120	0.18	0.118
Security	0.08	0.113	0.12	0.113
Self-Direction	0.07	0.115	0.09	0.113
Stimulation	0.08	0.114	0.09	0.114
Tradition	0.12	0.103	0.14	0.102
Universalism	0.11	0.102	0.13	0.101

10.5 Discussion

The main objective of this study was to assess the utility of demographic features in predicting stakeholder motivation, operationalized as the basic human values. We have shown through a set of experiments how these observable attributes can be utilized for predicting a subject's motivational profile. The results suggest that the overall predictability of these psychological variables from demographic features is relatively low, but that the usefulness of such assessments is highly dependent on the context in which the results are to be used. In cases where no prior information is available, even a slight reduction in uncertainty can be significant and worth the effort of gathering additional, easily observable features.

A study by Kosinski, Stillwell, and Graepelsing [18] has demonstrated how a set of psychological constructs (the Big 5 traits) can be predicted from online behavioral traces. Firstly, the study showed that certain differences can be expected among the Big 5 traits in their level of predictability: Openness (r = 0.43), Extraversion (r = 0.40), Neuroticism and Agreeableness (r = 0.3), and Conscientiousness (r = 0.29), covering a range between 8.41 and 18.49 in terms of the R^2. Considering

that the present study only relied on demographic features, the level of predictability matched closely, even though behavioral features might convey a lot more information about latent traits. Furthermore, the aforementioned study compared the predictive accuracy obtainable from observable features, to the predictive accuracy achievable by administering the same psychometric instrument for the same respondent at two points in time. The correlation between these scores (test-retest reliability) varies between r = 0.55-0.75, indicating a possible upper bound in terms of the predictability of relatively stable psychological traits by standard, validated instruments.

The experiments conducted with the ML approach established that the model's performance is superior to random guessing, as well as educated guessing (e.g. a guess of the group means), and that the LR approach had a higher level of performance when using different combinations of predictor variables, but also that most of these differences are only marginal. The differences could be attributed to the automated data preparation in the case of the LR approach, which shows the implementation's additional usefulness during the analysis of complex survey data.

In sum, country, age, and type of industry one is working for are the most important features that can be easily obtained and used for the prediction of the majority of basic values from the available set of features included in the ESS dataset. Therefore, identification and inclusion of other demographic features (which might be more difficult to obtain) do not necessarily provide additional predictive utility. This is important knowledge for an analyst when considering the cost-benefit of gathering a greater amount of descriptive data with the intention of achieving higher accuracy. In order to identify potentially more useful predictor variables, further studies will focus on features that reflect previous choices in a subject's history.

Legal and Ethical Considerations

It should be noted that there are important legal and ethical aspects when human subjects are involved both in research and in the real-world application of the described profiling method. For this reason it is necessary to outline and separate the conditions under which the method's application can be considered ethical or legal. While the distinction between law and ethics is often unclear, they are fundamentally different [16]. Both are normative, but ethical norms are formulated as guidelines rather than as prescriptions and prohibitions. Ethics is a collection of fundamental concepts and guidelines that informs individuals about desirable actions in certain situations. Legislation, on the other hand, refers to a systematic body of rules and regulations in written form that aim to govern the behavior of individuals within the boundaries of a particular organization (e.g. country) and unlawful activities are penalized and sanctioned. The difference between ethics and

law is also expressed in the corresponding documents.

Ethical guidelines (e.g. the Guidelines for Research Ethics in the Social Sciences, Humanities, Law, and Theology [16]) developed for conducting research with human participants require: respect for human dignity, privacy, safeguarding against harm, compliance with the duty to inform, and the obtaining of the participant's consent, especially in cases where sensitive personal data is collected. There are also exceptions from the main rule concerning informed consent e.g. observation in public arenas, public figures, if the research does not involve direct contact with the participants, and in cases where information cannot be provided before the research is initiated because it would affect the outcomes of the experiment. These exceptions must be justified by proving the they add value to the research and by demonstrating the lack of alternative options.

Laws vary with time and across territories; therefore, it is crucial to have an up-to-date and contextual understanding of the legal regulations concerning any activity. Different laws have been developed for the collection and protection of personal data across nations. Member states of the European Union (EU) and the European Economic Area (EEA) have opted for an all-encompassing regulation named the European General Data Protection Regulation (GDPR) [6]. The GDPR requires that the processing of personal (linkable to a person) and sensitive data (health, race or ethnic background, sexuality, political, or religious beliefs) should be done with free and informed consent, and that data processors are required to protect the privacy of respondents, and, therefore, ensure confidentiality. A different approach is used by the United States, which implements various sector-specific data protection laws that work together with state level legislation (e.g. HIPAA, NIST 800-171, the Gramm-Leach-Bliley Act, the Federal Information Security Management Act) [4].

The overview on the legal and ethical aspects aimed to highlight some important issues that have to be taken into consideration when it comes to either the development or the application of any profiling method.

10.6 Conclusions

This study aimed at increasing the real-world applicability of the CIRA method that addresses human-related risks within the domain of information security. The method focuses on stakeholder motivation and requires the inference of motivational profiles without direct involvement of the stakeholders. Therefore, we investigated the usefulness of easily observable demographic features for inferring stakeholder motivational profiles. By analyzing a high-quality dataset from representative European samples, and utilizing various data-analytic approaches, we showed that demographic features have some limited usefulness in terms of deriving stakeholder

motivation. While the analysis was limited to respondents from European countries, cultural differences account for the majority of variances explained. In sum, these results are useful for characterizing individuals' motivational profiles especially, when limited access to subjects is assumed, and in cases where subjects might be motivated to answer dishonestly to direct questions. While the primary application of these results is the CIRA method of risk analysis, other domains could benefit from predicting inaccessible subject's motivational profiles, especially where decisions are characterized by trade-offs between various objectives and have great potential impact (e.g. intelligence analysis, operations research, etc.). Future work may expand the analysis to include other regions of the world (e.g. USA, Eastern-cultures) to investigate whether the predictability of value profiles is affected by deeper cultural differences. Finally, these findings provide a solid benchmarking baseline for other future work, which will investigate other classes of observable features for inferring motivational profiles. More specifically, observables that represent the outcome of a conscious decision process (e.g. ownership of items, style, etc.) will be analyzed in terms of their capability to provide insight into the decision-maker's value structure.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions for improving the paper.

This work was partially supported by the project IoTSec – Security in IoT for Smart Grids, with number 248113/O70 part of the IKTPLUSS program funded by the Norwegian Research Council.

10.7 Appendix

	df					*Unstandardized Beta*		
	regression	*residual*	*F*	*adjusted* R^2	*Intercept*	*Country (coded as)*	*Age*	*Gender (coded as)*
Achievement	11	22,036	390.72	0.16	0.52	0.08 Finland (0)		
						0.15 United Kingdom (1)		
						0.15 Lithuania (2)		
						0.14 Netherlands (3)		
						0.01 Sweden (4)		
						0.08 Belgium, Switzerland (5)	-0.002	
						0.06 Spain, Poland (6)		
						0.10 Austria, Estonia, Italy, Russian Federation (7)		
						0.11 Czech Republic, Ireland (8)		
						0[a] Iceland, Norway (9)		
						0.05 Hungary, Slovenia (10)		
Benevolence	13	22,034	329.04	0.16	0.64	-0.10 Finland (0)		-0.03 Male (1)
						-0.04 United Kingdom (1)		0[a] Female (2)
						-0.06 Lithuania (2)		
						0.03 Netherlands (3)		
						-0.12 Sweden (4)		
						-0.10 Belgium, Switzerland (5)		
						-0.05 Spain, Poland (6)		
						-0.03 Austria, Estonia, Italy, Russian Federation (7)		
						-0.02 Czech Republic, Ireland (8)		
						0[a] Iceland, Norway (9)		
						0.01 Hungary, Slovenia (10)		
						-0.08 Germany, France, Israel, Portugal (11)		
Conformity	9	22,038	303.35	0.11	0.51	-0.10 Finland (0)		
						-0.03 United Kingdom (1)		
						-0.07 Lithuania (2)		
						-0.05 Netherlands (3)		
						-0.09 Sweden (4)	0.002	
						-0.06 Belgium, Switzerland (5)		
						-0.08 Spain, Poland (6)		
						-0.11 Austria, Estonia, Italy, Russian Federation (7)		
						0[a] Czech Republic, Ireland (8)		
Hedonism	13	22,034	341.90	0.18	0.60	-0.01 Finland (0)		
						-0.11 United Kingdom (1)		
						-0.07 Lithuania (2)		
						-0.13 Netherlands (3)		
						-0.05 Sweden (4)		
						0.01 Belgium, Switzerland (5)	-0.002	
						-0.06 Spain, Poland (6)		
						0.00 Austria, Estonia, Italy, Russian Federation (7)		
						-0.03 Czech Republic, Ireland (8)		
						0[a] Iceland, Norway (9)		
						-0.04 Hungary, Slovenia (10)		
						0.02 Germany, France, Israel, Portugal (11)		
Power	9	22,038	510.15	0.18	0.46	0.09 Finland (0)		
						0.13 United Kingdom (1)		
						0.05 Lithuania (2)		
						-0.03 Netherlands (3)		
						0.12 Sweden (4)		
						0.02 Belgium, Switzerland (5)		
						0.06 Spain, Poland (6)		
						-0.05 Austria, Estonia, Italy, Russian Federation (7)		
						-0.02 Czech Republic, Ireland (8)		
						0[a] Iceland, Norway (9)		

Note. [a] *reference variable; all SE B < .005; for all included variables p < .05*

Figure 10.4: Final regression models for each dependent variable (1/2).

	df												
	regression	*residual*	*F*	*adjusted* R^2	*Intercept*	*Country (coded as)*	*Age*	*Religion (coded as)*	*Level of education*	*Unstandardized Beta: NACE classification of economic activities (coded as)*		*Employment Relation (coded as)*	
Security	12	272,374	267.72	0.12	0.51	0.03	Finland (0)						
						0.10	United Kingdom (1)						
						-0.02	Lithuania (2)						
						-0.04	Netherlands (3)						
						0.08	Sweden (4)						
						0.08	Belgium, Switzerland (5)	0.001		-0.01			
						0.02	Spain, Poland (6)						
						0.04	Austria, Estonia, Italy, Russian Federation (7)						
						0.05	Czech Republic, Ireland (8)						
						0[a]	Iceland, Norway (9)						
						0.07	Hungary, Slovenia (10)						
Self-Direction	11	22,036	212.69	0.09	0.47	0.01	Finland (0)					0.04	Self-employed (0)
						-0.01	United Kingdom (1)					0[a]	Employee, Working for own family business (1)
						0.03	Lithuania (2)						
						0.06	Netherlands (3)						
						0.06	Sweden (4)			0.01			
						-0.04	Belgium, Switzerland (5)						
						0.04	Spain, Poland (6)						
						0.04	Austria, Estonia, Italy, Russian Federation (7)						
						0.01	Czech Republic, Ireland (8)						
						0[a]	Iceland, Norway (9)						
Stimulation	6	22,041	361.72	0.09	0.56	-0.02	Finland (0)						
						-0.04	United Kingdom (1)						
						-0.03	Lithuania (2)						
						-0.01	Netherlands (3)	-0.002					
						-0.04	Sweden (4)						
						0[a]	Belgium, Switzerland (5)						
Tradition	9	22,038	415.14	0.14	0.54	-0.03	Finland (0)		0.04	Yes (1)			
						-0.05	United Kingdom (1)		0[a]	No (2)			
						-0.05	Lithuania (2)						
						-0.01	Netherlands (3)	0.001			-0.01		
						-0.02	Sweden (4)						
						-0.03	Belgium, Switzerland (5)						
						0[a]	Spain, Poland (6)						
Universalism	16	22,031	269.70	0.13	0.51	-0.02	Finland (0)						
						-0.09	United Kingdom (1)						
						-0.08	Lithuania (2)						
						-0.04	Netherlands (3)						
						-0.05	Sweden (4)						
						-0.04	Belgium, Switzerland (5)	0.001					
						-0.10	Spain, Poland (6)						
						-0.01	Austria, Estonia, Italy, Russian Federation (7)						
						-0.04	Czech Republic, Ireland (8)						
						0[a]	Iceland, Norway (9)						
						-0.07	Hungary, Slovenia (10)						

	0.01 Mining and quarrying (0)
	-0.01 Water supply; sewerage, waste management and remediation activities, Information and communication (2)
	Professional, scientific and technical activities, Administrative and support service activities (2)
	0.02 Agriculture, forestry and fishing, Manufacturing, Accommodation and food service activities (3)
	0.03 Public administration and defence; compulsory social security, Education (4)
	Other service activities, Activities of extraterritorial organisations and bodies (4)
	-0.004 Human health and social work activities, Arts, entertainment and recreation, Activities of households as employers (5)
	0[a] Electricity, gas, steam and air conditioning supply, Wholesale and retail trade; repair of motor vehicles and motorcycles (6)
	Transportation and storage, Financial and insurance activities, Real estate activities (6)

Note. [a] *reference variable; all SE B < .005; for all included variables p < .05*

Figure 10.5: Final regression models for each dependent variable (2/2).

References

[1] Ross Anderson and Tyler Moore. 'Information security: where computer science, economics and psychology meet'. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 367.1898 (2009), pp. 2717–2727.

[2] Jaskiran Arora. 'Corporate governance: a farce at Volkswagen?' In: *The CASE Journal* 13.6 (2017), pp. 685–703.

[3] Leo Breiman. 'Random forests'. In: *Machine learning* 45.1 (2001), pp. 5–32.

[4] Andrada Coos. *EU vs US: How Do Their Data Protection Regulations Square Off?* 2018. URL: https://www.endpointprotector.com/blog/eu-vs-us-how-do-their-data-protection-regulations-square-off/ (visited on 15/09/2018).

[5] Allard E Dembe and Leslie I Boden. 'Moral hazard: a question of morality?' In: *New Solutions: A Journal of Environmental and Occupational Health Policy* 10.3 (2000), pp. 257–279.

[6] European Parliament, Council of the European Union. 'Regulation (EU) 2016/679 of the European Parliament and of the Council (GDPR)'. In: *Official Journal of the European Union* (2016). [Online; accessed 15. Apr. 2020]. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679#d1e40-1-1.

[7] NACE Eurostat. 'Rev. 2–statistical classification of economic activities in the european community'. In: *Office for Official Publications of the European Communities, Luxemburg* (2008).

[8] David L Forbes. 'Toward a unified model of human motivation'. In: *Review of general psychology* 15.2 (2011), p. 85.

[9] Zak Franklin. 'Justice for revenge porn victims: Legal theories to overcome claims of civil immunity by operators of revenge porn websites'. In: *California Law Review* (2014), pp. 1303–1335.

[10] Lewis R Goldberg et al. 'Demographic variables and personality: The effects of gender, age, education, and ethnic/racial status on self-descriptions of personality attributes'. In: *Personality and Individual differences* 24.3 (1998), pp. 393–403.

[11] Karen Grace-Martin. *Assessing the Fit of Regression Models*. [Online; accessed 05-July-2018]. 2008. URL: https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/.

[12] Frank L Greitzer and Ryan E Hohimer. 'Modeling Human Behavior to Anticipate Insider Attacks'. In: *Journal of Strategic Security* 4.2 (2011), pp. 25–48.

[13] Terry M Gudaitis. 'The missing link in information security: Three dimensional profiling'. In: *CyberPsychology & Behavior* 1.4 (1998), pp. 321–340.

[14] H2O.ai. *Distributed Random Forest (DRF)*. June 2018. URL: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html.

[15] H2O.ai. *H2O.ai*. June 2018. URL: https://www.h2o.ai/.

[16] Bjørn Hvinden et al. *Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theology*. 2016.

[17] Ariel Knafo and Lilach Sagiv. 'Values and work environment: Mapping 32 occupations'. In: *European Journal of Psychology of Education* 19.3 (2004), pp. 255–273.

[18] Michal Kosinski, David Stillwell and Thore Graepel. 'Private traits and attributes are predictable from digital records of human behavior'. In: *Proceedings of the National Academy of Sciences* 110.15 (2013), pp. 5802–5805.

[19] Brian Krebs. 'FBI: Smart Meter Hacks Likely to Spread'. In: (2012). [Online; accessed 26-June-2018]. URL: https://krebsonsecurity.com/2012/04/fbi-smart-meter-hacks-likely-to-spread.

[20] Brian W Kulik, Michael J O'Fallon and Manjula S Salimath. 'Do competitive environments lead to the rise and spread of unethical behavior? Parallels from Enron'. In: *Journal of Business Ethics* 83.4 (2008), pp. 703–723.

[21] Stan J Liebowitz and Stephen E Margolis. 'Network externality: An uncommon tragedy'. In: *Journal of economic perspectives* 8.2 (1994), pp. 133–150.

[22] N.A. 'European Social Survey Round 8 Data'. In: *NSD - Norwegian Centre for Research Data, Norway - Data Archive and distributor of ESS data for ESS ERIC* (2018).

[23] Nico JD Nagelkerke et al. 'A note on a general definition of the coefficient of determination'. In: *Biometrika* 78.3 (1991), pp. 691–692.

[24] S Patro and Kishore Kumar Sahu. 'Normalization: A preprocessing stage'. In: *arXiv preprint arXiv:1503.06462* (2015).

[25] Lisa Rajbhandari and Einar Snekkenes. 'Using the Conflicting Incentives Risk Analysis Method'. In: *Security and Privacy Protection in Information Processing Systems*. Ed. by Lech J. Janczewski, Henry B. Wolfe and Sujeet Shenoi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 315–329.

[26] Shalom Schwartz. *Computing Scores for the 10 Human values*. [Online; accessed 12-November-2019]. 2016. URL: https://www.europeansocialsurvey.org/docs/methodology/ESS1_human_values_scale.pdf.

[27] Shalom H Schwartz. 'An overview of the Schwartz theory of basic values'. In: *Online readings in Psychology and Culture* 2.1 (2012), pp. 2307–0919.

[28] Shalom H Schwartz. 'Are there universal aspects in the structure and contents of human values?' In: *Journal of social issues* 50.4 (1994), pp. 19–45.

[29] Shalom H Schwartz. 'Basic human values: Theory, measurement and applications'. In: *Revue française de sociologie* 47.4 (2007), p. 929.

[30] Shalom H Schwartz. 'Culture matters: National value cultures, sources, and consequences'. In: *Understanding Culture*. Psychology Press, 2013, pp. 137–160.

[31] Shalom H Schwartz. 'Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries'. In: *Advances in experimental social psychology*. Vol. 25. Elsevier, 1992, pp. 1–65.

[32] Shalom H Schwartz and Tammy Rubel. 'Sex differences in value priorities: Cross-cultural and multimethod studies'. In: *Journal of personality and social psychology* 89.6 (2005), p. 1010.

[33] Einar Snekkenes. 'Position paper: Privacy risk analysis is about understanding conflicting incentives'. In: *IFIP Working Conference on Policies and Research in Identity Management*. Springer. 2013, pp. 100–103.

[34] Adam Szekeres and Einar Arthur Snekkenes. 'Unobtrusive Psychological Profiling for Risk Analysis'. In: *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications - Volume 1: SECRYPT*. INSTICC. SciTePress, 2018, pp. 210–220.

[35] Hongwei Yang. 'The case for being automatic: introducing the automatic linear modeling (LINEAR) procedure in SPSS statistics'. In: *Multiple Linear Regression Viewpoints* 39.2 (2013), pp. 27–37.

Chapter 11

Article 3: Construction of Human Motivational Profiles by Observation for Risk Analysis

Adam Szekeres, & Einar Arthur Snekkenes. Construction of Human Motivational Profiles by Observation for Risk Analysis[1]. In: *IEEE Access, Vol. 8*. IEEE. 2020, pp. 45096–45107.

Abstract

This study aimed at analyzing the extent to which publicly observable pieces of information representing stakeholders' past and current choices can be utilized for the construction of motivational profiles. Motivation is operationalized by the theory of Basic Human Values, which organizes 10 values capturing distinct aspects of human motivation into a hierarchical order. The construction of motivational profiles for individual stakeholders is motivated by the need to enhance the existing decision-maker model in the Conflicting Incentives Risk Analysis (CIRA) method. This study utilized an online questionnaire to collect responses from participants (n = 331) about a wide range of habits and personal items that are easily observable in various contexts by an analyst. The validity of the set of observables as surrogate predictors of the motivational profiles is evaluated by various methods (i.e. comparison to previous results, cross-validation of models, comparison to test-retest reliability of the psychometric instrument) and techniques (calculation of

[1]This work was partially supported by the project IoTSec – Security in IoT for Smart Grids, with number 248113/O70 part of the IKTPLUSS program funded by the Norwegian Research Council.

prediction interval for individual profile scores). The assessment of the uncertainties associated with predicting motivational profiles is explored in detail. Additionally, an example illustrates how the profiles can be utilized for the assessment of action desirability (i.e. prediction of behavior) based on the utility calculations established in CIRA. The results contribute to an improved understanding about the accuracy with which human stakeholder motivation can be inferred from public observables and utilized within the context of information security risk analysis.

11.1 Introduction

Increasing levels of digitization affect more and more sectors, as well as critical infrastructures. A prominent emerging example is the Smart Grid, which represents the augmentation of the traditional electric grid with Internet of Things (IoT) devices enabling several desirable properties for various stakeholders, such as enhanced monitoring and control capabilities, the potential for more sustainable and eco-friendly energy utilization, new business opportunities, etc. [7]. The envisaged benefits can materialize given that the potential downsides introduced with novel technologies remain under control, and risks are mitigated. A complex system such as the Smart Grid has an inflated surface for cyber-attacks [9], and potential threats to privacy are increased [17]; thus, national security may be at risk when international conflict permeates to critical infrastructures [16].

Any stakeholder connected to the electric grid may be interested in getting answers to questions relating to their level of risk as a result of the actions or inactions of other parties. Homeowners may be interested in the privacy risks they face as owners of IoT smart appliances when service providers and manufacturers decide to merge [40]. Is it possible that threats to end-users' privacy observed in other sectors (e.g. toll booth use [38], health care [27]) are transferable to the Smart Grid when the need to process huge amounts of information at a low cost motivates companies to engage in outsourcing, which may expose millions of citizen records to parties whose interests may be difficult to monitor. Analysis of consumption data from Smart Meters enables profiling that can be used to identify unique devices used in the household, to reveal the number of occupants and other sensitive information not previously available from these sources of data. Such datasets are of great potential utility not only for the electricity provider, but other third parties (e.g. insurance companies, entertainment companies, and government authorities) [18]. Are the proper regulations in place, and are they enforced so that they prevent electricity companies from abusing their newly gained insights? Is it reasonable to assume that the information provided to prospective customers about the details of their contract is valid and reliable (i.e. integrity of information is appropriate)? Misinformation or deliberate withholding of pieces of key information can have

a negative impact on the organization, when misbehavior is revealed [24]. What are the key factors that motivate relevant decision-makers in an organization to invest scarce resources (e.g. time and money) to ensure that customer information is securely transmitted, processed, stored and erased throughout the entire lifecycle of the data [15]? From a national security perspective, it is important to understand whether all the stakeholders responsible for maintaining and developing the Smart Grid of the future act in accordance with national interests.

In order to answer such questions in highly complex systems, the Conflicting Incentives Risk Analysis (CIRA) method proposes a way to model risks in a novel way by focusing on the motivation of the relevant stakeholders within the scope of the analysis [25]. The method requires that stakeholder motivational profiles be constructed without direct interaction with the subjects (i.e. using unobtrusive measures). Therefore, the main objective and **research problem** of this paper is to explore the extent to which publicly observable pieces of information can be utilized for building individual motivational profiles. For the construction of the motivational profiles, this study focuses on two distinct types of information that are assumed to be easily available in any context and can be assessed with a high accuracy simply by observation of subjects:

- evidence of conscious choices from the past of the stakeholder (i.e. ownership of various items, buying decisions) and
- habits and activities in the present.

There is a growing collection of work that demonstrates how various personality features can be predicted from different behavioral traces: intelligence and Big 5 traits from Facebook likes [14], Big 5 traits from mobile phone use data [8], and so on; for an extensive review, see [35]. While these methods are unobtrusive in the sense that they do not rely on direct interaction with the subject, they are highly obtrusive since they require access to sensitive personal account information or behavioral characteristics available only in settings where a subject explicitly gives permission to an application or other data collecting service, which can be used to amass a vast amount of information about the subjects. It is, however, unreasonable to assume that such sources of information will be available in common risk analysis settings. Furthermore, respondents would not be legally obliged to provide such account information for the purpose of the analysis. Therefore, the utility of the previously mentioned unobtrusive profiling methods is highly limited for the purpose of a real-world risk analysis, since the methods require full access to devices and/or accounts and are dependent on specific services. The present analysis focuses on features that are independent of service providers and thus

aims for a wider range of applicability. The following **research questions** were formulated to address the overall research problem:

- **RQ 1:** How well can observable features predict stakeholder motivational profiles operationalized as the Basic Human Values?
- **RQ 2:** How much improvement can be expected from the present set of observable features compared to analyses using demographic features?

The paper is organized as follows. Section 11.2 presents a set of the most widely adopted information security risk assessment (ISRA) methods, which include human-related risks in the risk assessment procedure; describes key features of the CIRA method; and presents the theory used to operationalize human motivation. Section 11.3 describes the research method, including the data collection procedure and the instruments used. Section 11.4 presents the key findings, Section 11.5 discusses the relevance of the findings in the context of risk analysis, as well as the limitations and plans for further work. Finally, Section 11.6 provides a summary of the work.

11.2 Related work

Several risk analysis methods exist, but few address human motivation in detail. This section provides and overview of existing approaches for addressing human-related risks within information security as implemented in various ISRA methods. The primary resource for this overview is provided by [39], in which the most relevant ISRA methods are analyzed in detail, and the Core Unified Risk Framework is developed to aid practitioners in selecting the most appropriate method for the task at hand. The framework contains a total of nine ISRA methods, along with privacy and cloud risk assessment methods. The following overview focuses on a subset of the methods, including a discussion of human threats. Four methods were excluded from the present overview, due to their incompleteness on the following attributes according to the framework: threat willingness/motivation, threat capability, and threat capacity. Furthermore, the Risikovurdering av informasjonssystem (RAIS) method was also excluded due to its obsolescence and unavailability in English. A short summary of the strengths and weaknesses of the reviewed methods is provided in Table 11.1.

The **ISO 27005:2011** is one of the most widely used risk management frameworks, which in Annex C lists various threats that can guide an analyst through the threat assessment process [10]. Each *threat* is classified into one or more of the following groups: *accidental, deliberate*, and *environmental*. An additional table organizes

Table 11.1: Comparison of a representative set of ISRA methods with respect to their capability of dealing with human threats.

No.	Method	Pros	Cons
1.	ISO 27005:2011	Wide acceptance in community. List of most relevant human threat groups. List of attributes associated with each identified group.	Lack of guidelines on how to assess the attributes, how to derive valid probabilities, consequences, etc.
2.	FAIR	Based on solid quantitative theories and methodology. The risk landscape development is supported by software tools.	Difficult to check that the obtained parameters are correct. Extensive training is required to apply the method.
3.	OCTAVE-Allegro	Suitable for preliminary assessments especially when resources for conducting a risk analysis are limited. Requires no expert knowledge. Variety of supporting tools.	Lack of quantitative rigour. No systematic way for threat discovery. No guidance on mitigation strategies against human threats.
4.	NIST 800-30	Threat characteristics: adversary intent; adversary capability; adversary targeting. Provides human threat sources in a taxonomy. Designed for the needs of federal information systems.	Unclear how assumptions about threat characteristics could be verified. Potential for generating unmanageable amounts of threat scenarios.
5.	COBIT5/RISK IT	Emphasis on aligning business objectives and risk analysis objectives through a focus on critical assets. Basic classification of threat types including malicious, accidental human threats.	Lack of instructions about the procedures for assessing human-related risks.
6.	CORAS	Risk analysis aided by graphical representations, focus on re-usability of previous results. Input from various stakeholders with different knowledge and experience.	Success of risk assessment largely depends on subjective evaluations, and on the experience of participants of brainstorming sessions. Elementary conceptualization of human threats.
7.	CIRA	Suitable for emerging systems (without historical data). Addresses Opportunity Risk. Redefines risk as the misalignment between stakeholder motivations.	Requires enhancement to enable real-world applicability by characterizing stakeholders. Lacks validation in real settings.

human-related threats into five main groups by their origins (*hacker, computer criminal, terrorist, industrial espionage*, and *insiders*). Each group has an associated list of motivations, and the possible consequences of these threats are enumerated. Annex D also mentions several human-related vulnerabilities that span across issues related to personnel (e.g. lack of security awareness), organizational vulnerabilities (e.g lack of continuity plans), hardware and software (e.g. complicated user interfaces). The framework produces a risk matrix for further decision-making, which can be constructed by combining subjective and empirical measures, that fit well with the organization's objectives and available resources. In sum, the framework calls the analyst's attention to several human threats and provides general outlines about issues that should be considered during threat identification and vulnerability assessment, which can be useful during a high-level initial risk identification phase. However, the analyst is not provided with specific details about the complex motivational and cognitive processes that result in overt behavior. Since an analyst may have to resort to guesswork regarding human threats, the risk assessment procedure could result in ignoring or miscalculating human-related threats and risks.

The simulation-based Factor Analysis of Information Risk **(FAIR)** method was developed to measure and represent information security risk using quantitative methods and statistically sound mathematical calculations. Salient objects are identified in the environment, their characteristics are defined, and their interactions are modeled. The end result can be an integer, a distribution that represents the risk to information security. "Information risk occurs at the intersection of two

probabilities-the probability that an action will occur that has the potential to inflict harm on an asset, and the probable loss associated with the harmful event" [13]. A taxonomy for information risk includes elemental components (objects) that make up the information risk landscape, a set of variables that describe the characteristics of objects, a decomposition of the factors that drive information risk, and a description of the relationships between the factors. Humans are a type of object within the framework, and *threat agents* are special objects that can be categorized as: *humans, animals, environmental elements, and human-made objects*. Threat agents, which have the ability or tendency to inflict harm upon other objects, are characterized by a unique set of characteristics that captures a certain level of psychological realism, including *skill, knowledge, experience, resources, risk tolerance, primary and secondary motives, and intents*. A *threat community* provides a description for a set of threat agents that share some common characteristics. It is useful for defining threat agent characteristics based on group membership when individuals are unknown. FAIR takes the perspective of the threat agent when considering the value of the object, the vulnerability of the object and the level of risk to the threat agent with negative consequences. These considerations are included based on the specific threat community characteristics. To measure *threat capability*, a scale is constructed by combining three factors: knowledge, experience and resources. In sum, the method models human agents as a group within a threat community, where the parameters related to the specific threat community are *volume, activity level, capability, risk tolerance, selectiveness, primary intent, and secondary intent*.

The Operationally Critical Threat, Asset, and Vulnerability Evaluation **(OCTAVE-Allegro)** method was designed to optimize information security risk assessments, considering limited resources for the task. The methodology guides the analyst through the process by considering how people and technology contribute to business processes they support [2]. Several OCTAVE variants have been developed for the needs of organizations of various sizes. All variants aim at developing qualitative risk evaluation criteria, identifying assets that are crucial to the goals of the organization, identifying vulnerabilities and threats to those assets, and evaluating potential consequences to the organization if identified threats are realized. Some variants are based on workshops with interdisciplinary analyst teams or field experts. Allegro is specifically designed to guide risk assessment without extensive expert knowledge. The methodology is supported by worksheets and questionnaires. Human behavior is addressed in the *threat scenario identification* process, distinguishing between accidental and deliberate actions by human actors. Threats have the following properties: *asset, access, actor* (person who may violate security requirements), *motive* (intention of the actor), and *outcome*. Actors are further categorized according to their position as *inside, outside*. Threat identification largely

depends on incidental background knowledge (e.g. "John is the only employee who knows the production specs for producing widgets and he has been talking about leaving the company; if he does so, and the widget specs aren't obtained, we can't make widgets" [2]), which is brought to the analyst's attention by the use of threat scenario questionnaires. The scenario-based qualitative threat identification can be useful to highlight important aspects where more investigation is needed, but largely depends on the analyst's creativity or motivation and available resources to distinguish between realistic and unlikely threat events.

The **NIST 800-30** method developed by the National Institute of Standards and Technology of the U.S. Department of Commerce [12] was designed to assist with conducting risk assessments for federal information systems and organizations, with the aim of providing senior executives the information needed to determine appropriate courses of action when considering the identified risks. The impact of human behavior is discussed in *threat sources* (characterized by intent and method targeted at the exploitation of a vulnerability), which enables the development of corresponding *threat scenarios*. Types of threat sources may be: hostile cyber or physical attacks; human errors of omission or commission; structural failures of organization-controlled resources; and natural and man-made disasters, accidents, and failures beyond the control of the organization. Furthermore, when discussing vulnerabilities in the broader context, various types of vulnerabilities are enumerated that can be linked to human behavior (e.g. lack of effective risk management strategies and adequate risk framing; poor intra-agency communications; inconsistent decisions; misalignment of enterprise architecture to support mission/business activities; external relationships, such as dependencies on particular energy sources, supply chains, information technologies, and telecommunications providers, etc.). The assessment of incident likelihood for adversarial threats is based on: adversary intent, adversary capability, adversary targeting. Table D-2 in the Appendix provides a detailed taxonomy of adversarial Threat Sources at various levels (individual, group, organization, and nation-state), with further distinctions, such as *outsider, insider, trusted insider*, etc. Tables D3 to D5 describe relevant adversarial features (i.e. *capability, intent, targeting*), with descriptions of the meanings of the accompanying qualitative values (very low to very high). A lack of further guidance on what evidence is needed to support the adversarial models could hinder the effective assessment of human-related threats, and may turn risk assessment into an ad-hoc exercise, without empirical evidence supporting the assumptions. The volume of Threat Scenarios that can be potentially generated from the checklists may further complicate the execution of successful risk assessments.

The **COBIT5/RISK IT** framework developed by ISACA describes a process model for the management of information technology-related risk [11]. The document

emphasizes that risk management-related processes need to be connected to overall business objectives, and communication between stakeholders should be a continuous process. During risk-management activities the guidelines propose the development and use of risk scenarios to help overcome the challenges associated with identifying important and relevant risks amongst all that can possibly go wrong with the IT infrastructure. The scenarios are used during the risk analysis, where the frequency of a scenario actually happening and business impacts are estimated. Scenarios should contain *actors* (internal or external), *threat type* (malicious or accidental), *event* (e.g. disclosure of confidential information), *asset* (impacted by the event leading to business impact), and *timing*. The Risk IT framework is mainly focused on the management and governance perspective, and no further details are provided about how to assess human-related risks.

The graphical or model-based method for security risk analysis **CORAS**, was developed in order to provide a method that facilitates risk analysis by making previous results easily accessible and maintainable [4]. The method comprises a specific risk-modeling language, the step-by-step description of the risk-analysis process, and a software tool for documenting and maintaining the results of the analysis. It is based on meetings and workshops between relevant stakeholders and the analyst. Risks are identified through a process called structured brainstorming, where participants with different competences and backgrounds provide their perspectives on the target of analysis. The outputs of the brainstorming activities are threat diagrams where human threats (accidental or deliberate) are linked to vulnerabilities, threat scenarios and incidents. The next step focuses on risk estimation, in which participants provide likelihood estimates and consequence estimations for each threat scenario in the threat diagrams. For scenarios with difficult-to-estimate likelihoods, the analysis leader gives suggestions based on historical data, like security incident statistics or personal experience. Thus, risk assessment largely depends on subjective evaluations and on the breadth of knowledge possessed by the stakeholders invited to the brain-storming sessions.

In order to reduce the complexities associated with conducting the aforementioned risk-analysis methods, and to overcome some of their limitations, the Conflicting Incentives Risk Analysis **(CIRA)** method proposes a novel way to describe the risk situation. CIRA develops a different concept of risk, where risk is the result of misaligned stakeholder incentives [26] and risk is analyzed from the perspective of an individual facing a risk. Two types of risks are distinguished: **threat risk** -undesirable events- and **opportunity risk** - desirable events not realized. To characterize these risks, CIRA requires the identification of two classes of stakeholders: the **risk owner** and the **strategy owner**. Each stakeholder is characterized by a set of utility factors, which capture important aspects of their overall utility (e.g.

wealth, health, security, etc.). Actions available for the strategy owner can have a positive or negative impact on the risk owner's utility factors. Each action has a level of (un)desirability for the strategy owner, which (de)motivates the selection of that particular strategy (e.g. the prospect of a monetary reward). The analysis therefore requires a detailed description of the dependencies between stakeholders, their motivational profiles and the inference of perceived gains and losses from the perspective of the strategy owner to enable the assessment of action-desirability in order to completely characterize the risk situation. The method's real-world applicability is currently limited by the lack of procedures and guidelines for assessing stakeholder motivations in a reliable and valid manner. Consequently, the method requires extensive validation to gain acceptance in the professional community.

Based on the surveyed ISRA methods, it can be concluded that the impact of human behavior on the security of systems is widely recognized, as demonstrated by the inclusion of the issue in most well-established ISRA methods. However, a valid assessment and characterization of human-related risks is, to a great extent, missing in existing approaches. Moreover, most ISRA methods do not place human behavior at the center of the analysis. These issues may be attributed to the difficulties with operationalizing and measuring motivation, intention, capability and probability of goal execution (e.g. adversarial), and so on. While data may be abundant about potential internal threat actors through logging of various activities, the extent to which they are indicative of threat realization may remain unexplored. Additionally, most of the reviewed ISRA methods do not employ an interdisciplinary approach when investigating crucial aspects of human behavior when formulating risk estimations. This gap is partially addressed in this work by investigating how a specific motivational theory from psychology can be utilized to enhance the human model of the CIRA method, which exclusively focuses on deliberate (motivated) human behavior when addressing risks. Thus, the enhanced method could complement other methods with a more valid and practical characterization of human behavior supported by empirical data.

Capturing motivation: Theory of Basic Human Values

The theory of basic human values identifies 10 distinct values that serve as guiding principles throughout people's lives [33]. The values included in the theory are cross-culturally recognized, capture distinct aspects of human motivation, and each refers to desirable end goals that people strive for. The theory proposes a dynamic relationship among the values, which form a circular structure presented in Figure 11.1. When making a decision, values that are on opposite sides of the circle tend to conflict with each other, while adjacent values are more compatible, giving indications about the trade-offs a decision-maker may be willing to make. Thus, key decisions can be represented by combining an individual's value hierarchy with

the expected outcome of an action (i.e. is the duty to increase shareholder value for a CEO, more important than the desire to act lawfully?). Individuals and groups can be meaningfully characterized by their value hierarchies. Values possess the following six core features: 1. "Values are beliefs linked inextricably to affect. 2. Values refer to desirable goals that motivate action. 3. Values transcend specific actions and situations. 4. Values serve as standards or criteria. 5. Values are ordered by importance relative to one another. 6. The relative importance of multiple values guides action." [32]. Previous work has utilized the theory of Basic Human Values for operationalizing a Strategy Owner's motivation to enhance CIRA's real-world applicability [36].

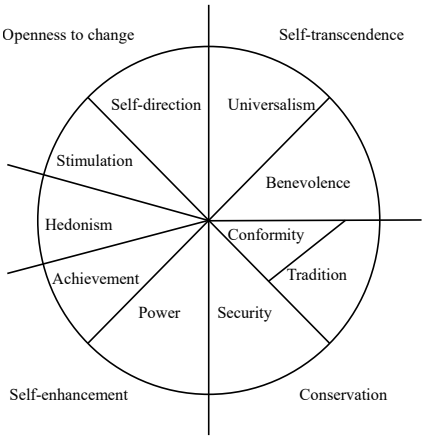


Figure 11.1: 10 basic human values with 4 higher dimensions forming a circular structure. Source: [32].

11.3 Materials and Methods

This section provides a detailed description about the data collection procedures, the sample and the instruments utilized for gathering the necessary information from respondents to address the research questions of the study.

Sample and Procedure

In order to reach a varied pool of respondents from the general population at a working age (above 18 years), a call for participation was distributed on several online channels. As the main objective of the study was to assess the utility of publicly observable pieces of information for the construction of stakeholder motivation profiles, the following channels were used for participant recruitment to ensure a wide coverage of respondents with various backgrounds: A pilot study was conducted on Amazon Mechanical Turk to test the feasibility of data collection

on the popular crowdsourcing platform; however it was assessed as inappropriate for the purpose of the present study since the majority of respondents are located in the U.S. (75%) and India (16%), working below median household incomes in the respective countries [5], which could hamper the transferability of conclusions to a different population (while not necessarily constraining model validation). Based on the results of the pilot study, modifications were implemented and links to the updated version of the survey were distributed on university and project-related mailing lists and on social media platforms. The survey was available in English and Norwegian, and the Norwegian translation was proof-read and finalized by a professional proofreading service. The data collection was open for a total of 114 days. The survey was implemented using the open-source Limesurvey tool and was hosted on internal servers provided by the university. The number of fully completed surveys is presented in Table 11.2, organized by the channels of recruitment.

Table 11.2: Number of completed surveys by distribution channels.

Distribution channel	Number of completed surveys
AmazonTurk	9
QR-invite	1
Social media	24
University e-mail list	332
Total	366

The validity of the final dataset was increased by removing responses below 10 minutes of completion time, which would indicate that respondents were not following the instructions carefully (estimated completion time: 20-30 minutes). Furthermore, extreme outliers with values exceeding three times the height of the boxes (25th-75th percentile) on each dependent variable's boxplot were identified and removed. The final sample **(n = 331)** consists of 173 males, 153 females, and 5 respondents with unspecified sex. The mean age is 40.28 years (SD = 13.27). Additional demographic descriptions of the sample are provided in Table 11.3. To compare the present sample to the general working-age population the information was obtained from the website of Statistisk Sentralbyrå (Statistics Norway) [20]. Compared to the general working-age population, in the present sample: males are slightly over-represented ($\approx$ 50% in the population vs. $\approx$ 52% in the sample), foreign citizens are over-represented ($\approx$ 11% in the population vs. $\approx$ 25% in the sample), the level of attained education is higher in the sample (tertiary: $\approx$ 37% in the population vs. $\approx$ 68% in the sample), and PhDs earned was higher ($\approx$ 1% in the population vs $\approx$ 23% in the sample). The ratio of employed/unemployed respondents is similar to the ratio found in the population considering the active

workforce ($\approx$ 3.7% in the population vs. $\approx$ 3.9% in the sample).

Table 11.3: Basic demographic description of the sample.

Highest level of education completed	**n**	**Employment status**	**n**
Secondary school	26	Employed for wages	292
Bachelor's	53	Self-employed or homemaker	3
Master's	174	Student	23
PhD	78	Currently not in work	13
Citizenship		**Type of industry**	
Norwegian	247	Information and communication	26
Other	84	Professional, scientific and technical activities	146
Marital status		Public administration and defence; compulsory social security	33
Single	79	Education	78
Married or in a long-term relationship	235	Human health and social work activities	14
Divorced or separated	17	Other	34

Measures

Motivational Profile - PVQ-21

Motivational hierarchy was assessed using the 21-item Portrait Value Questionnaire (PVQ). The PVQ was designed to measure the 10 basic value orientations and presents respondents with concrete and cognitively less-demanding tasks than previous instruments designed for measuring value structures. This makes the scale suitable for all segments of the population [31]. The PVQ includes short verbal descriptions of people with their goals and aspirations without explicitly identifying the values under investigation. Respondents answer by judging their own similarity to the portraits, and similarity judgments are transformed into a six-point numerical scale (reverse coded from the original as follows: 1 *(not like me at all)* to 6 *(very much like me)).* The PVQ's adequacy for measuring value structures is supported by adequate psychometric properties based on studies in several countries, and it is suitable for various forms of administration (e.g. face-to-face, by telephone, and online). As individuals may differ in their use of the response scale, centered scores were computed to correct for individual differences in response scale use, thus reflecting the relative importance of each value in the value system [30]. The original English version and the Norwegian version from the European Social Survey was used in this survey [21].

Everyday Choices and Habits Questionnaire

The next section of the survey collected information about various items and habits that are publicly observable. The aim of this part of the survey was to cover a wide range of items that can be observed in any situation without interaction with a respondent. Assessment of item ownership requires a single observation, while the assessment of habits may require observation over a longer period. The list of categories and the number of attributes collected per category are presented in Table 11.4. Questions designed to assess habits asked respondents to report the approximate frequency of the activity for the last year. Other questions used yes/no questions, numerical input, or a single-choice format. The PDF version of the survey (in English) is available as supplementary material. Note that some differences between the original online version of the survey and the PDF version may exist as a result of exporting and converting it into a different format.

Table 11.4: Categories of publicly observable pieces of information collected from respondents, with number of attributes per category.

Ownership		**Habits**	
Home	4	ConsumptionPreferences	17
MeansOfTransport	23	FreeTimeActivities	26
ITdevices	21	Style*	5
Accessories	14	DietChoice*	1
Pets	6	SportsActivities	17
Tattoo	8	MusicPreferences	14
SocialMediaPresence	11	ClothingChoices	23
Jewelery	11	**BasicDemographics**	8
SportEquipments	16		

* Corresponding questions were not formulated to assess frequency of activity.

11.4 Results

The dependent variables (DV) of interest are the 10 Basic Human Values, ground truth scores collected by the PVQ-21 instrument. Data on a total of 225 independent variables were collected, which resulted in 437 variables after categorical (i.e. nominal) variables were recoded into indicator variables (where 0 = no/attribute is not present for the respondent; 1 = yes/attribute is present). This procedure is recommended so that categorical variables with several levels can be included in regression models. Reliability of the instrument was tested through the internal consistency measure (Cronbach's alpha), by analyzing all items that measure the same value. Cronbach's alpha measures the extent to which certain items of a test measure the same construct by analyzing the inter-relatedness of the items

[37]. The analyses provided the following Cronbach's alpha scores for the 10 values: Conformity: .60, Tradition: .67, Benevolence: .56, Universalism: .52, Self-Direction: .36, Stimulation: .72, Hedonism .69, Achievement: .74, Power: .36, Security: .47. These results are similar to the reliability scores found in various nations [31]. It should be noted that the low alpha scores obtained in this and other studies (using the same instrument) may be attributed to the small number of questions (two or three for each dimension) measuring the same construct, which can decrease alpha scores [37]. By convention, alpha scores above .70 are preferred; however, there are no gold-standard levels of alpha, so even lower scores (.50) may be useful [28]. All the analyses were conducted using SPSS 25 by IBM, and scikit-learn, which is a free machine learning library for Python.

Feature selection and comparison with previous results

Following data pre-processing for each DV (10), several multiple linear regression models were built using the stepwise feature selection method in SPSS. This method searches among all independent variables that are not yet in the equation for the one which has the smallest probability of F ("The F-value is equivalent to the square root of the Student's t-value, expressing how different two data samples are, where one sample includes the variable and the other sample does not" [23]), and enters them into the equation if the inclusion criterion is met (p of entry set to < 0.05). Predictors in the regression equation were removed when their probability of F reached the criterion of exclusion (p of exclusion set to ≥ 0.1). The method stops when no predictor meets the inclusion/exclusion criteria. By tuning the exclusion and inclusion criteria, it is possible to control the final model's complexity. The procedure resulted in several models with increasing numbers of predictors and increasing levels of goodness of fit (R^2) associated with each model. The final set of predictors to be evaluated in the following step was selected from the model with the highest R^2 metric for each DV. R^2, or the coefficient of determination is calculated as $R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$, where SS_{res} is the sum of the residual squares and SS_{tot} is the total sum of squares, ranging between negative infinity and +1, which is a measure to assess the model's goodness of fit [22]. Table 11.5 summarizes details of the multiple linear regression models for each dependent variable. Adjusted R^2 scores represent a modified version of the R^2, which increases only when the additional terms improve the model more than expected by chance. Due to penalizing additional predictors the adjusted R^2 scores are always lower than corresponding R^2 scores for the same model. F-scores represent each model's improvement compared to the intercept-only models; df (degrees of freedom) signifies the number of predictors in each model.

Figure 11.2 reports the performance of each model (red bars). Grey bars represent a the predictive utility of demographic features for building motivational profiles

Table 11.5: Summary of multiple linear regression models for each dependent variable.

	R^2	Adjusted R^2	F	df
Conformity	0.33**	0.29	311	19
Tradition	0.44**	0.39	303	27
Benevolence	0.36**	0.32	308	22
Universalism	0.40**	0.35	306	24
Self-Direction	0.60**	0.54	282	48
Stimulation	0.40**	0.36	309	21
Hedonism	0.49**	0.44	299	31
Achievement	0.48**	0.42	295	35
Power	0.47**	0.42	300	30
Security	0.40**	0.35	303	27
$^{**}p < 0.01$				

established on the European Social Survey (ESS) dataset [36]. Differences between red and grey bars indicate the improvement between reported metrics from the two datasets. Across all DVs an average of 3.4-fold improvement is achieved by using the present class of predictor variables based on the R^2 metrics. Improvement for each DV was calculated as: $(R^2{}_{current}/R^2{}_{ESS})$, with AverageImprovement = Sum of improvements for each DV/10.

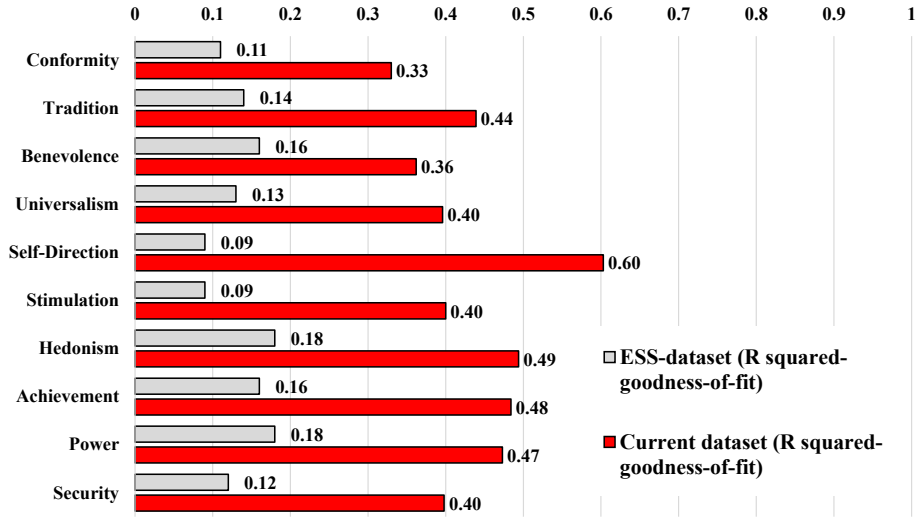


Figure 11.2: Prediction accuracy of Basic Human Values in terms of the R^2 metric. Red bars represent the maximum accuracy achieved after the models were built with the Stepwise feature selection algorithm in SPSS. Grey bars show the goodness of fit metrics for the same variables using demographic features from [36].

Model validation: expected performance on unseen data

The train-split re-sampling method was used to assess the proposed predictors' usefulness for predicting unobserved data points. The next set of experiments aimed at establishing the reliability of the regression models. This enabled the assessment of the model's performance on unseen data. Common practice is to evaluate the model, using only the goodness-of-fit metric; however, this generally leads to over-fitting, and cross-validation is rarely conducted in social science research [41]. "Stepwise regression and all subset regression are in-sample methods to assess and tune models. This means the model selection is possibly subject to overfitting and may not perform as well when applied to new data." [1]. In order to assess the model's predictive performance on previously unseen data, various validation techniques can be used. Due to the small sample size, validation was achieved by conducting several train-test split validations, which is a form of validation with replacement, where the model is trained on a random 80% partition of the dataset, and the predictive performance is evaluated on the remaining 20% that was not utilized for model training. This procedure was repeated 100 times to assess the overall performance more accurately. Figure 11.3 reports the goodness of fit metrics for each dependent variable in terms of R^2 scores. Since the predictions are not made on the part of the dataset which was used for training the model, a decrease in predictive accuracy is to be expected, which is represented by the difference between the grey (i.e. models without validation) and red bars (i.e. models with cross-validations). Table 11.6 provides the list of the top five predictors for each dependent variable.

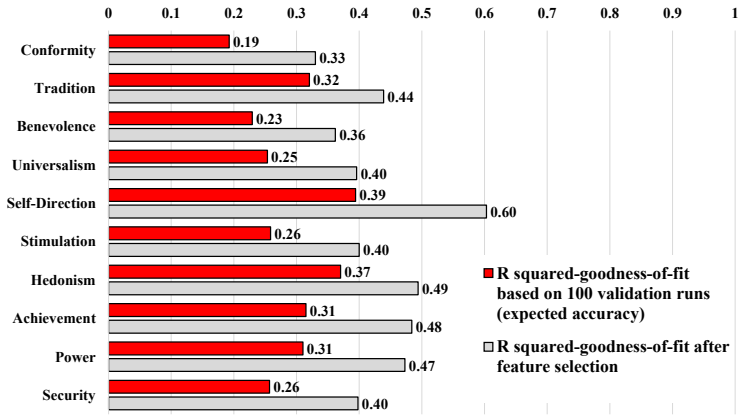


Figure 11.3: Prediction accuracy of Basic Human Values in terms of the R^2 metric. Grey bars represent the maximum accuracy achieved after the models were built with the Stepwise feature selection algorithm in SPSS. Red bars indicate the expected (mean) accuracy of the models after validation using 100 train-test split iterations.

Table 11.6: Top five features for predicting each dependent variable. Standardized Beta coefficients represent each independent variable's effect on the DVs.

Conformity		**Tradition**		**Benevolence**		**Universalism**		**SelfDirection**	
Activity_cigar	-0.18	Activity_presentation	-0.33	Item_carBrand	0.34	Activity_political	0.21	Activity_party	-0.26
Activity_music_alternative	-0.17	Item_numberOfCars	0.27	Activity_music_soundtrack	0.28	Activity_music_jazz	0.19	Item_carType_a	-0.25
Item_headphoneBrand	-0.17	Item_iceSkate	0.24	Activity_charity	0.21	Item _homeLocation	0.17	Item_carType_b	-0.24
Activity_music_electronic	0.15	Activity_highHeels	-0.21	Item_ski	-0.19	Activity_onlinePublishing	-0.17	Item_browser	-0.21
Activity_hunting	0.15	Item_socialMedia	-0.19	Item_searchEngine	0.17	Item_bicycleBrand	0.16	Item_headphoneBrand	-0.21
Stimulation		**Hedonism**		**Achievement**		**Power**		**Security**	
Activity_interview	0.27	Item_bicycleType	0.23	Activity_earring	-0.23	Demographic_citizenship	0.28	Activity_coffee	-0.27
Activity_cigarette	0.24	Item_homeOwnership	0.23	Activity_music_folk	-0.20	Activity_jacket	-0.25	Item_homeOwnership	0.21
Activity_music_alternative	0.18	Activity_fishing	-0.20	Item_bicycleType	-0.19	Item_phoneType	0.22	Activity_interview	-0.19
Activity_onlineForum	-0.18	Activity_learning	-0.19	Item_motorChoice	0.18	Item_homeLocation	-0.20	Activity_music_heavymetal	-0.19
Item_surf	0.15	Activity_snus	0.17	Item_laptopOS	0.17	Item_phoneColor	0.18	Activity_waterpolo	0.17

Model performance evaluation against PVQ-21 test-retest reliability

Following a similar approach which was utilized in [14] for assessing prediction accuracy, Figure 11.4 compares the accuracy of predicting the Basic Human Values scores expressed by the Pearson product-moment correlation coefficients between ground-truth and predicted scores (red bars), whereas grey bars represent the accuracy of the PVQ-21, when the same test is taken by the same individuals (test-retest reliability), and the resulting scores are correlated with each other. The reference PVQ-21 reliability scores are derived from [31], in which a German student sample completed the PVQ-21 two times, separated by an interval of six weeks, to assess the reliability of the questionnaire. The test-retest reliabilities obtained in that study were moderate to high.

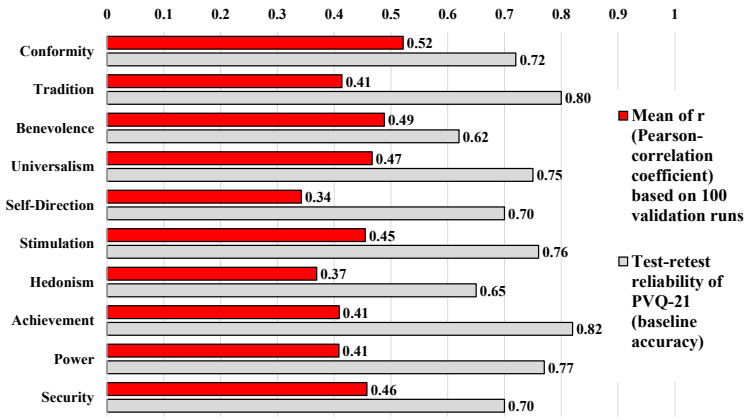


Figure 11.4: Prediction accuracy of Basic Human Values in terms of the Pearson correlation coefficients between predicted and ground-truth scores (red bars). The test-retest reliability is measure of correlation between the results of the PVQ-21 taken at different times by the same respondents (grey bars).

In the present sample Conformity achieved the highest accuracy (r = 0.52), fol-

lowed by Benevolence (r = 0.49), Universalism (r = 0.47), Security (r = 0.46), Stimulation (r = 0.45), Power, Achievement, Tradition (r = 0.41), Hedonism (r = 0.37), Self-direction (r = 0.34) expressed in terms of the Pearson correlation coefficient between ground-truth and predicted scores. The absolute difference is smallest for Benevolence and Conformity; thus, these models can predict the related concepts nearly as well as the PVQ-21 questionnaire, while for the other values, each regression model achieves around half the accuracy of the original questionnaire. Table 11.7 complements Figure 11.4 by providing the mean goodness-of-fit and model-accuracy metrics for each dependent variable. In addition, one-sample Kolmogorov-Smirnov (K-S) tests were run on all metrics to assess whether the distribution of metric scores follows a normal distribution. Cases that do not follow a normal distribution are marked with *.

Table 11.7: Measure of goodness of fit (R^2) and measure of prediction accuracy (r - Pearson-correlation coefficient between ground truth and predicted scores) over 100 train-test split iterations.

	Mean of R^2 measures	SD	Mean of r Pearson-correlation coefficients	SD
Conformity	0.192*	0.103*	0.522	0.086
Tradition	0.320	0.115	0.414	0.083
Benevolence	0.229	0.104	0.488	0.079
Universalism	0.253	0.100	0.467	0.072
Self-Direction	0.124	0.124	0.342	0.077
Stimulation	0.259	0.104	0.455*	0.080*
Hedonism	0.370	0.105	0.369	0.069
Achievement	0.315	0.112	0.409	0.072
Power	0.310*	0.123*	0.408*	0.076*
Security	0.257*	0.095*	0.458	0.072

* denotes cases where normality hypothesis was rejected by the one-sample Kolmogorov-Smirnov test.

The corresponding K-S test scores are as follows: R^2 scores for Conformity D(100) = 0.099, p = 0.016, Power D(100) = 0.124, p = 0.001, Security D(100) = 0.105, p = 0.008; r-scores for Stimulation D(100) = 0.097, p = 0.02 and Power = D(100) = 0.096, p = 0.022.

Example of predicting a single individual's profile scores

Based on the formula for multiple linear regression:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_k X_k + \hat{\epsilon} \tag{11.1}$$

Table 11.8 shows a working example of how an individual's Conformity score is predicted using the trained model. The codes and associated meaning for frequency (i.e. habits) and dummy variables (i.e. ownership of item/existence of attribute) are summarized in Table 11.9.

Table 11.8: Prediction of an individual's Conformity value based on 19 features. PI - Prediction Interval refers to the individual prediction error.

Predictors (valid scores)	**Unstandardized Coefficient** β	**RawScore** ($\mathbf{X}_k$)
Constant	0.218	
consumptionCigar (0-8)	-0.342	0
musicPreferenceAlternative (0-8)	-0.054	7
headPhoneBrand X (0-1)	-0.908	0
musicPreferenceElectronic (0-8)	0.050	1
sportsActivityHunting (0-8)	0.166	0
boatOwned (0-1)	-0.394	1
watchOwned (0-1)	0.291	1
clothWearSuit (0-8)	-0.073	2
laptopBrowser X (0-1)	0.545	1
bicycleBrand X (0-1)	-0.251	0
musicPreferenceSoundtrack (0-8)	-0.044	4
petsSmallMammal (0-1)	-1.428	1
socMedia X (0-1)	-0.222	1
tattooFig X (0-1)	-0.481	0
activityFrequency hairdresser (0-8)	0.073	5
phoneCoverColor X (0-1)	-0.395	1
carEnergy X (0-1)	-0.492	0
carType X (0-1)	0.233	1
watchBrand X (0-1)	0.419	1
IndividualPredictionError-Mean	0.077 (SD: 0.794)	
$\mathbf{Y}_{predicted}$ (95% CI)	**(-1.097) - (-0.785)**	

A prediction interval (PI) captures the uncertainty around the predicted score, which is attributed to uncertainty of coefficients and additional error of individual data points. The errors of individual point estimates are calculated using the residuals from the predicted values using the bootstrapping sampling method (number of re-sampling = number of observations). A bootstrap sample was taken from the data, the model was trained, and a new outcome was predicted. A random residual was taken from the original regression fit and added to the new value. The procedure was repeated for 100 iterations, and the resulting distribution of error terms was

Table 11.9: Explanation of raw variable scores.

Code	Meaning
0	Never in the last 12 months
1	Once in the last 12 months
2	Twice in the last 12 months
3	Three to six times in the last 12 months
4	Seven to 11 times in the last 12 months
5	One to three times a month
6	Once or twice a week
7	Three or four times a week
8	Every day or nearly every day
0	No
1	Yes

used to construct a variable with normal distribution that can be sampled randomly to capture the necessary error terms inherent in individual predictions (ϵ, PI) [1]. The one-sample Kolmogorov-Smirnov test did not reject the null-hypothesis (i.e. PIs are normally distributed D(100) = 0.081, p = 0.101).

Example scenario to assess action desirability

This section provides a simple example to assess the desirability of an action, which demonstrates how the method makes predictions about potential choices based on the derived value structure.

Predicted scores must be normalized by summing across all dimensions, then each score needs to be divided by the sum of scores, to quantify each value's relative importance. Formula: $w_i' = \frac{w_i}{\sum_{j=1}^{n} w_j}$ Thus, the normalized profile scores provide the necessary weights in Table 11.10. For the purpose of demonstration, the relative importance of values is taken from the pan-cultural empirical norms presented in Table 6 in [34]. The Strategy Owner faces a dilemma whether to implement an unconventional strategy that would provide significant personal gains and recognition from the organization's leaders, but which entails a misuse (secondary use) of customer data. An example of such a strategy considered by a stakeholder at an electric distribution system operator would be to use the detailed electricity consumption profiles of homeowners to infer their home occupancy patterns for promoting a novel home-surveillance service through personalized advertisements. Thus, the dilemma can be represented as *Option 0:* Do nothing - contributes positively to Conformity (i.e. restraint of actions that would harm or upset others), whereas Achievement values are unaffected; or *Option 1:* Implement strategy - contributes negatively to Conformity and contributes positively to Achievement (i.e. striving for personal success and recognition) values. For simplicity, the other utility factors

are assumed to be unaffected by the choice. In order to compute the desirability of each option for the Strategy Owner, the Multi-Attribute Utility Theory is used as proposed in [26], where the overall utility of an option is calculated as the weighted sum of the individual utility factors using the formula: $U = \sum_{k=1}^{m} w_k \cdot u(a_k)$, in which m equals the number of utility factors of the stakeholder; w_k is the derived weight of utility factor a_k while $U = \sum_{k=1}^{m} = 1$; and $u(a_k)$ is the utility function for the utility factor a_k. Thus, to compute the utility of an option the normalized weight of each utility factor is multiplied by the score that represents the contribution of that choice on that particular utility factor (i.e. Initial Value, Option 0 - Final Value, Option 1 - Final Value) and these products are summed over all utility factors. For demonstration, the utility calculation for **Option 1** is as follows: $0.11 \cdot 40 + 0.07 \cdot 50 + 0.12 \cdot 50 + 0.11 \cdot 50 + 0.12 \cdot 50 + 0.09 \cdot 50 + 0.1 \cdot 50 + 0.11 \cdot 90 + 0.06 \cdot 50 + 0.11 \cdot 50 = 53.20$. The process is repeated for each identified decision option to enable the comparison between the desirability of various actions. Table 11.11 presents the overall utility calculations for the identified options. The Strategy Owner is assumed to be utility maximizing, therefore selecting the option with the highest overall utility (Option 1). The differences between the utilities associated with the Initial State, Option 0, and Option 1 can be interpreted as the strengths of motivation at work when the Strategy Owner contemplates a particular course of action.

Table 11.10: Expected effects of implementing a strategy on the relevant utility factors. Affected utility factors are marked in **bold**.

Strategy Owner's options:			**Option 0**	**Option 1**
Utility Factors	Normalized Weights	Initial Value	Final Value	Final Value
Conformity (%)	0.11	50	**70**	**40**
Tradition (%)	0.07	50	50	50
Benevolence (%)	0.12	50	50	50
Universalism (%)	0.11	50	50	50
Self-Direction (%)	0.12	50	50	50
Stimulation (%)	0.09	50	50	50
Hedonism (%)	0.10	50	50	50
Achievement (%)	0.11	50	50	**90**
Power (%)	0.06	50	50	50
Security (%)	0.11	50	50	50

Table 11.11: Overall utilities associated with the initial state and with making a choice. The outcome with the greatest expected utility is assumed to be selected by the Strategy Owner (i.e. Option 1 in this example).

Overall utility in Initial State	50.00
Overall utility of Option 0	52.11
Overall utility of Option 1	**53.20**

11.5 Discussion

Modern societies keep on designing and implementing complex systems to fulfill certain goals with increasing efficiency (e.g. legal systems, markets for trading, voting, etc.). Most systems critical for modern life are enabled and dependent on innovations from information and communication technologies. The field has developed a variety of risk assessment methods and tools to deal with unexpected events by assessing the probability of such events and the consequences associated with them. Relatively less attention has been given to the consciously active part of the system - the human decision-maker with its unique motivations. This work aimed at improving the state of knowledge in relation to modeling human decision-makers for the purpose of risk analysis. More specifically, the study aimed at exploring the usefulness of easily observable pieces of information connected to potential decision-makers for inferring individual motivational profiles. This aim is supported by the requirements of the Conflicting Incentives Risk Analysis (CIRA) method, which uses stakeholder motivation to characterize risks. The results present the extent to which these features are valid predictors of the motivational profiles operationalized as the Basic Human Values. Furthermore, the results showed the added utility of this set of features in comparison to previous results using demographic data for the same purpose [36]. The reliability of profile predictions was assessed by various techniques (i.e. cross-validation, comparison with the personality test's test-retest reliability, and calculation of prediction error (prediction interval) for predicting an individual's score). Some aspects of the motivational profile can be predicted nearly as well from the observable features as from the original psychometric instrument (Conformity and Benevolence).

While various steps were taken to include a diverse sample within the data collection, the relatively small sample size can be considered an important limitation, when the generalizability of the findings is considered. In replication studies, it would be desirable to have at least 10-20 unique observations for each category of the independent variables to ensure that inferences made from the sample are valid and robust for the target population. The external validity of the results could be improved using strict probability sampling, since most of the respondents were

recruited through the university's e-mail list, which may result in a biased sample. Furthermore, the length of the survey needs to be reduced to increase respondent retention. Future studies may benefit from converting the obtained categorical data (e.g. type of phone) into corresponding retail prices to enhance the information content of the independent variables. The suitability of the established method for capturing action desirability for the stakeholders (i.e. computing the utilities according to Multi-Attribute Utility Theory) has to be investigated in future work. Choices of human stakeholders can be analyzed in real-world or in experimental settings to assess the procedure's applicability for capturing stakeholder intentions in various choice situations. The procedure's correctness would be verified if the investigation reveals a high degree of overlap between predicted (calculated on the basis of utility calculations) and actual choices made by subjects.

The agenda proposed in [41] calls for a shift in research strategy for psychology, with an increased focus on the prediction of behavior as opposed to explanation. The paradoxical state in which a good explanatory model is not necessarily good at predicting real-world behavior needs to be considered. While the objectives of the traditions may be different, methodological issues are enumerated as the reason for the discrepancy (e.g. p-hacking or lack of model validation on out-of-sample data). The paper proposes that the methodological shift should be aided by relying on machine learning (ML) methods that have been designed and used efficiently in various fields of computer science for the explicit purpose of generating predictive models that perform well on unobserved data as well. It is important to note that the present study utilized a traditional data analysis technique (using cross-validation to ensure reliability) instead of a complex ML method. This represents a conscious choice, where the transparency and interpretability of a simpler model is given a higher priority than the potential predictive improvement enabled by a complex ML method, which operates as a black box. The potential dangers of using black-box models for predictions affecting humans may result in gender or racial bias in case of school admission decisions [29], decisions about risk of re-offending behavior, risk of illness estimations, etc. [3]. Furthermore, European legislation also requires that algorithmic decisions that "significantly affect" subjects are to be explainable [6]. Easy interpretation of the model may increase the risk of manipulation and deception by motivated subjects, which has to be considered for real-world applications [19].

11.6 Conclusion

This paper aimed at investigating the relevance of observable personal items and habits (public observables) for the construction of stakeholder motivational profiles. The stakeholder profiling method presented in this work is expected to complement

the CIRA method, which focuses on stakeholder motivation to characterize risks within the domains of privacy and information security. The real-world applicability of the method depends on the accuracy with which stakeholder motivational profiles can be constructed without direct access to subjects. This paper assessed the predictive accuracy of publicly observable pieces of information associated with individual choices. It was demonstrated that these features are significantly better for profile-building than the most basic features that can be assessed by observation in any context (i.e. demographic features). Several comparisons and evaluations have been presented to assess the validity and reliability of the resulting profiles, and the uncertainty associated with the resulting profile scores has been assessed by the bootstrapping method (i.e. calculation of Prediction Intervals). The error associated with each predicted motivational score is modeled as a random variable with corresponding parameters from a normal distribution. Finally, a demonstration was presented using the utility calculations proposed in CIRA to assess the desirability of the options as perceived by the Strategy Owner in a potential choice situation. The presented work's main contribution is an enhanced understanding of the applicability of stakeholder motivational profiling for the purpose of risk analysis.

Acknowledgment

A. Szekeres would like to thank Dóra Szekeres for the initial Norwegian translation of the survey, Vebjørn Slyngstadli and Eigil Obrestad, for their help with setting up the hosting service and Jag Mohan Singh for useful discussions. We would like to thank the reviewers for their valuable comments which improved the overall quality of the paper.

References

[1] Peter Bruce and Andrew Bruce. *Practical statistics for data scientists: 50 essential concepts*. " O'Reilly Media, Inc.", 2017.

[2] Richard Caralli et al. *Introducing OCTAVE Allegro: Improving the information security risk assessment process*. Tech. rep. CMU/SEI-2007-TR-012. Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, May 2007, p. 154. URL: `https://resources.sei.cmu.edu/asset_files/TechnicalReport/2007_005_001_14885.pdf`.

[3] Danton S Char, Nigam H Shah and David Magnus. 'Implementing machine learning in health care—addressing ethical challenges'. In: *The New England journal of medicine* 378.11 (2018), p. 981.

[4] Folker Den Braber et al. 'Model-based security analysis in seven steps—a guided tour to the CORAS method'. In: *BT Technology Journal* 25.1 (2007), pp. 101–117.

[5] Djellel Difallah, Elena Filatova and Panos Ipeirotis. 'Demographics and dynamics of mechanical Turk workers'. In: *Proceedings of the eleventh acm international conference on web search and data mining*. ACM. 2018, pp. 135–143.

[6] Finale Doshi-Velez and Been Kim. 'Towards a rigorous science of interpretable machine learning'. In: *arXiv preprint arXiv:1702.08608* (2017).

[7] Olav B Fosso et al. 'Moving towards the smart grid: The Norwegian case'. In: *Power Electronics Conference (IPEC-Hiroshima 2014-ECCE-ASIA), 2014 International*. IEEE. 2014, pp. 1861–1867.

[8] Nan Gao, Wei Shao and Flora D Salim. 'Predicting Personality Traits From Physical Activity Intensity'. In: *Computer* 52.7 (2019), pp. 47–56.

[9] Adam Hahn and Manimaran Govindarasu. 'Cyber attack exposure evaluation framework for the smart grid'. In: *IEEE Transactions on Smart Grid* 2.4 (2011), pp. 835–843.

[10] International Organization for Standardization. 'ISO 27005:2011'. In: *Information technology - Security techniques - Information security risk management* (2011).

[11] ISACA. *The Risk IT Framework*. Risk IT. ISACA, 2009. URL: https://books.google.no/books?id=tG7VMihmwtsC.

[12] Joint Task Force Transformation Initiative. *Guide for conducting risk assessments, NIST 800-30*. Tech. rep. National Institute of Standards and Technology, 2012.

[13] Jack Jones. *Factor analysis of information risk*. US Patent App. 10/912,863. Mar. 2005.

[14] Michal Kosinski, David Stillwell and Thore Graepel. 'Private traits and attributes are predictable from digital records of human behavior'. In: *Proceedings of the National Academy of Sciences* 110.15 (2013), pp. 5802–5805.

[15] Barbara Krumay. 'The E-Waste-Privacy Challenge'. In: *Privacy Technologies and Policy*. Ed. by Stefan Schiffner et al. Cham: Springer International Publishing, 2016, pp. 48–68.

[16] Gaoqi Liang et al. 'The 2015 Ukraine blackout: Implications for false data injection attacks'. In: *IEEE Transactions on Power Systems* 32.4 (2016), pp. 3317–3318.

[17] Patrick McDaniel and Stephen McLaughlin. 'Security and privacy challenges in the smart grid'. In: *IEEE Security & Privacy* 7.3 (2009), pp. 75–77.

[18] Ramyar Rashed Mohassel et al. 'A survey on advanced metering infrastructure'. In: *International Journal of Electrical Power & Energy Systems* 63 (2014), pp. 473–484.

[19] Christoph Molnar. *Interpretable Machine Learning*. [Online; accessed 12-November-2019]. 2019. URL: https://christophm.github.io/interpretable-ml-book/interpretability-importance.html.

[20] N.A. *Statistisk Sentralbyrå - Statistics Norway*. [Online; accessed 18-November-2019]. 2019. URL: https://www.ssb.no/en.

[21] N.A. *Norwegian version of PVQ-21*. Available at https://www.europeansocialsurvey.org/docs/round8/fieldwork/norway/ESS8_questionnaires_NO.pdf/. 2016.

[22] Nico JD Nagelkerke et al. 'A note on a general definition of the coefficient of determination'. In: *Biometrika* 78.3 (1991), pp. 691–692.

[23] Robert Nisbet, John Elder and Gary Miner. *Handbook of statistical analysis and data mining applications*. Academic Press, 2009.

[24] Johan Nordstrøm. *Forbrukerombudet: – Strømselskap driver med ulovlig telefonsalg*. [Online; accessed 7-November-2019]. 2017. URL: https://e24.no/privatoekonomi/i/naMbeL/forbrukerombudet-stroemselskap-driver-med-ulovlig-telefonsalg.

[25] Lisa Rajbhandari. 'Risk analysis using 'conflicting incentives' as an alternative notion of risk'. PhD thesis. Gjøvik, Norway: Gjøvik University College, 2013.

[26] Lisa Rajbhandari and Einar Snekkenes. 'Using the Conflicting Incentives Risk Analysis Method'. In: *Security and Privacy Protection in Information Processing Systems*. Ed. by Lech J. Janczewski, Henry B. Wolfe and Sujeet Shenoi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 315–329.

[27] Anne Cecilie Remen and Line Tomter. *Helse Sør-Øst: Innrømmer at utenlandske IT-arbeidere fikk tilgang til sensitive pasientdata*. [Online; accessed 7-November-2019]. 2017. URL: https://www.nrk.no/norge/helse-sor-ost_-innrommer-at-utenlandske-it-arbeidere-har-hatt-tilgang-til-pasientjournaler-1.13478443.

[28] Neal Schmitt. 'Uses and abuses of coefficient alpha'. In: *Psychological assessment* 8.4 (1996), p. 350.

[29] Oscar Schwartz. *Untold History of AI*. [Online; accessed 12-November-2019]. 2019. URL: https://spectrum.ieee.org/tag/AI+history.

[30] Shalom Schwartz. *Computing Scores for the 10 Human values*. [Online; accessed 12-November-2019]. 2016. URL: https://www.europeansocialsurvey.org/docs/methodology/ESS1_human_values_scale.pdf.

[31] Shalom H Schwartz. 'A proposal for measuring value orientations across nations'. In: *Questionnaire Package of the European Social Survey* (2003), pp. 259–290.

[32] Shalom H Schwartz. 'An overview of the Schwartz theory of basic values'. In: *Online readings in Psychology and Culture* 2.1 (2012), pp. 2307–0919.

[33] Shalom H Schwartz. 'Basic human values: Theory, measurement and applications'. In: *Revue française de sociologie* 47.4 (2007), p. 929.

[34] Shalom H Schwartz and Anat Bardi. 'Value hierarchies across cultures: Taking a similarities perspective'. In: *Journal of cross-cultural psychology* 32.3 (2001), pp. 268–290.

[35] Michele Settanni, Danny Azucar and Davide Marengo. 'Predicting individual characteristics from digital traces on social media: A meta-analysis'. In: *Cyberpsychology, Behavior, and Social Networking* 21.4 (2018), pp. 217–228.

[36] Adam Szekeres, Pankaj Shivdayal Wasnik and Einar Arthur Snekkenes. 'Using Demographic Features for the Prediction of Basic Human Values Underlying Stakeholder Motivation'. In: *Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 2: ICEIS*. SciTePress, 2019, pp. 377–389.

[37] Mohsen Tavakol and Reg Dennick. 'Making sense of Cronbach's alpha'. In: *International journal of medical education* 2 (2011), p. 53.

[38] Maria Knoph Vigsnæs, Peter Svaar and Vegards Venli. *Slike bilder sender bomselskap til Kina: Nå går Datatilsynet inn i saken*. [Online; accessed 7-November-2019]. 2019. URL: https://www.nrk.no/norge/slike-bilder-sender-bomselskap-til-kina_-na-gar-datatilsynet-inn-i-saken-1.14754918.

[39] Gaute Wangen, Christoffer Hallstensen and Einar Snekkenes. 'A framework for estimating information security risk assessment method completeness'. In: *International Journal of Information Security* (2017), pp. 1–19.

[40] Kyle Wiggers. *iRobot partners with Google to improve smart home devices with indoor maps*. [Online; accessed 11-November-2019]. 2018. URL: https://venturebeat.com/2018/10/31/irobot-partners-with-google-to-improve-smart-home-devices-with-indoor-maps/.

[41] Tal Yarkoni and Jacob Westfall. 'Choosing prediction over explanation in psychology: Lessons from machine learning'. In: *Perspectives on Psychological Science* 12.6 (2017), pp. 1100–1122.

Chapter 12

Article 4: A Taxonomy of Situations within the Context of Risk Analysis

Adam Szekeres & Einar Arthur Snekkenes. A Taxonomy of Situations within the Context of Risk Analysis. In: *Proceedings of the 25th Conference of Open Innovations Association FRUCT*. FRUCT Oy, Helsinki, Finland. 2019, pp. 306–316.

Abstract

Prediction of deliberate human decisions with potential negative impact on others would have great practical and scientific utility. The Conflicting Incentives Risk Analysis (CIRA) method defines risk as a result of misaligned incentives between various stakeholders. The method makes predictions based on action desirability from the perspective of the individual in the position to implement the action. Therefore, in order to assess action desirability it is necessary to characterize stakeholders and their perceptions about the situation as well. While classification systems and taxonomies related to stakeholder attributes are well-established, systematic classifications of situational aspects are underdeveloped in the literature. Therefore, the main objective of this paper is to present a classification of situational variables in the form of a taxonomy capturing key situational features that exert influence on decision-makers. The development of the taxonomy begins with mapping two major types of risks distinguished in the CIRA method to relevant psychological constructs. The principled, systematic development of dilemmas enabled by the

taxonomy allows researchers to investigate the predictability of stakeholder behavior which may result in various types of risks. The taxonomy is extensible, thus additional concepts and variables can be included depending on the needs of the analysis and according to future developments within the fields of psychology and information security.

12.1 Introduction

Frederiksen's overview on approaches for predicting individual behavior explains that the need for predicting the behavior of single individuals arose and received a great deal of scientific attention during the cold-war era following the realization that an individual could initiate economic and military actions with serious negative consequences for millions of people [11]. In a world characterized by increasing levels of interconnectedness and inter-dependency, where decision-makers and the people affected by critical decisions are linked together and separated by layers of complex technical solutions, there is a pressing need to understand and predict key decision-makers' behavior. The Conflicting Incentives Risk Analysis (CIRA) method was developed for the analysis of risks arising from deliberate human decisions, and re-conceptualizes risk, within the domains of information security and privacy [37]. CIRA requires the identification of two classes of stakeholders, their relevant utility factors and the actions that can be implemented to describe the risk situation. Stakeholder classes are: *Strategy Owner:* the person capable of executing an action and *Risk Owner:* the person(s) enjoying the benefits/suffering the consequences of the actions. To analyze risks resulting from intended human actions which impact the utility factors of the respective stakeholders CIRA asks the question from the perspective of the *Risk Owner*: are we in equilibrium? More specifically, CIRA analyzes situations such as the following: can those that are in the position to implement an action obtain a significant benefit and at the same time cause damage to the *Risk Owner* (in terms of loss of utility)? Such situations are defined as *Threat Risks*. *Opportunity Risks* may result from (in)actions that one can reasonably expect that the *Strategy Owner* should take, but for which the *Strategy Owner* would have to take a loss in utility and the *Risk Owner* has the prospect of a gain [44]. There is a need to enhance the method's applicability by including relevant situational and personality variables which enable predictions with respect to the Strategy Owner's choices in strategic settings. This work contributes to CIRA's ongoing enhancement by proposing a taxonomy of situations -built on existing literature and extending on established results- which enables the systematic manipulation of relevant situational variables for effective dilemma development. The dilemmas created by utilizing the taxonomy facilitate further research attempts to test and fine-tune CIRA's predictive capabilities.

The paper is structured as follows: Section 12.2 provides an overview about existing research work related to the overall objectives and about previous attempts for developing taxonomies of situations. Section 12.3 explains the development of the taxonomy in detail, Section 12.4 presents a set of dilemma examples to demonstrate the usefulness of the taxonomy, Section 12.5 provides the evaluation of the taxonomy. Section 12.6 gives a summary about the relevance and limitations of the proposed taxonomy and identifies venues for further improvements and Section 12.7 concludes the work.

12.2 Related work

A detailed overview is provided on the potential approaches for behavior prediction from a psychological perspective in [11]. It is noted that the scientific perspective is more concerned with generalizations that hold for a large number of people rather than for a single individual, with the exception of clinical applications. The method which relies on individual differences (e.g. aptitude, personality, attitudes, personal history) works well when comparative statements need to be made about the probable performance of many individuals. However, the method fails when the problem is to predict a single individual's behavior across situations over time, since "individual differences" do not exist for the specific person (i.e. lack of comparability). Three potential solutions are presented for the problem of individual behavior prediction:

1. **Personnel psychologist's approach:** requires a measure of a criterion performance y, and at least one measure of personal characteristic x, which is correlated with y. The regression of y on x provides the prediction of criterion performance. Similarly, an analogous procedure would require criterion behaviors measured on many occasions and the predictor variables would have to be personal characteristics that vary over time.

2. **Situational variables approach:** the criterion performance y is predicted by ratings of situational variables which correlate well with the criterion variable over occasions. This method would require extensive assessments of situations across settings.

3. **Clinician's approach:** relies on careful study of the individual and tries to predict (using subjective evaluation) the behavior in previously unobserved situations. This approach is often utilized in clinical settings, for parole decisions, assessment of re-offending behavior, etc. The clinician makes a judgment which implicitly states how the subject with a given set of personal characteristics placed in a specific situation will likely behave. Thus, the

clinician's judgment implies interactions between personal and situational variables.

The first two approaches (i.e. personnel psychologist's, situational variables) correspond to the mechanical approach, while the third one corresponds to the clinical approach in the literature. For detailed discussions about the relative superiority of the mechanical prediction approach over the clinical approach see: [27, 18, 46]. There has been an increased research interest in the interactionist attempts to the behavior-prediction problem (formalized versions of the clinical judgment) but their ineffectiveness might be due to the fact that there is a lack of classification of situations that would enable a systematic way of conceptualizing situations and situational variables. Thus, taxonomies that have been very efficient in classifying variables related to individual differences need to be developed in the domain of situations as well. Taxonomies would allow for a satisfactory and systematic conceptualization of the environment by dimensional analysis of the stimulus variables [41]. "The purpose of a taxonomy is twofold: (1) to structure a domain of objects in order to efficiently handle its information content, and (2) predictive power; if we know that an object belongs to a particular taxon we can immediately predict a number of characteristics that it is expected to possess" [50]. Taxonomies are widely utilized across disciplines for organizing information in a systematic way and for presenting it efficiently and coherently. Taxonomies have been developed to classify: cognitive skills [2], personality attributes [15], information system artifacts [33], network attacks [29], clustering algorithms [9], intrusion detection systems [7], privacy violations [45], etc.

The following overview focuses on attempts for systematically identifying psychologically relevant situational dimensions and for developing taxonomies of situations. The review is restricted to taxonomies of situations constructed within the field of psychology focusing on the individual's perception of the situation. The literature search was conducted with the keywords "taxonomy" AND "situations" in the title in the following databases: Google Scholar, ScienceDirect. Research papers available in English were considered for inclusion. Furthermore, based on the review in [50] additional taxonomies are presented that were otherwise inaccessible in full text. The overview's primary purpose is to demonstrate previous approaches and theoretical considerations, without aiming for completeness. Comprehensive taxonomies are presented first, which aim at capturing influential situational factors across various domains, followed by domain-specific taxonomies which are characterized by a narrower scope, based on the context of application. Table 12.1 presents a classification of the articles included in this overview, based on the breadth of the situations analyzed and approaches for development.

Table 12.1: Classification of existing taxonomies of situations based on the breadth of situations included and the approaches chosen for taxonomy development. The location of the taxonomy developed in this paper is marked with **X** among the existing taxonomies of situations.

		Breadth	
		Comprehensive	*Domain-specific*
Approach for development	*Theoretical*	[24], [22]	[26], **X**
	Empirical	[32], [51], [1], [55], [13], [30]	[8], [25], [12], [28], [35], [10]

Comprehensive taxonomies

An early taxonomy of social situations was developed theoretically [24], guided by ideas from ecological psychology and it identifies seven classes of behavioral settings across various domains: (1) joint working; (2) trading; (3) fighting; (4) sponsored teaching; (5) serving; (6) self-disclosure; (7) playing. According to the theory every person is capable of objectively categorizing a given situation into one of the seven classes and behaves according to the contextual, cultural and role requirements invoked by the given situation.

The Atlas of Interpersonal Situations [22] focuses on the interpersonal aspects (as opposed to impersonal features, e.g. physical) of situations by developing a framework systematically and theoretically. The framework includes 21 frequently occurring situations that can be discriminated and classified according to their conceptual properties, thus the taxonomy does not aim to achieve completeness in terms of all potential situations but aims to focus on factors that are most likely to dominate the individual's attention and behavior according to interdependence theory. Interdependence theory provides a tool for analyzing situations in which individuals influence each other's outcomes. The atlas provides detailed analyses for the 21 situations through interdependence theory's lens.

A taxonomy is constructed from a factor analysis of respondents' descriptions about the relevant situational traits (i.e. persons involved, time and place of the event), feelings and behaviors [32]. The analysis of four participants' responses generated four different taxonomies for each respondent but aggregating them together resulted in the following six situational dimensions: (1) Home and family; (2) Friends and peers; (3) Relaxation, recreation and play; (4) Work; (5) School and (6) Alone.

The lexical approach was utilized by [51] for the development of an empirical tax-

onomy which contains a broad range of objectively defined (i.e. ignores individual differences) situational attributes generated from nouns used for the description of various situations. The cluster analysis revealed the following ten situation dimensions: (1) interpersonal conflict; (2) joint working, exchange of thoughts, ideals and knowledge; (3) intimacy and interpersonal relations; (4) recreation; (5) travelling; (6) rituals; (7) sport; (8) excesses; (9) serving; (10) trading.

The joint taxonomy of traits and situations [1] aims to consider how traits get expressed in various situations, and how situations differ in the type and number of traits that are expressible in them. Based on the Big Five trait taxonomy, situations were generated by participants considering the expression of the given trait in various situations. A reduced set of situations was evaluated by the probability of a trait-related behaviors' occurrence. The principal component analysis revealed five situation dimensions named as: (1) adversity; (2) amusement; (3) positioning; (4) conduct; (5) daily routine.

Another taxonomy using the lexical approach on Chinese idioms is presented in [55]. Based on participant's judgment of the idioms content it was revealed that goal processes (i.e. what impact a given situation had on the goals of the people described in the idioms) was a major distinguishing factor between situations. On the broadest level, people distinguish situations along the success-failure dimensions (the situation's impact on the goals), while at more fine-grained level 17 factor solutions were deemed best, based on various statistical considerations.

A cross-cultural (U.S. and Japan) study using the Riverside Situational Q-sort method shows preliminary evidence that both cultures assess the importance of two dimensions similarly when evaluating situations [13]. The relevant dimensions identified in the study are: (1) presence of a member of the opposite sex; (2) and the experience of being criticized by others.

The CAPTION-model presented in [30] is one of the latest attempts for constructing a comprehensive situation taxonomy through factor analysis of in-situ qualitative descriptions provided by respondents (i.e. using the lexical approach). The basis for the work was the lexical corpus of U.S. movie subtitles with 51 million words, which was screened by Amazon Turk workers. The study constructs a framework by identifying key similarities and differences among a wide collection of situation characteristics. CAPTION refers to the 7 situation dimensions which emerged after applying several data-analytic techniques: (1) Complexity; (2) Adversity; (3) Positive Valence; (4) Typicality; (5) Importance; (6) Humor; (7) Negative Valence. Additionally, the study presents the assessment of the psychometric properties (e.g. internal factor structure, convergent-discriminant validity, predictive validity) of the measure developed from the taxonomy.

Domain-specific taxonomies

The theoretically constructed, domain specific taxonomy of high-risk situations for relapse in relation to alcohol abstinence was proposed in [26]. The taxonomy is built from accumulated research results to enable the identification and classification of situations increasing the probability of a relapse. It identifies five sub-categories under the "Intrapersonal determinants" (e.g. Urges and temptations), and three sub-categories within the "Interpersonal determinants" dimension (e.g. Social pressure). The taxonomy allows practitioners to develop cognitive-behavioral interventions matching specific categories, to which patients may be exposed. Furthermore, the taxonomy enables the targeted training of specific coping strategies needed to deal with specific high-risk situations.

The taxonomy presented in [8] organizes a total of 11 situations according to their potential for evoking anxiety in subjects. The taxonomy is based on factor analysis of responses and distinguishes three classes of situations based on their anxiety-provoking potential: (1) interpersonal situations; (2) dangerous situations without social aspects; (3) ambiguous situations. The selection of situations was guided by intuitive attempts to present respondents with a variety of situations that most people have experience with.

Another empirically developed situational taxonomy is presented in [25], based on similarity judgments of situations in an academic setting. The hypothesis upon which the study builds supposes that people distinguish between situations along unique cognitive dimensions, which raises the problem that the structure is flexible and changes across domains between individuals as well as within individuals. The factor analysis of participants' responses identified 5 dimensions: (1) positive situations; (2) negative situations; (3) passive situations; (4) social situations; and (5) active situations.

The empirical taxonomy in [12] is based on the idea that the similarity of situations should be assessed on the basis of the elicited behaviors. The taxonomy is created by using a three-dimensional data matrix which consists of individuals, situations, and elicited behaviors. The matrix is collapsed across people, thus ignoring individual differences and is factor analyzed to reveal clusters of situations invoking similar behaviors. The domain of the taxonomy is based on hypothetical work tasks that respondents had to solve assuming a chief executive role. The following six factors emerged in this specific taxonomy of executive tasks: (1) evaluation of procedures for accomplishing organizational goals; (2) routine solution; (3) solution of inter-organizational problems; (4) solution of personnel problems; (5) change in policy; (6) conflicting demands on staff time.

Individual's perception about the psycho-social features (i.e. perceived climates) of various social environments form the basis of the taxonomy in [28]. The taxonomy implicitly takes into account the personality of the respondents and it identifies three dimensions: (1) relationship; (2) personal development; (3) system maintenance and system change.

The taxonomy presented in [35] is based on the appropriateness of behaviors in various situations. The situations and behaviors were generated from university students' diaries, and respondents had to judge the resulting combinations in terms of the appropriateness of the behavior in various situations. The matrix was cluster-analyzed and resulted in the following four homogeneous situation-clusters based on their specific behavioral content: (1) park, sidewalk, football game; (2) dating, family dinner, movies; (3) bar, elevator, job interview, restroom; (4) class, church, bus, dorm lounge, own room.

A taxonomy of social episodes is presented in [10], which is based on the individual's perception of recurring interaction sequences, which are defined by symbolic, temporal and physical boundaries. A student and a housewife sample generated lists of adjectives describing their interactions over the course of a day which were used to form the hypothetical dimensions. The relatedness of the episodes was Q-sorted by participants and resulted in a two-dimensional configuration for housewives and a three-dimensional configuration for students. The episode structure according to the perception of the housewives is governed by (1) perceived intimacy, involvement, and friendliness of episode; (2) subjective self-confidence, or competence of the actors related to the episodes. For the student sample the following structure emerged: (1) involvement; (2) pleasantness; (3) knowledge about how to behave.

Based on the overview of existing attempts at developing taxonomies of situations a few things may be noted: the environment and situations are rich (i.e. abundant with features), which results in a high degree of incompatibility across taxonomies. This may reflect the complexity associated with situational aspects; the difficulty associated with objectivist descriptions of situational attributes which exert influence on the behavior irrespective of personality characteristics; and that the goals of the taxonomy (i.e. application domain), as well as the personal history of the researchers largely influences which situational aspects, methods and analytical procedures are evaluated as appropriate for solving a given research problem. Despite efforts aiming for comprehensive situational taxonomies the field is still characterized by perplexity. Existing taxonomies vary significantly in their perspectives on the relevant situational features. The overview suggests that a feasible approach for developing a practically useful situational taxonomy starts by investigating the domain of application extensively. Next, it should consider existing research results

that capture specific situational features assumed to be relevant within the field; and finally synthesizes the results in a concise manner.

12.3 Development of the Proposed Taxonomy

Key requirements for the taxonomy are as follows:

1. to systematically categorize situations based on a subset of their attributes, which have been demonstrated to exert influence on decision-makers.
2. to enable the development of dilemmas which can be used for testing and improving the predictive capabilities of CIRA.
3. to operationalize risk concepts in CIRA (i.e. Threat/Opportunity Risk) and connect them to existing research traditions.

This section describes the method of the taxonomy development, starting by mapping CIRA's risk concepts and major psychological constructs. A definition for each dimension's meaning with reference to previous research results is provided and the section ends with the presentation of the proposed taxonomy.

A taxonomy classifies objects of interest (such as animals and plants, etc.) into groups within a larger system according to their similarities and differences [5]. However, classification systems are always somewhat arbitrary [11]. A taxonomy which successfully classifies objects may have useful implications improving theories and facilitating discoveries (i.e. the periodic table of elements, Carl Linnaeus's taxonomies). The most widely used techniques for empirically developing taxonomies in the field of individual differences, (e.g. abilities, intellect, personality) are factor analysis or clustering analytic methods that rely on a vector of attribute scores for individuals. For cluster analysis the measure of similarity for a pair of individuals is not the correlation coefficient (as opposed to inverse factor analysis), but the number of shared features, an aggregated similarity judgment of objects, the Euclidean distance between two vectors or any other sophisticated, generalized distance metric [11]. Taxonomies, however can be constructed by theoretical considerations as well. Such taxonomies can be built by taking all the possible combinations of identified attributes, while keeping in mind that this method may result in a large number of categories, or categories that do not exist in real life [11].

The present development followed the theoretical approach by identifying and combining relevant situational attributes based on existing and related research results. During the development, the following factors were considered:

- the domain of application for the proposed taxonomy (i.e. human-related

risk analysis in the field of information security and privacy as defined by the CIRA method's purpose),

- theoretical considerations and analysis of the underlying mechanism of decision-making relevant for the risk concepts identified in CIRA,
- compatibility of existing and well-established psychological constructs with CIRA concepts,
- existing empirical results about relevant situational dimensions for decision-making associated with the previously identified concepts and constructs.

Taken together, these considerations lead to a mapping between two central concepts defined within CIRA and established psychological constructs: **Threat risk** was mapped to the concept of a **Moral dilemma**, and **Opportunity risk** was mapped to the construct of **Altruism**. The mapping enables the operationalization of the two distinct risk types as established psychological constructs for research purposes, however other operationalizations are also conceivable. The mapping enabled the identification of existing research results in these separate domains of scientific inquiry which were combined for the construction of the taxonomy. The initial conceptual model of the proposed taxonomy of situations and mapping of psychological constructs to key CIRA concepts is presented in Fig. 12.1. Further steps of the development of the taxonomy are presented below by presenting additional dimensions, with their relevance supported by empirical results and theoretical considerations.

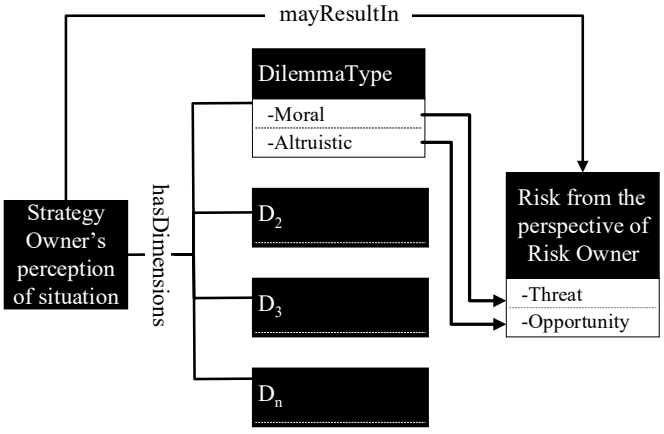


Figure 12.1: Initial conceptual model of the situation taxonomy with mapping of psychological constructs to risk types distinguished by CIRA.

DilemmaType

The dilemma type dimension was developed as follows: the moral dilemma questionnaire (used in cognitive neuroscience for the investigation of dual process theories in moral judgments) presented in [16] served as the starting point for the development of the taxonomy. The original objective of the questionnaire was to enable investigations into the neural correlates of moral decision-making (i.e. Is it acceptable to inflict harm upon a victim for the benefit of others?). Moral philosophers have identified discrepancies between responses given to dilemmas that are identical in terms of their objective outcomes, but differ in the level of engagement required by the decision-maker (the prototypical dilemmas are known as the *Trolley dilemma* and the *Footbridge dilemma* see: [16]). These dilemmas are characterized by their difficulty which is attributed to the conflict between dissociable psychological processes. These processes yield different solutions to the problem based on a utilitarian (i.e. consequentialist) and a non-utilitarian (emotion-driven, deontological, rule-based) assessment [17]. These dilemmas are especially hard since, no matter which solution is selected, the other system will be dissatisfied [6]. The original dilemmas [16] were reused in several studies and got refined over time to allow more detailed investigations. Altruistic dilemmas proposed in this taxonomy represent counterparts of moral dilemmas. Based on the structural features of moral dilemmas, altruistic dilemmas were introduced, in which the respondent has to decide whether to implement a self-sacrificing act for the benefit of others. The crucial difference is that altruistic choices require that the decision-maker take a loss in a broad sense (e.g. in terms of money, time, health, etc.) in order to provide a benefit for others (in a broad sense as well). Thus, for altruistic situations conflict arises between immediate self-interest and between the potential benefit provided for others. Inspiration was taken from the Altruistic Personality Scale presented in [39], but the dilemmas developed using the proposed taxonomy do not aim to assess altruism as a personality trait. A third dilemma type was also considered (previously termed as non-moral dilemmas in [16]) which require the weighing of costs and benefits for the decision-maker only, thus dilemmas in this category have no influence on others, except the decision-maker. The proposed taxonomy identifies the following three *DilemmaTypes*: **Moral**, **Altruistic** and **Rational**.

Context

Context refers to salient features of the environment that may impact behavior in predictable ways by activating short-term goals in a given role. The inclusion of this dimension is supported by evidence that there is significant within-person variability of expressed and experienced personality states across situations throughout extended periods of time [19] and across roles [42, 36]. Some proposed models aim at capturing and integrating how social roles are associated with different

types of short-term goals which represent important aspects of situations which in turn exert influence on expressed personality traits [20, 38, 54]. Management of role requirements in various work settings is a central topic in economics and is known as the principal-agent problem. Research in the field focuses on ways to achieve alignment between the interests of workers and employers using proper incentives [34]. In its current form the proposed taxonomy distinguishes between two *Contexts*: **Private** and **Professional**.

PhysicalDistance

The physical distance dimension matches with the classification used in the refined Greene-dilemmas [17]. It has been shown that impersonal dilemmas (i.e. there is no physical contact with the victim) increases the tendency to use the utilitarian decision-making approach compared to personal dilemmas (i.e. harm is directly inflicted upon somebody). The distinction applies to dilemmas in the *Altruistic* category such that altruistic personal dilemmas imply that a benefit is provided to someone else through direct physical interaction, while impersonal altruistic dilemmas introduce physical separation between the decision-maker and the potential beneficiary. The *Rational* dilemma type has no corresponding *PhysicalDistance* dimension, since it captures decisions that require pure cost-benefit analysis, which have no direct or indirect impact on others than the decision-maker. The taxonomy identifies the following levels of the *PhysicalDistance* dimension: **Personal** and **Impersonal**.

LevelOfConflict

The updated set of moral dilemmas in [17] distinguishes between high- and low-conflict dilemmas only in the case of personal dilemmas. High-conflict dilemmas mean that the two parallel evaluative processes provide contradictory answers, while in general for low-conflict dilemmas the suggestion from the rule-based (deontological) system overrides the utilitarian system's suggestion or they are in alignment. The original categorization is now extended to the **Impersonal** level such that **Impersonal** High-conflict dilemmas would entail an indirect loss inflicted upon others for a greater good, while impersonal low-conflict dilemmas would require an indirect loss inflicted upon others for a selfish gain in case of moral dilemmas. For altruistic choices the *LevelOfConflict* signifies a high or low cost for the self, given that the action is initiated. *Rational* dilemmas have no corresponding *LevelOfConflict* dimension. The taxonomy identifies the following levels within the *LevelOfConflict* dimension: **High** and **Low**.

Fig. 12.2 shows the overall structure of the proposed taxonomy, which enables the systematic manipulation of situational variables thus allowing the construction

of specific dilemmas for each leaf node. This results in a taxonomy with 18 leaf nodes in total. The main objective of the taxonomy is that it provides a structured, principled way to develop dilemmas which can be used to test and fine-tune CIRA's predictive capabilities.

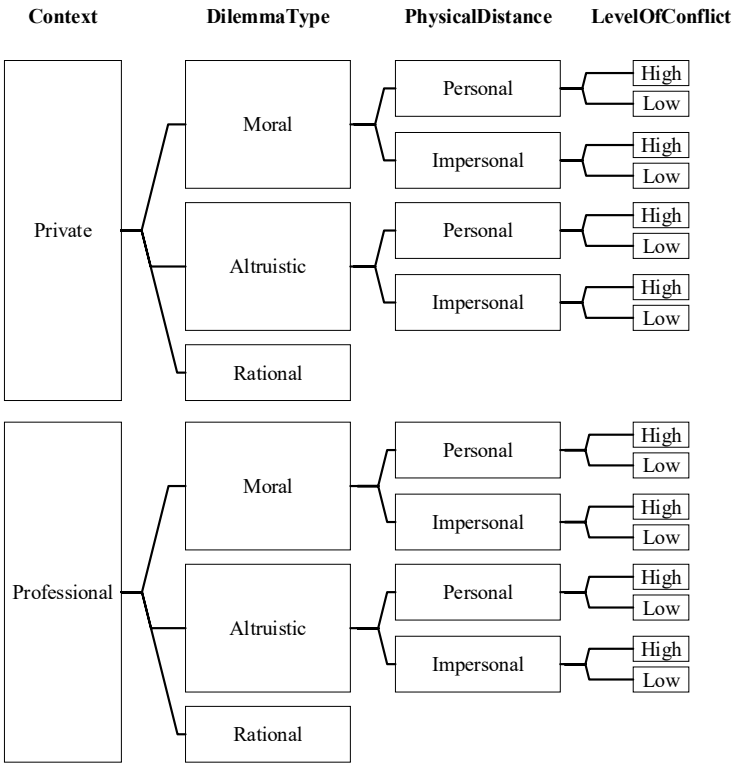


Figure 12.2: Structure of the proposed taxonomy of situations.

12.4 Illustrative scenarios

For each leaf node two different dilemmas were constructed to enable the prediction of subject's responses (36 dilemmas in total). For the purpose of demonstration and due to space limitations only three dilemmas are presented here. One of them -which was previously developed in [16]- is used for demonstrating how existing dilemmas can be categorized according to the taxonomy; and two new dilemmas demonstrate how the taxonomy enables the creation of novel dilemmas in a systematic manner.

Table 12.2 provides an overview about the dilemmas by specifying the leaf-node (category), the identified Risk Owner(s), and the corresponding type of risk. In an experimental setting a respondent takes the role of the Strategy Owner. The

dilemmas are identified by their leaf-nodes, using the following abbreviations: Priv. (**Private**), Prof. (**Professional**) for *Context*; M (**Moral**), A (**Altruistic**), R (**Rational**) for *DilemmaType*; P (**Personal**), I (**Impersonal**) for *PhysicalDistance*; H (**High**), L (**Low**) for *LevelOfConflict*.

Table 12.2: Dilemma examples constructed by using the taxonomy. Risk Owner(s) and the type of risk experienced is specified in advance. The dilemma marked with * is taken from [16].

Leaf-node code	Risk Owner(s)	Type of risk (Threat/Opportunity)	Title of dilemma
Priv-M-P-H	Child vs. group	T	Crying baby*
Prof-M-I-L	Shareholders, employees	T	CEO
Prof-A-I-H	Society that values privacy	O	Whistleblower

Classification of an existing dilemma

Categorization of existing dilemmas and real-world situations can be achieved by analyzing the case according to the taxonomy's structure. By identifying the dimensions and the levels associated with the dimensions it is possible to assign cases into a unique category specified by the taxonomy. This is illustrated by the Crying baby dilemma from [16], which can be placed into the category defined by the **Private** *Context*, **Moral** *DilemmaType*, **Personal** *PhysicalDistance* and **High** *LevelOfConflict* dimensions. The dilemma goes as follows:

Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables. Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death. Is it appropriate for you to smother your child in order to save yourself and the other townspeople?

Dilemma generation

The first step in generating novel dilemmas is to specify which types of risk (Threat or Opportunity) needs to be operationalized as a dilemma. This determines the level of the *DilemmaType* dimension. Next, dilemmas can be generated by relying on the taxonomy's structure and asking questions such as: What are the potential situations that a decision-maker may encounter in a work setting (*Context* - **Professional**) which require no direct physical interaction (*PhysicalDistance* - **Impersonal**) with the Risk Owner and would require a choice between decreasing the utility of the

Risk Owner while providing benefit for the Strategy Owner (*LevelOfConflict* - **Low**)? A potential answer is illustrated by the CEO dilemma:

You are the CEO of a giant IT company which specializes in the development and production of microprocessors. Based on internal communications you become aware of a major vulnerability in one of your products. If this vulnerability becomes public knowledge, it will have a very negative impact on the share prices, and there is no way to prevent this from becoming public knowledge soon. As you own a large number of the company's stocks and options you would lose a lot of money. Based on your knowledge about the situation you consider selling $24 million worth of your shares and options. Would you sell your stocks of the company you are a leader of before the vulnerability hits the news?

A dilemma which operationalizes an Opportunity Risk can be developed by asking the following question based on the taxonomy's structure, given that *DilemmaType* is set to **Altruistic**: What are the potential situations that a decision-maker may encounter in a work setting (*Context* - **Professional**) which require no direct physical interaction (*PhysicalDistance* - **Impersonal**) with the Risk Owner and would require a choice between increasing the utility of the Risk Owner while causing a significant loss of utility for the Strategy Owner (*LevelOfConflict* - **High**)? A potential answer is illustrated by the Whistleblower dilemma:

You work as an information technology service contractor for various governmental organizations. Your work is strictly confidential and classified, you are legally obliged not to talk about the details of your job to anyone neither privately nor publicly. During your work you realize that the material you are working on reveals the extent and sophistication with which your government monitors digital communications between its citizens. If you reveal these secret documents to the public you will instantly receive huge media attention, charges will be pressed against you for theft of government property and you may have to flee your country to avoid going to prison. Would you reveal the secret governmental documents to the public if you were sure that serious consequences for yourself would be unavoidable?

The questions accompanying the taxonomy narrow down the search-space sufficiently and guide researchers so that dilemmas can be generated by systematically manipulating each level associated with the dimensions. Control over the variables, and systematic manipulation would not be feasible without an explicitly and properly defined structure.

12.5 Evaluation of the proposed taxonomy

A key component of the design science research methodology is the evaluation of the resulting artifact. Despite this key requirement and despite the popularity of taxonomies, there is a lack of consensus on how to evaluate taxonomies according to the literature review provided in [49]. The paper therefore constructs a framework for taxonomy evaluation, which is used in this section for evaluating the proposed taxonomy of situations. Furthermore, several points are considered from the article available at [43].

Based on the framework proposed by [49] it is possible to analyze the evaluation procedure by answering the following three questions: Who was involved in the evaluation (i.e. subject)? What type of objects were used for the evaluation (i.e. object)? How was the evaluation performed (i.e. method)?

Evaluation was performed by two persons, each with different academic backgrounds (i.e. psychology and computer science), one being involved in the development of the taxonomy. Objects used for building the taxonomy (i.e. existing dilemmas) and objects not used (i.e. dilemmas generated from the taxonomy) for taxonomy construction were utilized during the evaluation, therefore the coverage of objects can be characterized as selective, but not exhaustive. The evaluation relies on the logical argument and illustrative scenario methods.

The taxonomy's face validity (i.e. compatibility with existing theories and ability to capture relevant concepts in a field [43]) was assessed subjectively as satisfactory, as it creates a link between well-established research results and various types of risk identified in the CIRA method. However, it should be noted that the particular operationalization of risks proposed in this paper is not the only one conceivable. Logical argument revealed that the taxonomy may suffer from a lack of completeness, so that certain dilemmas may arise that do not fit well in the existing taxonomy. The issue of reciprocal altruism (i.e. the decision-maker takes a short-term loss, with the expectation that at a later point it will be reciprocated by the other party) arose during logical arguments. Currently the taxonomy does not have a corresponding class. This could be alleviated by the inclusion of an additional **Reciprocal** sub-class for the **Altruistic** *DilemmaType* dimension. While it may be possible that an act of reciprocal altruism is motivated by rational cost-benefit analysis; **Rational** dilemmas by definition have no impact on other stakeholders. Furthermore, by definition **Moral** dilemmas refer to potential losses exerted on others, while **Altruistic** dilemmas refer to choices that increase the benefit of others, therefore the categories fulfill the requirement of mutual exclusivity [43]. The illustrative scenario method was used to demonstrate how existing dilemmas and situations can be classified according to the taxonomy. Furthermore, the method

presented how previously non-existent dilemmas can be generated in a systematic and principled way by manipulating the levels associated with each dimension of the taxonomy. Therefore, the usability property has been demonstrated, while more rigorous assessments may be advantageous in the future.

Overall, the taxonomy fulfills the key requirements by enabling the creation of novel dilemmas in a systematic and principled way; by providing a way to operationalize both types of risks identified in the CIRA method; by enabling the classification of existing dilemmas and real-world situations into unique categories.

12.6 Discussion

In order to predict stakeholder behavior, it is crucial to obtain information about the individual and the context in which a decision-maker operates from a person-situation interactionist perspective. During the development of the CIRA method, which focuses on human motivation for the purpose of risk analysis, characterization of the decision-maker's motivation received more attention than relevant aspects of the situation. This resembles the state of the scientific literature which is abundant with systematic and well-tested personality and trait theories, whereas the description of situational aspects is less advanced and far from being unified [50].

This work contributes to the development of the CIRA method in the following ways: by developing a taxonomy of situations based on a review of previous approaches; by identifying relevant psychological constructs and establishing a mapping between these and CIRA's key risk-concepts. The selection of situational dimensions included in the taxonomy is supported by empirical evidence and theoretical considerations. The key utility of the proposed taxonomy is that it allows the systematic manipulation and control of situational factors, thus enables the principled development of hypothetical scenarios which will be used in future investigations to test and improve the existing framework's predictive capabilities.

Several widely-publicized, high-impact decisions (e.g. diesel emission-scandal [21], bribery [52], Watergate-scandal [53], insider trading [14], etc.) with negative outcomes for various classes of Risk Owners fall in the **"Professional-Moral-Impersonal-Low-conflict"** category according to the proposed taxonomy. While this category was not explicitly defined in previous studies, the potential effect of anxiety (as experienced by the decision-maker when contemplating the consequences of the actions) on choices was investigated in various studies. Researchers have found that both high-anxiety and low-anxiety psychopaths were more likely than participants in a control group to endorse harmful impersonal acts which cause indirect or remote harm to others [23]. Additionally, low-anxiety psychopaths were more likely than control subjects or high-anxiety psychopaths to enact

harmful behavior in personal dilemmas. Another study found that the anti-anxiety drug lorazepam caused a dose-dependent increase in participants' willingness to engage in harmful actions in the personal condition (for high-conflict and low-conflict situations as well), but it did not significantly change responses in case of impersonal dilemmas [31]. These results suggest that-since impersonal situations are less anxiety-provoking (compared with personal situations), -detachment from consequences in itself has important implications for moral decision-making for a variety of settings where Strategy Owners and Risk Owners are interconnected (and at the same time separated) by sophisticated technical means. The importance of understanding how human moral judgement is influenced by situational factors becomes increasingly important as more and more autonomous systems will have to rely on some sort of simulated human judgement when making their choices on behalf of others. Due to the fact that several problems in real-life have no objectively defined criteria which could be used to evaluate whether a decision is right or wrong, systems may have to use human judgements as the gold standard [3].

Limitations and further work

While the benefits of the proposed taxonomy (i.e. enabling systematic development of dilemmas, and classification of situations) have been demonstrated through the examples, there are limitations which have to be considered. Empirical tests are needed to assess whether the taxonomy allows useful deductions regarding the predictability of subjects' choices. If the taxonomy successfully captures the decision-makers' mental model, systematic differences may emerge from the responses, thus valid predictions on probable subject behavior could be made simply by matching the leaf-nodes in the taxonomy with real-world situations. This empirical test represents planned future work. The taxonomy is constructed theoretically and by considering previous research results, thus enumerates dimensions and specifies the associated levels on each dimension. This method may result in leaf-nodes that are rare or non-existent in realistic settings, and quickly leads to a combinatorial explosion as the number of dimensions increases, making it unmanageable for human experts. However, the inclusion of additional dimensions may be necessary to capture other forms of dilemmas that may arise in realistic settings. Inclusion of the *AmountOfBenefitProvided* dimension in case of **Altruistic** dilemmas would enable manipulation of the amount of benefit provided by an action, which could also be an important factor from the decision-maker's perspective. Furthermore, incorporating a *Reciprocal* sub-dimension for **Altruistic dilemmas** would potentially improve the taxonomy's capability to classify actions with hidden motives. The overall appropriateness of included dimensions should be judged by considering the purpose of the application and the existing domain-specific research results.

The proposed taxonomy has been evaluated using various methods, however more rigorous evaluations may be carried out in the future by applying the taxonomy in a real-world context using the action research method (i.e. asking practitioners/researchers to generate dilemmas using the taxonomy) or using the case study method over an extended period of time, for real-world applications to evaluate its performance more independently [49].

During dilemma development care must be taken to control for several undesirable effects that may threaten the validity of the measure (e.g framing effects; descriptions suggestive of the trade-offs assumed implicitly by the researcher; and to avoid lengthy dilemmas resulting in respondent fatigue [4]). Furthermore, it is especially challenging to control for spill-over effects across contexts (i.e. a choice in a professional setting may have important implications for the private context as well). Finally, it should be mentioned that it is possible to construct the same scenario both as an **Altruistic** dilemma (i.e. providing benefit for others at own expense) and as a **Moral** dilemma (i.e. decreasing the utility of others) by manipulating the *Risk Owner* variable (e.g. Whistleblower-dilemma can be turned into a special kind of **Moral** dilemma -in which both stakeholders would have to take a loss- when the previously identified *Risk Owner* is replaced by the employer who will be negatively impacted). For real-world applications which aim at simulating the Strategy Owner's mental model of the situation, it would be crucial to understand which framing is more active from the set of potential mental representations. The decision-maker's value hierarchy obtained through unobtrusive measures [48, 47] may enable inferences about which cognitive representation is more active (i.e. how does the *Strategy Owner* actually perceive the situation?). The topic requires extensive future work and needs to be guided by relevant results obtained from investigations into how values get activated, how they motivate behavior, and how they relate to pro-social and moral decision-making, since the hierarchy of values fundamentally influences how a situation is perceived by individuals [40].

Taken together, these observations lead to the conclusion that challenging dilemmas are not just hard to solve but are hard to develop as well. Furthermore, real-world applications need to combine several research results in order to predict individual choices in specific situations, where complex interactions between personal, intrapersonal and situational factors produce observable outcomes.

12.7 Conclusions

This paper aimed at proposing and developing a taxonomy of situations, based on existing literature and theoretical considerations by identifying limitations in existing solutions and by extending on well-established research results. The need for the development of a domain-specific taxonomy of situations arises from the

fact that predicting the choices of key decision-makers is a central aim of the CIRA method. While personality and trait theories are suitable for characterizing individuals, they cannot account for the intra-individual personality-state variability expressed in various situations and across different social roles. The taxonomy of situations proposed in this work incorporates key situational attributes which have significant influence on decision-makers, as demonstrated by existing research results. The taxonomy proposes a novel way to operationalize risks identified in the CIRA method, thus providing a connection between separate areas of scientific inquiry. Additional benefits of the proposed taxonomy include: enabling the creation of novel dilemmas in a systematic and principled way; categorization of existing dilemmas and real-world situations based on their attributes. The dilemmas generated by utilizing the taxonomy's structure enable further empirical assessments and improvements related to CIRA's predictive capabilities. This work contributes to the inter-disciplinary effort which aims at developing novel tools for improved decision-making by focusing on human-related risks in the context of information security risk analysis.

Acknowledgements

This work was partially supported by the project IoTSec – Security in IoT for Smart Grids, with number 248113/O70 part of the IKTPLUSS program funded by the Norwegian Research Council.

References

[1] Maaike Ten Berge and Boele De Raad. 'The construction of a joint taxonomy of traits and situations'. In: *European Journal of Personality* 15.4 (2001), pp. 253–276.

[2] Benjamin S Bloom et al. 'Taxonomy of educational objectives. Vol. 1: Cognitive domain'. In: *New York: McKay* (1956), pp. 20–24.

[3] Jean-François Bonnefon, Azim Shariff and Iyad Rahwan. 'The social dilemma of autonomous vehicles'. In: *Science* 352.6293 (2016), pp. 1573–1576.

[4] Julia F Christensen et al. 'Moral judgment reloaded: a moral dilemma validation study'. In: *Frontiers in psychology* 5 (2014), p. 607.

[5] CollinsDictionary. *Taxonomy definition and meaning*. URL: `https://www.collinsdictionary.com/dictionary/english/taxonomy`.

[6] Fiery Cushman and Joshua D Greene. 'Finding faults: How moral dilemmas illuminate cognitive structure'. In: *Social neuroscience* 7.3 (2012), pp. 269–279.

[7] Hervé Debar, Marc Dacier and Andreas Wespi. 'A revised taxonomy for intrusion-detection systems'. In: *Annales des télécommunications*. Vol. 55. Springer. 2000, pp. 361–378.

[8] Norman S Endler, JMCV Hunt and Alvin J Rosenstein. 'An SR inventory of anxiousness'. In: *Psychological Monographs: General and Applied* 76.17 (1962), p. 1.

[9] Adil Fahad et al. 'A survey of clustering algorithms for big data: Taxonomy and empirical analysis'. In: *IEEE transactions on emerging topics in computing* 2.3 (2014), pp. 267–279.

[10] Joseph P Forgas. 'The perception of social episodes: Categorical and dimensional representations in two different social milieus'. In: *The Psychology of Social Situations*. Elsevier, 1981, pp. 80–94.

[11] Norman Frederiksen. 'Toward a taxonomy of situations'. In: *American Psychologist* 27.2 (1972), p. 114.

[12] Norman Frederiksen et al. *Prediction of organizational behavior*. Pergamon, 1972.

[13] David Funder et al. 'The person-situation debate and the assessment of situations'. In: *The Japanese Journal of Personality* 21.1 (2012), pp. 1–11.

[14] Sean Gallgaher. *Intel CEO sold all the stock he could after Intel learned of security bug*. URL: `https://arstechnica.com/information-technology/2018/01/intel-ceos-sale-of-stock-just-before-security-bug-reveal-raises-questions/`.

[15] Lewis R Goldberg. 'An alternative description of personality: the big-five factor structure'. In: *Journal of personality and social psychology* 59.6 (1990), p. 1216.

[16] Joshua D Greene et al. 'An fMRI investigation of emotional engagement in moral judgment'. In: *Science* 293.5537 (2001), pp. 2105–2108.

[17] Joshua D Greene et al. 'Cognitive load selectively interferes with utilitarian moral judgment'. In: *Cognition* 107.3 (2008), pp. 1144–1154.

[18] William M Grove et al. 'Clinical versus mechanical prediction: a meta-analysis'. In: *Psychological assessment* 12.1 (2000), p. 19.

[19] Daniel Heller, Jennifer Komar and Wonkyong Beth Lee. 'The dynamics of personality states, goals, and well-being'. In: *Personality and Social Psychology Bulletin* 33.6 (2007), pp. 898–910.

[20] Daniel Heller, Wei Qi Elaine Perunovic and Daniel Reichman. 'The future of person-situation integration in the interface between traits and goals: A bottom-up framework'. In: *Journal of Research in Personality* 43.2 (2009), pp. 171–178.

[21] Russell Hotten. *Volkswagen: The scandal explained*. URL: `https://www.bbc.com/news/business-34324772`.

[22] Harold H Kelley et al. *An atlas of interpersonal situations*. Cambridge University Press, 2003.

[23] Michael Koenigs et al. 'Utilitarian moral judgment in psychopathy'. In: *Social cognitive and affective neuroscience* 7.6 (2011), pp. 708–714.

[24] Merton S Krause. 'Use of social situations for research purposes'. In: *American Psychologist* 25.8 (1970), p. 748.

[25] David Magnusson. 'An analysis of situational dimensions'. In: *Perceptual and motor skills* 32.3 (1971), pp. 851–867.

[26] G Alan Marlatt. 'Taxonomy of high-risk situations for alcohol relapse: evolution and development of a cognitive-behavioral model'. In: *Addiction* 91.12s1 (1996), pp. 37–50.

[27] Paul E Meehl. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press, 1954.

[28] Rudolf H Moos. 'Conceptualizations of human environments'. In: *American psychologist* 28.8 (1973), p. 652.

[29] Noureldien A Noureldien. 'A novel taxonomy of MANET attacks'. In: *2015 International Conference on Electrical and Information Technologies (ICEIT)*. IEEE. 2015, pp. 109–113.

[30] Scott Parrigon et al. 'CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics'. In: *Journal of Personality and Social Psychology* 112.4 (2017), p. 642.

[31] Adam M Perkins et al. 'A dose of ruthlessness: Interpersonal moral judgment is hardened by the anti-anxiety drug lorazepam'. In: *Journal of Experimental Psychology: General* 142.3 (2013), p. 612.

[32] Lawrence A Pervin. 'A free-response description approach to the analysis of person-situation interaction'. In: *The Psychology of Social Situations*. Elsevier, 1981, pp. 40–55.

[33] Nicolas Prat, Isabelle Comyn-Wattiau and Jacky Akoka. 'A taxonomy of evaluation methods for information systems artifacts'. In: *Journal of Management Information Systems* 32.3 (2015), pp. 229–267.

[34] Canice Prendergast. 'The provision of incentives in firms'. In: *Journal of economic literature* 37.1 (1999), pp. 7–63.

[35] Richard H Price. 'The taxonomic classification of behaviors and situations and the problem of behavior-environment congruence'. In: *Human Relations* 27.6 (1974), pp. 567–585.

[36] Dawn Querstret and Oliver C Robinson. 'Person, persona, and personality modification: An in-depth qualitative exploration of quantitative findings'. In: *Qualitative Research in Psychology* 10.2 (2013), pp. 140–159.

[37] Lisa Rajbhandari and Einar Snekkenes. 'Using the Conflicting Incentives Risk Analysis Method'. In: *Security and Privacy Protection in Information Processing Systems*. Ed. by Lech J. Janczewski, Henry B. Wolfe and Sujeet Shenoi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 315–329.

[38] Brent W Roberts and Eileen M Donahue. 'One personality, multiple selves: Integrating personality and social roles'. In: *Journal of Personality* 62.2 (1994), pp. 199–218.

[39] J Philippe Rushton, Roland D Chrisjohn and G Cynthia Fekken. 'The altruistic personality and the self-report altruism scale'. In: *Personality and individual differences* 2.4 (1981), pp. 293–302.

[40] Shalom H Schwartz. 'Basic values: How they motivate and inhibit prosocial behavior'. In: *Prosocial motives, emotions, and behavior: The better angels of our nature*. Ed. by Mario Ed Mikulincer and Phillip R Shaver. American Psychological Association, 2010, pp. 221–241.

[41] Saul B Sells. 'An interactionist looks at the environment'. In: *American Psychologist* 18.11 (1963), p. 696.

[42] Kennon M Sheldon et al. 'Trait self and true self: Cross-role variation in the Big-Five personality traits and its relations with psychological authenticity and subjective well-being'. In: *Journal of personality and social psychology* 73.6 (1997), p. 1380.

[43] Steven Shorrock. *Twelve Properties of Effective Classification Schemes*. URL: https://humanisticsystems.com/2018/08/31/twelve-properties-of-effective-classification-schemes/.

[44] Einar Snekkenes. 'Position paper: Privacy risk analysis is about understanding conflicting incentives'. In: *IFIP Working Conference on Policies and Research in Identity Management*. Springer. 2013, pp. 100–103.

[45] Daniel J Solove. 'A taxonomy of privacy'. In: *University of Pennsylvania law review* 154 (2005), p. 477.

[46] Paul M Spengler. 'Clinical versus mechanical prediction'. In: *Handbook of Psychology, Second Edition* 10 (2012).

[47] Adam Szekeres and Einar Arthur Snekkenes. 'Unobtrusive Psychological Profiling for Risk Analysis'. In: *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications - Volume 1: SECRYPT*. INSTICC. SciTePress, 2018, pp. 210–220.

[48] Adam Szekeres, Pankaj Shivdayal Wasnik and Einar Arthur Snekkenes. 'Using Demographic Features for the Prediction of Basic Human Values Underlying Stakeholder Motivation'. In: *Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 2: ICEIS*. SciTePress, 2019, pp. 377–389.

[49] Daniel Szopinski, Thorsten Schoormann and Dennis Kundisch. 'Because Your Taxonomy is Worth IT: towards a Framework for Taxonomy Evaluation'. In: *27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS 2019, Stockholm and Uppsala, Sweden, June 8-14, 2019*. 2019. URL: https://aisel.aisnet.org/ecis2019%5C_rp/104.

[50] Maaike A Ten Berge and Boele De Raad. 'Taxonomies of situations from a trait psychological perspective. A review'. In: *European Journal of Personality* 13.5 (1999), pp. 337–360.

[51] Guus L Van Heck. 'The construction of a general taxonomy of situations'. In: *Personality psychology in Europe: Theoretical and empirical developments* 1 (1984), pp. 149–164.

[52] Bertrand Venard. *Lessons from the massive Siemens corruption scandal one decade later*. URL: http://theconversation.com/lessons-from-the-massive-siemens-corruption-scandal-one-decade-later-108694.

[53] Wikipedia. *Watergate scandal*. URL: https://en.wikipedia.org/wiki/Watergate%5C_scandal.

[54] Dustin Wood and Brent W Roberts. 'Cross-sectional and longitudinal tests of the Personality and Role Identity Structural Model (PRISM)'. In: *Journal of Personality* 74.3 (2006), pp. 779–810.

[55] Yu Yang, Stephen J Read and Lynn C Miller. 'A taxonomy of situations from Chinese idioms'. In: *Journal of Research in Personality* 40.5 (2006), pp. 750–778.

Chapter 13

Article 5: Prediction of threat and opportunity risks: evaluation of a psychological approach using attributes of persons and situations

Adam Szekeres & Einar Arthur Snekkenes. Prediction of threat and opportunity risks: evaluation of a psychological approach using attributes of persons and situations. Under review In: *Risk Analysis: An International Journal*. Wiley-Blackwell. 2020.

Abstract

Information security is rife with human-related threats. The Conflicting Incentives Risk Analysis (CIRA) method distinguishes between threat and opportunity risks, covering a broad range of motivationally and psychologically different human threats to information security. The method's real-world applicability depends on its capability to predict conscious human choices resulting in these risks. Traditional approaches for behavior prediction utilize personal attributes only and achieve low prediction accuracies in general. Therefore, the primary objective of this exploratory study is to evaluate another approach for behavior prediction, which utilizes attributes of persons and situations to achieve improved predictive accuracy. The second objective is to estimate the method's practical feasibility when the

decision-maker's value trade-offs need to be assessed by an observer (i.e. risk analyst). Data was collected from 59 subjects using an online survey to address the research objectives. Results show that the proposed behavior prediction approach outperforms the traditional approach across a wide range of choice situations. The method's real-world performance may be negatively impacted by analysts' limited capability to objectively assess value trade-offs elicited by situations. Potential research directions are outlined to reduce errors associated with analyst subjectivity.

13.1 Introduction

A central ambition of science is to make increasingly accurate predictions. As sciences progress, previously unexplained phenomena become predictable and controllable by humans and it is reasonable to expect that a deeper understanding of risks gives rise to improved techniques for managing and mitigating undesirable events. Since human decision-makers are often identified as the root causes of disastrous incidents, predictions about the future should consider the behavior of key stakeholders responsible for the system's behavior. Active and latent failures have been distinguished by Reason when analyzing human contributions to the breakdown of complex social-technical systems from the 80s [55]. Active failures have an immediate adverse effect and are attributed to direct operators of a system. Latent failures, on the other hand refer to decisions made a long time before incidents and their negative consequences manifest later in combination with other triggers. Latent failures are associated with the decisions of people who are more removed (both in space and time) from the direct human-machine interface such as regulators, managers. A key characteristic of latent failures is that they were present in the system well before the onset of the incident, thus a challenge for risk analysis is to identify and mitigate latent decision failures before they combine with local triggers leading to disastrous consequences.

Since the time of Reason's analysis, great technological advances have been made in the information and communication domain (e.g. widespread internet, IoT, etc.) which became crucial for supporting all aspects of life. The number of tightly coupled social-technical systems rises and decisions with negative consequences for a large number of people have become commonplace. Information security (IS) incidents can be observed in all domains affected by digitization. Individuals on social media are exposed to threats when their privacy is violated by a single unsuspecting node in their personal network which can expose the entire social graph for exploration and exploitation [5]. Major data leaks endanger the privacy of entire voting age populations in certain countries [13]. Trust between countries may be undermined when compromised cryptographic devices are sold by one party to spy on the others [33]. Hacked smart electricity meters may cause great

financial losses for utility companies [36]. An inappropriate strategic decision (e.g. software installed on cars to evade emission tests) can have ripple effects affecting the reputation and sales of other organizations associated with a perpetrator only by nationality [6]. When a cryptographic algorithm endorsed by standardizing agencies is fundamentally flawed, any application relying on that standard becomes vulnerable and opens a back door for abuse [28].

A crucial question is whether such incidents can be predicted and prevented by focusing on the people making the choices. Would a different person make a different choice in a similar situation? Would everyone make similar choices in the same situation irrespective of their personal differences? What are the necessary conditions to predict human choices that affect others? This study aimed at exploring the usefulness of analyzing personal and situational aspects of the decision-making process for the prediction of choices affecting other stakeholders. This study assumes that certain risks are fundamentally attributed to individual stakeholders. This view is well-aligned with organizational practices which specify chains of command to delegate responsibilities and duties to individuals through their roles across the organizational hierarchy [61]. Furthermore, legal systems identify the individual subject (i.e. a natural person) as a responsible and accountable entity for his/her actions and suitable for punishment [47]. The fact that legal systems often penalize organizations (i.e. legal persons) for the actions of their agents may contribute to sub-optimal individual decisions and requires the attention of legal scholars [24].

Motivation - Conflicting Incentives Risk Analysis and human motivation

The Conflicting Incentives Risk Analysis (CIRA) method relies on the assessment of action-desirability of key decision-makers to characterize risks in the domain of IS [54]. Risk is conceptualized as the misalignment between stakeholder incentives, where one party's actions or inactions may be (un)desirable for another party [65]. Two stakeholder categories are distinguished: *Risk owner* - the person facing a risk due to exposure to the (in)actions of the *Strategy owner* - the person who is in the position to take actions which have a potential positive or negative impact on the risk owner. Actions with potential negative consequences are *threat risks*, while actions with positive outcomes which may not get realized represent *opportunity risks*, encompassing a wide variety of motivationally and psychologically distinct behaviors. Both stakeholders are represented by their overall utility, comprising of various utility factors. The strategy owner's motivation is captured by the expected change in its overall utility as a result of implementing an action. Choices are predicted by the analyst who infers the utility factors and their weights from publicly available pieces of information to construct the motivational profile of the strategy owner [71]. Next, the analyst intuitively estimates how potential actions modify the relevant utility factors to assess the desirability of various actions [65].

The method's real-world applicability depends on the effectiveness with which an analyst can predict the strategy owner's behavior.

Previous work has utilized the theory of basic human values (BHV) as a comprehensive organizing framework of human motivation to operationalize utility factors in CIRA [72, 71]. The comprehensiveness of the theory ensures its applicability to a broad range of motivationally distinct conscious behaviors within the scope of CIRA. Values represent abstract trans-situational goals which vary in importance for individuals. Values serve as guiding principles which remain relatively stable during the course of a lifetime [60]. Figure 13.1 shows the circular arrangement of ten values forming a motivational continuum and the four higher-level dimensions are marked outside of the circle [60]. Values close to each other on the circle are more compatible whereas opposing values tend to be in conflict. The theory proposes that conscious choices are guided by the trade-offs between the personal values of a decision-maker. The terminology used throughout this study is summarized in Table 13.1.

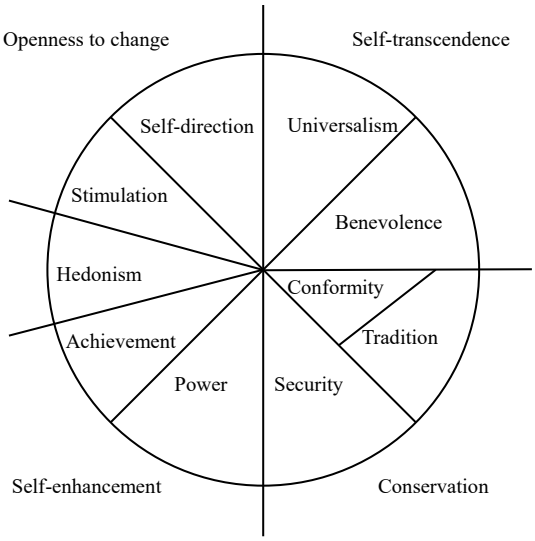


Figure 13.1: Basic human value structure based on [60].

Table 13.1: Terminology used throughout the study.

Term	*Description*
Personal attributes	Characteristics of an individual/person. Utility factors in CIRA. Operationalized throughout the study by the basic human values (BHV).
Situational attributes	Characteristics of a choice situation exerting influence on the decision-maker (e.g. presence of other people, rules, time pressure, etc.).
Value trade-offs	The measurable effects of the situational attributes on the subjective perception of option-desirability in a choice situation. Expected gains/losses due to changes in the values of the relevant utility factors. A subject's perception about the costs/benefits associated with an option of a dilemma.
Traditional prediction approach	Predictive models which utilize personal attributes only as predictors for predicting the outcome of interest (i.e. choice).
Person-situation (P-S) interactionist approach	Combined models which utilize personal and situational attributes together as predictors for predicting the outcome of interest (i.e. choice).

Problem statement

To date, empirical tests are lacking about CIRA's performance for threat and opportunity risks. This work is also motivated by the need to improve behavior prediction capabilities, which are generally low, when traditional methods are used (e.g. 19%-38% of the variance explained by a state-of-the-art behavior prediction method [69]). Thus, an approach which uses personal and situational attributes is proposed and evaluated to test whether improved predictions are achievable. Furthermore, the extent to which value trade-offs can be accurately assessed by analysts is explored to explore the practical feasibility of the approach. The following research questions were formulated:

Research Questions

- RQ 1: To what extent can a P-S interactionist approach provide improved prediction accuracy compared to a traditional approach for predicting stakeholder behavior?
- RQ 2: To what extent is the P-S interactionist approach feasible in real-world settings when P-S interactions need to be objectively assessed by a risk analyst?

Thus, this study has two main objectives. First, to explore and report the potential benefits of using a P-S interactionist approach for behavior prediction. Second, to investigate potential analyst performance with respect to producing accurate, objective assessments about the strategy owner's value trade-offs (i.e. how actions modify the strategy owner's utility factors).

The paper is organized as follows: Section 13.2 presents existing results relevant to the paper's topics. An overview about various human-related threats within IS is

provided, followed by a discussion about situational attributes relevant for decision-making. Section 13.3 presents the characteristics of the sample, the instruments and procedures used for data collection. Results are presented in Section 13.4 organized according to the main research questions. A discussion of the results, their implications and the limitations of the study are presented in Section 13.5. The key findings are summarized in Section 13.6 along with directions for further work.

13.2 Related work

The purpose of this section is to provide an overview of the major classes of conscious human behaviors within CIRA's scope, referred to as threats in the domain of IS, representing a broad range of motivationally and psychologically distinct categories. Conscious decisions can have negative impact on IS in a variety of ways: deliberate misuse (i.e. insiders and hackers); negligence or lack of motivation to comply with policies at the operational level [10], negative side effects of decisions at the strategic level (i.e. externalities) [2]. The overview focuses on the breadth of approaches for predicting undesirable outcomes. The presentation is restricted to conscious, intentional behaviors; excluding the human error paradigm, which is by definition concerned with unintentional, accidental incidents [16]. The final part of the section surveys research results focusing on situational aspects of decision-making from the broader field of psychology. Depending on the maturity of the fields, literature reviews were selected as starting points for demonstrating key research directions or highly influential papers were identified to demonstrate the breadth of approaches.

Predicting human behavior within IS

The need for secure communication may be as old as war [64]. Modern IS is rooted in computer security, dating back to World War II military operations, where key threats were mainly restricted to the physical domain (e.g. theft, espionage and sabotage of computing equipment). With the introduction of networked computing systems during the Cold War-era, IS started to cover a wider range of issues than the physical protection of the machinery. According to [83], the Rand Report R-609 from 1967 was among the first documents to highlight that security risks can no longer be mitigated by practices focusing on physical and hardware security. Thus, the goal became the protection of data by considering technical and personnel-related issues across the organization. Multidisciplinary research activities started addressing the challenges since the beginning of the 2000s [3], but human behavior is still among the top challenges for IS [56].

Malicious intention: insider and external threats

Prediction of human behavior at the operational level is most often investigated in connection with insider threats. Insiders are individuals who break IS rules deliberately. The actions can be motivated by a variety of reasons (e.g. financial gain, curiosity, ideology, political, revenge) [41]. Early investigations took a descriptive approach focusing on associations observed in historical incidents between the insider's activity and their demographic features. The type of insider activity was linked to the position held by the insider in the organization; correlations were found between the insider's age and amount of financial loss caused [50]. A systematic literature review [23] revealed that most of the insider threat prediction applications focus on patterns of online activity as key features, whereas personal attributes of individuals are investigated less frequently.

The subset of empirical works focusing on the psychological attributes of insiders aim at identifying and assessing personal or behavioral characteristics which are valid and reliable indicators of forthcoming incidents. The insider threat prediction model prepared for the U.S. Department of Energy identifies 12 unique psychosocial behavioral indicators (e.g. disgruntlement, stress, absenteeism, etc.) that may be indicative of an insider threat [26]. The U.S. Department of Homeland Security [46] advises organizations to focus on specific personal characteristics to successfully detect insider threats. Personal characteristics include introversion, greed, lack of empathy, narcissism, ethical flexibility, while certain behavioral indicators are also listed (e.g. using remote access, interest in matters outside of scope of duties, risk-taking behavior, etc.). Some of the most useful personal features for predicting insider threats are personality traits (e.g. Big Five, Dark Triad [22], sensation-seeking, etc.), emotions (e.g. hostility, anger) and mental disorders (e.g. paranoia, depression, etc.). Sources of information for assessing these personality features include real-time monitoring of network use [34], social networking sites, criminal record histories, clinical diagnoses, website visit logs, lexicographic analysis of personal communications [8]. Despite research efforts, several challenges impair efficient behavior predictions such as non-stationary data, lack of real-world datasets, high number of interacting features, class imbalances, improper assessment of uncertainty [23].

In the case of external threats, predictions are less prevalent and potentially more difficult due to the unavailability of subjects. However, a taxonomy which categorizes hackers according to their properties (motivations, capabilities, triggers, methods) observed in historical incidents can be useful for defense planning and forensic investigations. Four broad motivational categories have been distinguished for external human threats: revenge, curiosity, financial gain, and notoriety (i.e. fame) [27].

Policy violations

The purpose of a policy is to prescribe expected behaviors and to specify the consequences of undesirable behavior. Even organizations that have IS policies in place, are exposed to risks due to inappropriate employee behavior. Several empirical studies investigate conscious security-related behaviors using the Theory of Planned Behavior (TPB), which is regarded as the most mature model for attitude-based behavior prediction within psychology [1]. TPB specifies that *attitudes*, *subjective norms*, and *perceived behavioral control* are the most important antecedents of *behavioral intention*, which is the ultimate antecedent of actual behavior. The utility of the theory is twofold: (non)-compliant behavior can be predicted from the combination of the factors; interventions can be targeted at specific factors to motivate desirable behavior.

It has been established that the TPB has similar efficiency in predicting behavioral intentions in the domain of IS as in other application domains [66]. Most of the research studies analyzed in [39] and [14] found that all three main constructs of TPB have significant associations with behavioral intentions, providing support for the model's suitability within IS. However, it is important to note that most investigations measure behavioral intentions, not actual behavior. TPB's prediction accuracy is generally measured by the R^2 metric (variance explained) averaging around $R^2 = 0.42$ across studies for intentions [66]; but when actual behavior is measured the prediction accuracy may decrease to as low as $R^2 = 0.1$ [62], which may raise questions about its practical utility.

Negative externalities and moral hazard

Political and economic motives are highly influential in IS. Concepts from economics have been utilized to explain various phenomena at the strategic level of decision-making. Negative externalities and moral hazard are key concepts which can be linked to individual decision-makers [2, 3]. Negative externalities refer to negative side effects of transactions (i.e. private and public decisions with undesirable consequences for other parties, who did not choose to be involved in the transaction). Moral hazard may arise in contractual relationships between a principal and an agent, where the agent's actions may or may not be in the best interest of the principal [45].

Most of the studies use the formalism of game theory to study the relevant concepts. The utility of game theory for analyzing several important problems in IS is demonstrated with examples in [4], also showing how dynamic issues can be modeled by combining evolutionary game theory and the study of network topology. The motivations in the software industry for producing insecure products are modeled

and validated in [84]. Misalignment of incentives between end-users and system operators can be analyzed through the lens of the principal-agent paradigm and game theory can be used to simulate the effects of various interventions (i.e. penalty [80] or positive reinforcement [31]) on end-user behavior. The topic of inter-dependency between organizations has been addressed in several studies [38]. When organizations are highly interdependent free riding may occur (under-investment by some players), whereas low degree of interdependence may result in over-investment in IS, which can be corrected by central authorities [17].

Some artifacts have been developed to detect, analyze or modify misaligned incentives among stakeholders. Individual risk perception is a central topic for identifying perverse incentives among individual stakeholders [18]. An economic modeling framework is proposed in [32] to assist decision-makers with optimal investment in IS.

Only a few studies take an empirical approach, but one example focuses on software vendors who have little incentives to produce secure software, reinforced by the customers' costs associated with switching supplier. Empirical data showed that a software vendors' stock market value significantly drops following the disclosure of a vulnerability in their products [73]. The results can be interpreted as the market value of security. A study from the perspective of legislation analyzes IS breach notifications in addition to court and government records. The analysis highlights the legal system's inefficiency in mitigating agency problems and negative externalities within IS [51].

Situational aspects of decision-making

Situations have fundamental impact on decision-making and behavior. Behaviorism aimed at controlling and predicting human behavior by exploring basic stimulus-response relationships between environment and organisms [81]. The field of persuasion is concerned with creating situations which increase compliance by triggering various psychological processes (e.g. need for consistency, reciprocation, social proof, etc.) [12]. Humans are sensitive to the mere presence of others when performing a task [86]. Human preferences can be reversed by manipulating how the same information is presented using the framing effect [76]. Despite the vast knowledge base of situational influences on behavior accumulated over the decades, systematization of the literature is lacking. The lack of consensus on how to conceptualize, define and measure situations is attributed to the complex and multifaceted nature of situations [44]. Despite the challenges, several situation taxonomies have been developed from various theoretical foundations [70]. Most taxonomies focus on situational features as perceived by the individual in the situation, assuming that behavioral incentives are subjective rather than objective

[74].

Situations have been classified based on their ability to enable or inhibit goal-directed behavior suggesting that situations are perceived and evaluated by individuals relative to their goals [85]. Since the Dark Triad traits are frequently associated with harmful workplace behavior [67], a taxonomy was developed which identifies situational triggers that facilitate the manifestation of these traits [49], which can be used for the development of situational interventions to mitigate risks. A study investigates the impact of situational attributes on leader's decision-making from an ethical perspective [68]. The results showed how the presence of an authoritative figure resulted in ethically questionable decisions and several interactions have been observed between situational attributes (e.g. performance pressure, interpersonal conflict) and the quality of the final decision.

The theory of BHV proposes that the trade-off between competing values guide behavior when the relevant values are activated [60]. However, relatively few studies investigated explicitly the link between situational attributes, value-activation and value trade-offs [19, 78]. One study [19] showed that the valences (attractiveness) assigned for alternative courses of actions correlate well with values in choice situations, which were specifically designed to activate certain target values. In a series of studies using consumer choice problems, it was concluded that values do not influence behavior by default but only when activated (i.e. attention was drawn to value-relevant information by priming stimuli), and when the activated values were central to the self-concept (i.e. important to the individual) [78]. There is some evidence about the practical utility of values for predicting unethical behavior at the individual-level [20], and it has been demonstrated that voting can be predicted by values at group-level [9].

Summary of related work

Psychological and empirical approaches for behavior prediction are widely used in the operational context of IS, specifically in connection with insider threats and compliance. Several research attempts focus on the identification and assessment of personal attributes for the prediction of undesirable behaviors. However, behavior prediction approaches require improvements to increase their practically utility, since accuracies are low when actual behavior is considered. Investigations into negative externalities and moral hazard are dominated by game theoretic approaches, focusing at the organizational level. Only a few studies use empirical methods to investigate these issues and almost none of them focus on the psychological attributes of individual decision-makers, who are ultimately responsible for making decisions. While situational attributes have fundamental influence on decision-making and behavior, the literature lacks systematic, unified theories and

applications to assess these attributes. Value activation and trade-offs represent under-investigated areas within the theory of BHV. Thus, one possibility to achieve improved predictions would require the integration of separate results exploring how choices can be predicted by using personal and situational attributes together. The P-S interactionist view [40] proposes that behavior is a function of personal and situational attributes as perceived by a subject [35].

13.3 Materials and methods

The main objective of the online questionnaire was to collect two types of behavioral responses from subjects: perceptions of value trade-offs in dilemmas representing threat and opportunity risks to model the decision-making process and explicit choices as outcomes to be predicted using two different approaches. Motivational profile information was collected for the traditional prediction approach. Behavioral data was used for the P-S interactionist approach. To assess the practical feasibility of the P-S interactionist approach, the extent of objectivity in value trade-offs across raters needs to be explored. The questionnaire was completely anonymous, no personally identifiable information was collected, participants were required to express consent to participate. The questionnaire was implemented in Limesurvey and was hosted on servers provided by the university. Sections were presented in the following order to maximize the number of tasks between behavioral tasks to increase validity:

1. Evaluation of dilemma-options (value trade-offs).
2. Basic demographic data and personal attributes (BHV profiles).
3. Explicit choice between dilemma-options.

Sample

Based on the sample size recommendations for logistic regression analyses, the data collection aimed at a minimum of 50 fully completed questionnaires [77]. In the first wave of the survey distribution a random sample of university students received an invitation to take part in the online survey, which resulted in 22 fully completed surveys. Therefore, in the next wave, 40 additional respondents were recruited through the Amazon Mechanical Turk (MTurk) online workplace, where subjects receive compensation for completing various human intelligence tasks (HITs). Each respondent who completed the survey received 4 USD net compensation distributed through the MTurk system, which equals to an hourly rate of 12-16 USD. In addition to the higher-than average compensation [30], additional options were selected to ensure data quality: the survey was available only for MTurk workers

with a HIT Approval Rate greater than 90%, and only to Masters (MTurk's quality assurance mechanism). Completed surveys below 9 minutes of completion time were removed to increase the quality of the dataset. Thus, the final convenience sample comprised of 59 respondents with a mean age of 34 years (S.D. = 10.44) including 27 females and 32 males. Citizenship of the respondents was as follows: 53% U.S., 25% Norway, 14% India, 8 % other. Most respondents had bachelor's degree (46%), followed by a completed upper secondary education (36%), master's degree (17%) and lower secondary education (2%).

Measures

Dilemmas representing threat and opportunity risks

The dilemmas were constructed using a previously proposed taxonomy of situations for risk analysis which established a connection between situational attributes and the risk concepts of CIRA by operationalizing threat risks as moral dilemmas and opportunity risks as altruistic dilemmas [70]. The dilemmas aimed at covering the breadth of motivationally distinct behaviors resulting in threat and opportunity risks, which were all presented as riskless choices (i.e. consequences are specified with certainty, as opposed to probabilistic outcomes [15]) with two mutually exclusive options for each dilemma. Table 13.2 provides a short description of the nine dilemmas included in the survey. Some of the dilemmas were inspired by real cases receiving significant media coverage as they resulted in negative outcomes for certain risk owners, representing a decision to increase the ecological validity of the stimuli [11]. The dilemmas were used at the beginning of the questionnaire to collect evaluations (i.e. value trade-offs), as well as at the end of the questionnaire to collect explicit choices from participants taking the role of the strategy owner. It was assumed that no special training or knowledge is required to provide evaluations on the dilemma-options or to make a choice. Each dilemma represented a specific type of risk from the perspective of the risk owner; had a clearly defined victim/beneficiary exposed to the consequences to focus the respondents' attention to the social consequences of their decisions.

Personal attributes - motivational profile

Individual motivational profiles were collected using the Portray Value Questionnaire (PVQ-21), which is a 21-item questionnaire designed for self-assessment [59]. The instrument captures ten BHVs, which were computed according to the instructions provided in [57]. Cronbach-alpha scores measuring the reliability of the instrument were as follows: self-direction 0.52, power 0.69, universalism 0.50, achievement 0.83, security 0.63, stimulation 0.85, conformity 0.64, tradition 0.50, hedonism 0.71, benevolence 0.76. Five value dimensions were created by com-

Table 13.2: Short description of the main theme of the dilemmas included in the survey.

Dilemma Number	Potential trade-offs	Short summary
20	Cause one death actively vs. cause many deaths passively	Kill an injured person to save rest of crew?
22	Abuse of power vs. sexual excitement	Approach employees with sexual offer looking for promotion?
23	Avoid mayhem vs. cause death indirectly	Distribute electricity to residents instead of hospital during electricity crisis?
26	Financial gain vs. risk of punishment	Reprogram customer's Smart Meters for a fee?
28	Responsibility vs. following rules	Inform contractors about security issues identified at employer?
29	Effort vs. benefit for others	Include a patient in clinical trial through a difficult procedure?
32	Living in exile vs. freedom of research	Create paywall bypassing website to make research results freely available?
34	Productivity lost vs. responsibility	Running a virus scan for colleagues?
36	Number of holidays vs. salary	Accept unfavorable job offer?

puting the mean of the corresponding values as follows: *self-enhancement:* power and achievement, *self-transcendence:* universalism and benevolence, *openness to change:* self-direction and stimulation, *conservation:* security, tradition, conformity, while *hedonism* was treated as a separate dimension, due to its instability in the value hierarchy [59].

Value trade-offs

To capture the perceived losses/benefits obtained from a particular choice, participants were asked to evaluate both options of all dilemmas on five value dimensions of the BHV theory (the evaluations were unrelated to subjects' motivational profiles collected by the PVQ-21 instrument). Value trade-offs were collected by using continuous sliding scales ranging from negative 100 through 0 to positive 100, with textual anchor labels at the two endpoints and at the mid-point of the scale (-100: Maximum possible decrease; 0: No impact; 100: Maximum possible increase). Value dimensions were presented as textual descriptions of desirable end-goals associated with the dimensions (i.e. *Experiencing pleasure, success, social status and prestige.* - representing self-enhancement) [58]. For each dilemma-option the following text was presented once: *By considering the consequences please rate how XX would influence each of the following factors from your perspective compared to your state before the decision*, where XX was replaced by the action/inaction described in the dilemma and ratings were requested on each of the five factors/value dimensions (i.e self-enhancement, self-transcendence, openness

to change, conservation, hedonism). The overall utility of each dilemma-option was calculated by summing the evaluations across attributes as follows: $U_{\text{total}} = U_{\text{self-enhancement}} + U_{\text{self-transcendence}} + U_{\text{openness to change}} + U_{\text{conservation}} + U_{\text{hedonism}}$, using an unweighted version of the multi-attribute utility theory (MAUT) [21] implemented in CIRA [54]. Since respondents were required to provide subjective evaluations for the dilemma-options it was assumed that evaluations represent the combined effects of personal values and the contributions of the options on the utility factors. In short, the evaluation method required participants to explicitly rate five value dimensions for both options of all dilemmas which were used to compute the utility associated with each dilemma-option.

Data processing

Internal consistency of choices.

A choice was considered internally consistent when the explicitly selected option (section 3 of questionnaire) got a higher calculated utility score than the other option of the dilemma using the evaluations (section 1 of the questionnaire). This metric may be an indication of data validity (i.e. respondents were following instructions and providing evaluations based on their preferences), subject rationality (i.e. making choices according to stated preferences) and difficulty of making a choice.

Choice-matched evaluations.

The combined approach for prediction used the value trade-offs (section 1 of questionnaire) matched with the chosen option (section 3 of questionnaire) for each subject across dilemmas. Thus, if a subject selected *option 0* in the forced-choice task on *dilemma d*, the value trade-offs for *option 0* of *dilemma d* were used as predictors, even if the total utility was higher for *option 1* of *dilemma d*. Note that the chosen option was not necessarily the option which received a highest overall utility score thus, internal inconsistency was permitted.

Analyst performance for producing objective value trade-offs.

For the purpose of understanding the accuracy with which raters can objectively assess the value trade-offs in a dilemma-option, data was prepared as follows. Since dilemma options represent independent objects that were rated by subjects, for each dilemma-option (18 in total) a separate dataset was created using the value trade-offs provided by respondents. Each subject was represented in the columns and the value trade-offs were entered as rows in each dataset following the guidelines of [63].

13.4 Results

This section reports the results of all the analyses addressing the main research questions. The analyses were conducted in IBM SPSS 25. Figure 13.2 presents an overview of dilemma characteristics and descriptive statistics across all dilemmas. Red bars indicate the percentage of affirmative choices provided by subjects in the third section of the questionnaire (i.e. threat risk realized for risk owner, opportunity risk avoided for risk owner) across dilemmas. Blue bars represent the percentage of internally consistent choices (when the value trade-offs collected in the first section of the questionnaire are compared with explicit choices collected in the third section of the questionnaire).

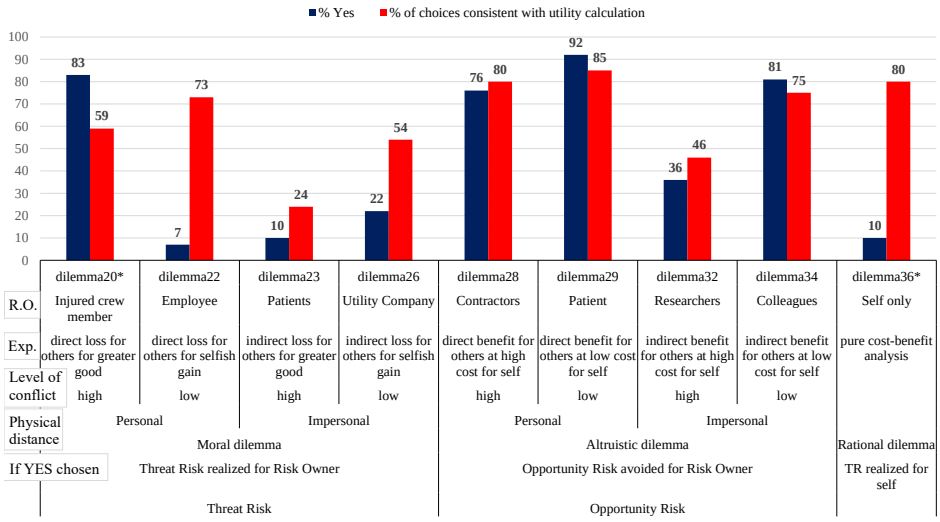


Figure 13.2: Overview of dilemma characteristics and descriptive statistics about choices across dilemmas. Percentage of affirmative choices and percentage of choices consistent with utility calculations derived from subjective value trade-offs across dilemmas. The YES options capture threat risks realized and opportunity risks avoided for the risk owners. Dilemmas are organized according to risk types. Dilemmas marked with * were taken from [25]. *R.O.: risk owner, Exp.: explanation.*

RQ 1: Comparison of approaches for prediction

In order to evaluate the prediction accuracies between the traditional approach and the P-S interactionist approach two separate sets of analyses were conducted. The following sections present the set of analyses and the comparison separately. The traditional approach uses personal attributes for predicting the outcomes, the P-S interactionist approach uses value trade-offs for predicting the same outcomes. Dilemmas represent the unit of analysis for the purpose of choice prediction.

Traditional approach

Nine binary logistic regression models were built (one for each dilemma) using personal attributes as independent variables and explicit choices as dependent variables. Table 13.3 presents each model with the regression coefficients and corresponding tests of significance for each predictor. In total, five out of the nine predictive models are significantly better than the intercept-only models based on the Overall model evaluation row of the table. Predictive performance for each model is assessed by two variants of the R^2 (total variance explained) metric: Cox & Snell R^2, Nagelkerke R^2. Significant coefficients may indicate which values are activated in specific dilemmas.

Table 13.3: Summary of nine binary logistic regression models for each dilemma. Each model uses personal attributes (BHVs) of subjects as independent variables (predictors) to predict choices (outcome).

Predictor	β	*SE* β	p
Constant	9.04[a], -12.83[b], -5.53[c], -4.39[d], -1.67[e], 7.74[f], -7.94[g], 5.58[h], -10.72[i]	4.01[a], 7.42[b], 3.34[c], 2.54[d], 2.14[e], 4.64[f], 2.86[g], 3.31[h], 5.30[i]	**0.02***[a], 0.08[b], 0.10[c], 0.09[d], 0.44[e], 0.10[f], **0.01***[g], 0.09[h], **0.04***[i]
Prof_Self-enhancement	-0.36[a], 3.12[b], 1.31[c], 0.44[d], -0.02[e], -0.61[f], 0.45[g], -1.48[h], 1.01[i]	0.65[a], 1.90[b], 0.92[c], 0.52[d], 0.45[e], 0.91[f], 0.44[g], 0.64[h], 0.87[i]	0.58[a], 0.10[b], 0.15[c], 0.40[d], 0.97[e], 0.50[f], 0.30[g], **0.02***[h], 0.24[i]
Prof_Self-transcendence	-0.62[a], -0.47[b], -1.13[c], -0.75[d], 0.67[e], 1.18[f], 0.17[g], 0.11[h], -0.51[i]	0.74[a], 1.34[b], 0.79[c], 0.60[d], 0.52[e], 1.11[f], 0.58[g], 0.79[h], 1.11[i]	0.40[a], 0.73[b], 0.15[c], 0.21[d], 0.20[e], 0.29[f], 0.76[g], 0.89[h], 0.65[i]
Prof_Openness to change	0.53[a], -2.18[b], -0.63[c], -0.01[d], 0.39[e], -1.03[f], 0.33[g], -0.26[h], 0.83[i]	0.71[a], 1.80[b], 0.93[c], 0.62[d], 0.49[e], 1.25[f], 0.53[g], 0.76[h], 1.14[i]	0.46[a], 0.23[b], 0.50[c], 0.99[d], 0.43[e], 0.41[f], 0.54[g], 0.73[h], 0.47[i]
Prof_Conservation	-1.49[a], 1.43[b], 0.30[c], 1.00[d], -0.06[e], -1.06[f], 0.71[g], -0.19[h], 0.49[i]	0.67[a], 1.45[b], 0.71[c], 0.52[d], 0.40[e], 0.82[f], 0.42[g], 0.47[h], 0.66[i]	**0.03***[a], 0.32[b], 0.67[c], **0.05***[d], 0.87[e], 0.19[f], 0.09[g], 0.69[h], 0.46[i]
Prof_Hedonism	0.27[a], 1.11[b], 1.38[c], 0.32[d], -0.39[e], 0.01[f], 0.25[g], 0.43[h], 0.46[i]	0.39[a], 0.98[b], 0.74[c], 0.39[d], 0.36[e], 0.66[f], 0.34[g], 0.48[h], 0.67[i]	0.49[a], 0.26[b], 0.06[c], 0.41[d], 0.27[e], 0.99[f], 0.47[g], 0.37[h], 0.49[i]
Test	χ^2	*df*	p
Overall model evaluation	*11.80[a], 11.14[b], 11.15[c], 10.08[d], 4.51[e], 7.86[f], 14.17[g], 11.89[h], 10.90[i]*	*5[a, b, c, d, e, f, g, h, i]*	***0.04***[a], **0.05***[b], **0.05***[c], 0.07[d], 0.48[e], 0.16[f], **0.02***[g], **0.04***[h], 0.06[i]*
Goodness-of-fit-tests:			
Cox and Snell R^2	0.18[a], 0.17[b], 0.17[c], 0.16[d], 0.07[e], 0.13[f], 0.21[g], 0.18[h], 0.17[i]		
Nagelkerke R^2	0.30[a], 0.44[b], 0.36[c], 0.24[d], 0.11[e], 0.28[f], 0.29[g], 0.30[h], 0.35[i]		

Note. ***p $\leq$ 0.05.**
a = dilemma20, b = dilemma22, c = dilemma23, d = dilemma26, e = dilemma28, f = dilemma29, g = dilemma32, h = dilemma34, i = dilemma36
Prof: profile scores from PVQ-21.

Person-situation interactionist approach

This set of analyses aimed at exploring the extent of potential improvements that can be expected when the value trade-offs (representing the combined effect of personal and situational attributes) are used to predict the same outcomes. Table 13.4 presents the details of the nine logistic regression models which relied on the choice-matched subjective value trade-offs as predictors. Regression coefficients and corresponding tests of significance for each predictor are presented for all dilemmas. With the exception of models *b* and *f*, all predictive models are significantly better than the intercept-only models as demonstrated by the Overall model evaluation row of the table. Predictive performance of each model is evaluated by Cox & Snell R^2, Nagelkerke R^2 metrics.

Table 13.4: Summary of nine binary logistic regression models for each dilemma. Each model uses the subjective value trade-offs assessed on the five basic human value dimensions as independent variables (predictors) for predicting the outcome (choice).

Predictor	β	*SE* β	*p*
Constant	0.90[a], -4.66[b], -2.87[c], -2.41[d], -0.63[e], 0.14[f], -1.46[g], 1.04[h], -3.31[i]	0.64[a], 1.72[b], 0.90[c], 0.66[d], 0.72[e], 0.96[f], 0.60[g], 0.48[h], 1.11[i]	0.16[a], **0.01***[b], **0.00***[c], **0.00***[d], 0.38[e], 0.89[f], **0.02***[g], **0.03***[h], **0.00***[i]
Eval_Self-enhancement	-0.04[a], 0.02[b], -0.01[c], 0.01[d], -0.02[e], -0.02[f], 0.02[g], -0.02[h], -0.06[i]	0.01[a], 0.02[b], 0.02[c], 0.02[d], 0.02[e], 0.03[f], 0.03[g], 0.02[h], 0.03[i]	**0.01***[a], 0.34[b], 0.45[c], 0.75[d], 0.40[e], 0.55[f], 0.48[g], 0.23[h], **0.03***[i]
Eval_Self-transcendence	0.02[a], 0.00[b], 0.04[c], 0.03[d], 0.06[e], 0.05[f], -0.03[g], 0.02[h], 0.03[i]	0.01[a], 0.02[b], 0.02[c], 0.01[d], 0.02[e], 0.03[f], 0.03[g], 0.01[h], 0.02[i]	0.08[a], 0.99[b], **0.02***[c], **0.02***[d], **0.00***[e], 0.10[f], 0.26[g], 0.12[h], 0.23[i]
Eval_Openness to change	-0.02[a], 0.06[b], 0.00[c], -0.01[d], 0.00[e], 0.01[f], 0.06[g], 0.00[h], 0.08[i]	0.01[a], 0.03[b], 0.02[c], 0.02[d], 0.03[e], 0.02[f], 0.03[g], 0.01[h], 0.03[i]	0.32[a], **0.05***[b], 0.97[c], 0.43[d], 0.99[e], 0.81[f], **0.01***[g], 0.91[h], **0.01***[i]
Eval_Conservation	0.02[a], -0.03[b], 0.01[c], -0.03[d], 0.00[e], 0.01[f], -0.03[g], 0.03[h], 0.00[i]	0.01[a], 0.02[b], 0.02[c], 0.01[d], 0.02[e], 0.02[f], 0.02[g], 0.01[h], 0.02[i]	0.12[a], 0.19[b], 0.48[c], **0.03***[d], 0.82[e], 0.68[f], **0.03***[g], **0.03***[h], 0.96[i]
Eval_Hedonism	0.01[a], -0.02[b], -0.01[c], 0.04[d], 0.00[e], 0.01[f], 0.02[g], -0.02[h], -0.02[i]	0.01[a], 0.02[b], 0.02[c], 0.02[d], 0.02[e], 0.02[f], 0.02[g], 0.01[h], 0.02[i]	0.36[a], 0.37[b], 0.69[c], **0.04***[d], 0.92[e], 0.65[f], 0.29[g], 0.22[h], 0.28[i]
Test	χ^2	*df*	*p*
Overall model evaluation	*21.00[a], 10.40[b], 19.45[c], 22.97[d], 37.13[e], 8.50[f], 47.65[g], 19.15[h], 12.64[i]*	*5[a, b, c, d, e, f, g, h, i]*	***0.00**[a], 0.07[b], **0.00***[c], **0.00***[d], **0.00***[e], 0.13[f], **0.00***[g], **0.00***[h], **0.03***[i]
Goodness-of-fit-tests:			
Cox and Snell R^2	0.30[a], 0.16[b], 0.28[c], 0.32[d], 0.47[e], 0.13[f], 0.55[g], 0.28[h], 0.19[i]		
Nagelkerke R^2	0.50[a], 0.41[b], 0.58[c], 0.50[d], 0.70[e], 0.31[f], 0.76[g], 0.45[h], 0.40[i]		

Note. ***p $\leq$ 0.05.**
a = dilemma20, b = dilemma22, c = dilemma23, d = dilemma26, e = dilemma28, f = dilemma29, g = dilemma32, h = dilemma34, i = dilemma36
Eval: subjective value trade-off evaluations collected in section 1 of the questionnaire.

Comparison of approaches

Table 13.5 presents a summary of the nine logistic regression models' predictive performance across dilemmas using personal attributes only ("Personal attributes only" columns) for predicting the outcomes and subjective value trade-offs provided by subjects for the chosen option ("Person-situation attributes" columns). Traditional models were outperformed by P-S interactionist models across all dilemmas with the exception of dilemma 22.

Table 13.5: Comparison of the two approaches for predicting identical outcomes.

	% of overall correct classification		% of variance explained (Nagelkerke's R^2)	
	Personal attributes only	Person-situation attributes	Personal attributes only	Person-situation attributes
Dilemma20	86.4	**89.8***	30	**50***
Dilemma22	93.2	91.5	44	41
Dilemma23	86.4	**94.9***	36	**58***
Dilemma26	74.6	**81.4***	24	**50***
Dilemma28	76.3	**91.5***	11	**70***
Dilemma29	89.8	**93.2***	28	**31***
Dilemma32	71.2	**91.5***	29	**76***
Dilemma34	81.4	**86.4***	30	**45***
Dilemma36	91.5	**93.2***	35	**40***
Note. * improvement of predictive accuracy from personal attributes-only model				

RQ 2: Practical feasibility of the P-S interactionist approach

The second research question is concerned with exploring the potential accuracy which could be expected when value trade-offs must be assessed by a risk analyst (i.e. subjective evaluations are not available). This task requires the identification of the affected values in the motivational hierarchy (value activation) when different outcomes are evaluated. Furthermore, it requires assessing the magnitude of impact on the affected values (i.e. value trade-offs). The agreement between observers about the value trade-offs can indicate the extent of objectivity observable in situations. From the practical perspective it is necessary to identify the extent to which different risk analysts may arrive at similar evaluations about the value trade-offs that are involved in the decision-making process.

In order to answer this research question the dilemma-options were the units of analysis (i.e. two options for each dilemma) and inter-rater reliabilities were analyzed for each dilemma-option. Intraclass correlations (ICC) are used as estimates of

inter-rater reliability, a technique which is useful for understanding the proportion of reliable ("real") estimates provided by independent raters about a construct or a combination of constructs [37]. As respondents represent a sample from the population of potential respondents and all dilemma options were evaluated by all raters a Two-way random analysis was selected ICC(2), which assumes that the variance of raters adds noise to the estimation of objects, where errors even out as the number of raters is increased [43]. The Two-way random effects technique assumes random effects for raters as well as for the objects being rated, and that both raters and objects are randomly drawn from a larger population of raters and objects. Two types of reliability scores (consistency and absolute agreement) were computed to assess the accuracies of the ratings using ICC(2,1). The difference between the consistency and absolute agreement measures is that "if two variables are perfectly consistent, they don't necessarily agree. For example, consider Variable 1 with values 1, 2, 3 and Variable 2 with values 7, 8, 9. Even though these scores are very different, the correlation between them is 1 – so they are highly consistent but don't agree" [37]. Thus, absolute agreement is a more restrictive measure of inter-rater reliability which may be more relevant for practical settings where the magnitude of a choice's impact on the utility factors is crucial.

Figure 13.3 presents the intraclass correlation scores as a measure of interrater reliability for all dilemmas. Consistency (red bars) refers to the extent of agreement about the direction of the value trade-offs from a randomly selected analyst. Absolute agreement (blue bars) represents the expected accuracy when a single analyst estimates the exact magnitude of the value trade-off. The interpretation of the results is as follows: "an interrater reliability estimate of 0.80 would indicate that 80% of the observed variance is due to true score variance or similarity in ratings between coders, and 20% is due to error variance or differences in ratings between coders" [29]. Except for dilemma_29's No option, all intraclass correlations were statistically significant (i.e. the probability of observing these results due to random chance is $\leq 5\%$).

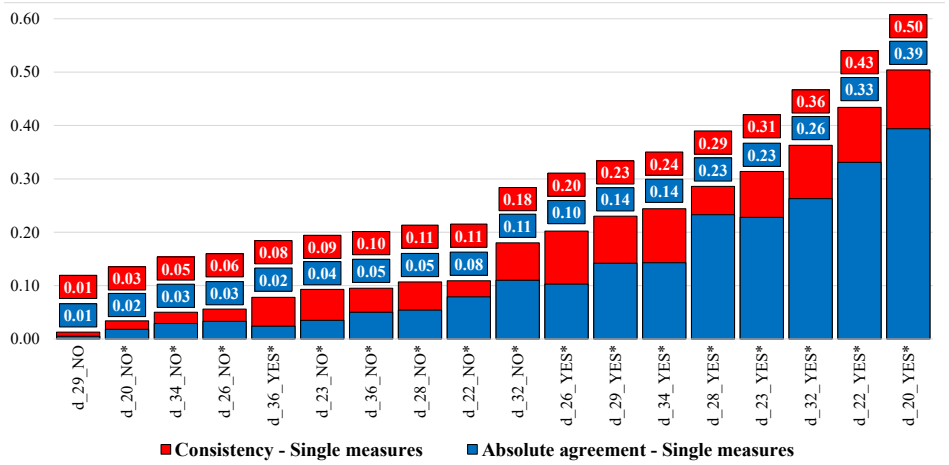


Figure 13.3: Interrater reliability estimates across all dilemma options sorted according to increasing levels of agreement in terms of the consistency and absolute agreement measure. Red bars indicate the consistency with which a randomly selected rater could capture the direction of value-trade-offs, blue bars represent absolute agreement. Dilemma-options marked with * are statistically significant at $p \leq 0.05$.

13.5 Discussion

Predictability is the essence of security. The antecedents and consequences of human behavior are well-explored in the operational domain, where policy violations and threats to IS can be operationalized and measured relatively well. However, attempts to predict decisions at the strategic level are mostly restricted to simulations and game theoretical models at the organizational level. The dearth of empirical studies focusing on the prediction of individual stakeholder behavior may be attributed to several factors: ambiguity and complexity of the environment in which strategic decisions take place [53]; ill-defined measures of success and good decisions; lack of empirical data about decisions. Even though sophisticated tools (e.g. Analytic Hierarchy Process) have been developed to aid decision-makers, a person has to develop a set of measures and evaluate them to compare alternatives, thus subjective judgements and considerations are unavoidable and are inherent in every decision. In summary, while tools can improve decisions, they do not replace the decision-making individual [7]. Furthermore, there is evidence suggesting that real-world IS decision-makers do not utilize standardized decision processes developed by academia; evaluation processes, security metrics hardly exist; and learning takes place in an ad-hoc fashion [82], indicating that there is a serious need to decrease the gap between theories and practice to improve decisions [52]. Finally, since latent failures often creep in at the managerial level of decision-making [55]

and individuals are fundamentally responsible for decisions, more empirical work is needed in the field which can be transferred to practical contexts.

Therefore, this study aimed at exploring the predictability of individual's choices through dilemmas which were designed to capture the breadth of motivationally distinct risk types identified in the CIRA method. The first research question focused on exploring the extent of predictive improvement which can be achieved when traditional approaches to prediction (personal attributes only) are complemented by value trade-offs (a choice's impact on the utility factors). Personal attributes were operationalized using the theory of BHVs. Two sets of analyses were conducted to enable a clear comparison between the two approaches. The overall percentage of correct classifications ranges between 71.2%-93.2% when only personal attributes are utilized, and between 81.4%-94.9% using a P-S interactionist approach. Predictive performance in terms of Nagelkerke's R^2 performance metric consolidates the findings for the combined method's superiority. Nagelkerke's R^2 scores for the BHV-only models range between: 11%-44%, while for the P-S interactionist approach, performance ranges between: 31%-76%.

The second research question aimed at exploring the extent to which analysts can objectively assess how situations impact the decision-maker's value hierarchy and subsequent decisions. Intraclass correlations were used as estimates of interrater reliability to explore the extent of agreement between subjects about value-activation and value trade-offs. The highest accuracy (0.5 consistency, 0.39 absolute agreement) was achieved for the YES option of dilemma20 (shooting an injured crew member to save the rest of the crew), which is a classical dilemma from moral decision-making research. Dilemma36, which was included as a purely rational control dilemma, received a high number of internally consistent (signifying that evaluations provided by subjects were valid), correct responses (selecting the option with higher utility) which signifies that most of the respondents were following instructions properly. However, the agreement for both options of this dilemma were relatively low (8% and 10%), indicating that objectively well-quantifiable aspects of a situation (e.g. amount of salary traded off for number of vacation days) are perceived largely subjectively by respondents giving rise to significant disagreement.

Threat risks and opportunity risks in CIRA are emergent properties resulting from the interaction between the strategy owner's behavior and the risk owner's exposure to the consequences of actions. Due to this complexity, it is challenging to operationalize such risks succinctly, which represents a limitation in the methodology. The dilemmas developed for this study aimed at solving this problem by mapping the two risk types to moral and altruistic dilemmas. While several dilemmas were developed from realistic historical cases, they were presented as hypothetical stories

to respondents. Even though self-report questionnaires are the most-widely used formats for assessing attitudes and personal attributes, they may be prone to several problems (e.g. socially desirable responding, lack of required self-awareness to report why certain actions are chosen, etc.) hampering the validity of the results [42]. To ensure validity, this study used a validated questionnaire (PVQ-21) for the assessment of motivational profiles and the anonymous nature of the online data collection procedure could facilitate the expression of socially undesirable intentions, thus contributing to increased validity [48]. Another limitation is related to the composition and size of the sample. The main purpose of the study was to evaluate the predictive capabilities of two different approaches; therefore, a convenience sample was used, which can be useful for validating the approaches, but may have limited transferability to other populations.

Since risk analysis is a high-risk modeling activity, it is important to identify potential sources of error which may contribute to decreased behavior prediction capabilities in real settings. Based on [79] the following types of errors can be distinguished:

- **1.** Model error refers to the correctness or completeness of the included variables in the model.
- **2.** Parameter error refers to uncertainty of measurements. This error may arise due to limited amount of data and in case of dynamic systems, future states may be difficult to estimate.
- **3.** Stochastic error refers to other errors even when the model and parameters are correct.

These error categories are illustrated in Figure 13.4 which represents the model of the strategy owner's decision-making process from the perspective of the risk analyst.

Model error captures potential inaccuracies with which the decision-making process is modeled. If the model's components are weak predictors of decisions, then assessment of other personal attributes or inclusion of additional variables may be necessary to improve the model. There are two main parameter errors which arise in practical situations. The first one is related to the uncertainty with which stakeholder motivational profiles can be constructed from observational data. These errors have been explored in [71]. A part of the present study (RQ 2.) explored the second source of uncertainty in the category of parameter errors, which is related to the accuracy with which situational impacts can be accurately and objectively assessed to capture value trade-offs. Another part of the present study explored the

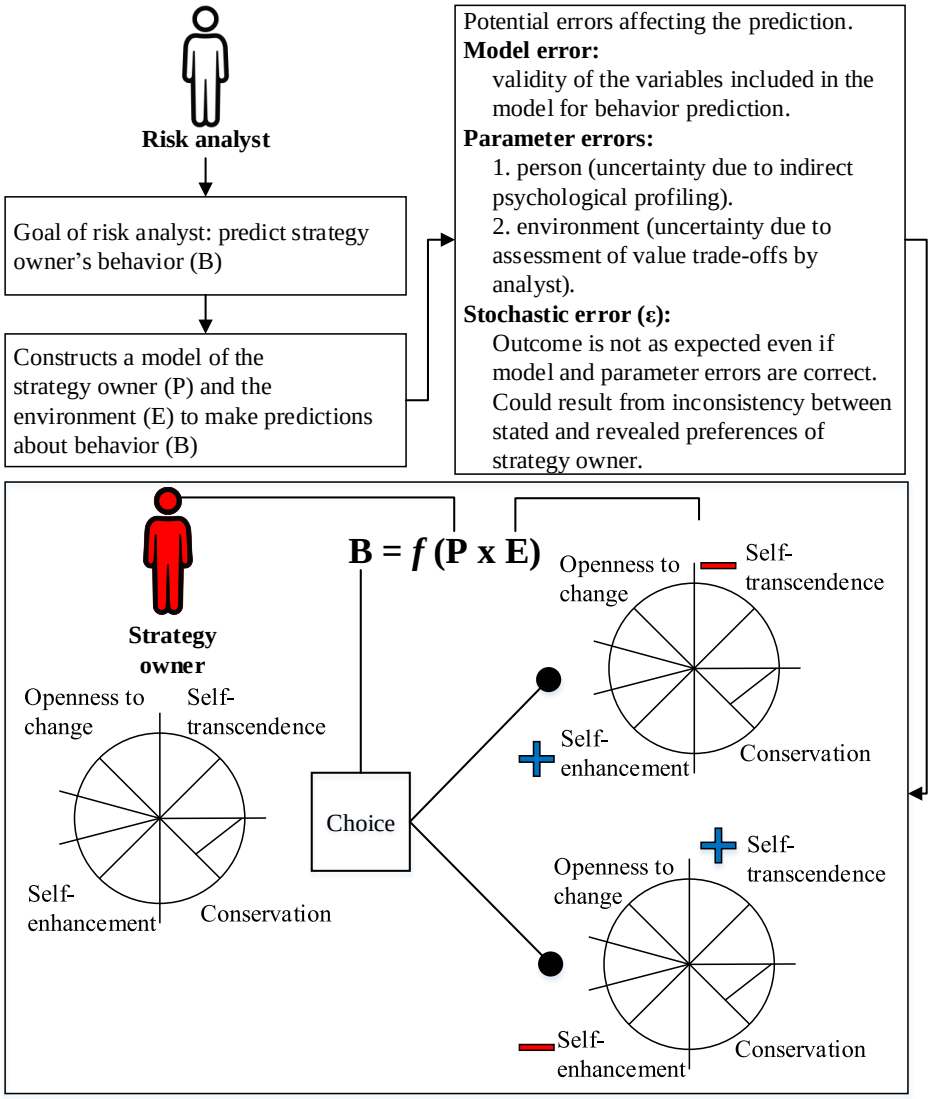


Figure 13.4: Abstraction of the Strategy owner's decision-making process by the risk analyst highlighting three main sources of potential errors (i.e. model error, parameter error and stochastic error according to [79]). Behavior (B) is assumed to result from the interaction between attributes of the person (P) and attributes of the environment/situation (E) [40].

errors related to the consistency between stakeholder preferences and actual choices (i.e. internal consistency of choice). Internal inconsistencies can be categorized

into the stochastic error category which may be inherent in human decision-making processes.

13.6 Conclusions and further work

IS is a complex field, in which people and technology are intertwined in a variety of ways. There is clearly a need for predicting human behavior at all levels of the interaction. An overview of the relevant literature showed that attempts for predicting human behavior at the strategic level in IS seldom focus on the psychology of individual decision-makers who are ultimately responsible for the highest-impact outcomes. Therefore, this study aimed at contributing to the field by enhancing the predictive capabilities of the CIRA method, which focuses on decision-makers' motivation when defining risk. This study proposed and evaluated a behavior prediction approach, which uses personal and situational attributes in combination. The feasibility of assessing situations by the elicited value trade-offs was explored, which has key implications for practical applications. While the utility of the P-S interactionist approach was demonstrated, some issues require further investigations.

Replication studies may benefit from using probability sampling methods from specific populations. While minimum sample size requirements have been fulfilled, uncertainties can be decreased by collecting data from more respondents. Future studies could explore whether it is possible to increase the accuracy of value trade-off assessments by observers. This could be achieved by training analysts in situation-assessment and by developing methods which specify more precisely the mappings between quantifiable situational aspects (e.g. amount of salary vs. number of vacation days) and motivational constructs. Furthermore, the development of automated situation-assessments would be necessary to increase reliability. The present study used dilemmas in which "the decision's impact on other people" can be considered the most salient feature of the environment. While most strategic decisions in for-profit organizations are made on the basis of the expected financial impacts, future studies could explore the effect of presenting everyday dilemmas to decision-makers by emphasizing the implications of a decision on the people affected. In other words, it would be important to conduct more work to study the effects of presenting risks to strategy owners as they impact potential risk owners (in contexts where such interaction is possible). Evidence shows that salient features in the problem formulation (framing of a decision) have very important implications for the outcomes [75]. Would it be possible to use this effect to mitigate some of the risks by modifying the way information is presented to decision-makers?

References

[1] Icek Ajzen. 'The theory of planned behavior'. In: *Organizational behavior and human decision processes* 50.2 (1991), pp. 179–211.

[2] Ross Anderson. 'Why information security is hard-an economic perspective'. In: *Seventeenth Annual Computer Security Applications Conference*. IEEE. 2001, pp. 358–365.

[3] Ross Anderson and Tyler Moore. 'Information security: where computer science, economics and psychology meet'. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 367.1898 (2009), pp. 2717–2727.

[4] Ross Anderson et al. 'Incentives and information security'. In: *Algorithmic Game Theory* (2007), pp. 633–649.

[5] Edmund L. Andrews. 'The Science Behind Cambridge Analytica: Does Psychological Profiling Work?' In: *Stanford Graduate School of Business* (Apr. 2018). [Online; accessed 14. Feb. 2020]. URL: `https://www.gsb.stanford.edu/insights/science-behind-cambridge-analytica-does-psychological-profiling-work`.

[6] Rüdiger Bachmann, Gabriel Ehrlich and Dimitrije Ruzic. *Firms and collective reputation: The Volkswagen emission scandal as a case study*. Tech. rep. CESifo Working Paper, 2017.

[7] Lawrence D Bodin, Lawrence A Gordon and Martin P Loeb. 'Evaluating information security investments using the analytic hierarchy process'. In: *Communications of the ACM* 48.2 (2005), pp. 78–83.

[8] Christopher R Brown, Alison Watkins and Frank L Greitzer. 'Predicting insider threat risks through linguistic analysis of electronic communication'. In: *2013 46th Hawaii International Conference on System Sciences*. IEEE. 2013, pp. 1849–1858.

[9] Gian Vittorio Caprara et al. 'Personality and politics: Values, traits, and political choice'. In: *Political psychology* 27.1 (2006), pp. 1–28.

[10] James L Cebula and Lisa R Young. *A taxonomy of operational cyber security risks*. Technical note CMU/SEI-2010-TN-028. Hanscom, MA: Carnegie-Mellon Software Engineering Institute, 2010.

[11] Naomi Chaytor, Maureen Schmitter-Edgecombe and Robert Burr. 'Improving the ecological validity of executive functioning assessment'. In: *Archives of clinical neuropsychology* 21.3 (2006), pp. 217–227.

[12] Robert B. Cialdini. *Influence: The psychology of persuasion*. HarperCollins, New York, 2007.

[13] Catalin Cimpanu. 'Voter records for 80% of Chile's population left exposed online'. In: *ZDNet* (Aug. 2019). URL: https://www.zdnet.com/article/voter-records-for-80-of-chiles-population-left-exposed-online.

[14] W Alec Cram, Jeffrey Proudfoot and John D'Arcy. 'Seeing the forest and the trees: A meta-analysis of information security policy compliance literature'. In: *Proceedings of the 50th Hawaii International Conference on System Sciences*. 2017, pp. 4051–4060.

[15] Ward Edwards. 'The theory of decision making'. In: *Psychological bulletin* 51.4 (1954), p. 380.

[16] Ivan Enrici, Mario Ancilli and Antonio Lioy. 'A psychological approach to information technology security'. In: *3rd International Conference on Human System Interaction*. IEEE. 2010, pp. 459–466.

[17] Mansooreh Ezhei and Behrouz Tork Ladani. 'Interdependency analysis in security investment against strategic attacks'. In: *Information Systems Frontiers* (2018), pp. 1–15.

[18] Fariborz Farahmand, Mikhail Atallah and Benn Konsynski. 'Incentives and perceptions of information security risks'. In: *ICIS 2008 Proceedings* (2008), p. 25.

[19] Norman T Feather. 'Values, valences, and choice: The influences of values on the perceived attractiveness and choice of alternatives'. In: *Journal of personality and social psychology* 68.6 (1995), p. 1135.

[20] Gilad Feldman et al. 'The motivation and inhibition of breaking the rules: Personal values structures predict unethicality'. In: *Journal of Research in Personality* 59 (2015), pp. 69–80.

[21] Gregory W Fischer. 'Experimental applications of multi-attribute utility models'. In: *Utility, probability, and human decision making*. Springer, 1975, pp. 7–46.

[22] Donelson R Forsyth, George C Banks and Michael A McDaniel. 'A meta-analysis of the Dark Triad and work behavior: a social exchange perspective'. In: *Journal of applied psychology* 97.3 (2012), p. 557.

[23] Iffat A Gheyas and Ali E Abdallah. 'Detection and prediction of insider threats to cyber security: a systematic literature review and meta-analysis'. In: *Big Data Analytics* 1.1 (2016), p. 6.

[24] Gregory M Gilchrist. 'Individual Accountability for Corporate Crime'. In: *Georgia State University Law Review, Forthcoming* (2017).

[25] Joshua D Greene et al. 'An fMRI investigation of emotional engagement in moral judgment'. In: *Science* 293.5537 (2001), pp. 2105–2108.

[26] Frank L Greitzer et al. *Identifying at-risk employees: A behavioral model for predicting potential insider threats*. Technical note PNNL-19665. Richland, WA: Pacific Northwest National Lab.(PNNL), 2010.

[27] Sara LN Hald and Jens M Pedersen. 'An updated taxonomy for characterizing hackers according to their threat properties'. In: *Advanced Communication Technology (ICACT), 2012 14th International Conference on*. IEEE. 2012, pp. 81–86.

[28] Thomas C Hales. 'The NSA back door to NIST'. In: *Notices of the AMS* 61.2 (2013), pp. 190–19.

[29] Kevin A Hallgren. 'Computing inter-rater reliability for observational data: an overview and tutorial'. In: *Tutorials in quantitative methods for psychology* 8.1 (2012), p. 23.

[30] Kotaro Hara et al. 'A data-driven analysis of workers' earnings on Amazon Mechanical Turk'. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pp. 1–14.

[31] Tejaswini Herath and Raghav Rao. 'Control mechanisms in information security: a principal agent perspective'. In: *International Journal of Business Governance and Ethics* 5.1-2 (2010), pp. 2–13.

[32] Borka Jerman-Blažič et al. 'An economic modelling approach to information security risk management'. In: *International Journal of Information Management* 28.5 (2008), pp. 413–422.

[33] Michael Kan. *CIA used Swiss firm to spy on allies foes via hacked encryption*. [Online; accessed 14. Feb. 2020]. Feb. 2020. URL: https://uk.pcmag.com/cpus-components/124845/report-cia-used-swiss-firm-to-spy-on-allies-foes-via-hacked-encryption-tech.

[34] Miltiadis Kandias et al. 'An insider threat prediction model'. In: *International Conference on Trust, Privacy and Security in Digital Business*. Springer. 2010, pp. 26–37.

[35] John F Kihlstrom. 'The Person–Situation Interaction'. In: *The Oxford Handbook of Social Cognition*. Oxford University Press, 2013, pp. 786–805.

[36] Brian Krebs. 'FBI: Smart Meter Hacks Likely to Spread — Krebs on Security'. In: (Apr. 2012). [Online; accessed 14. Feb. 2020]. URL: https://krebsonsecurity.com/2012/04/fbi-smart-meter-hacks-likely-to-spread.

[37] Richard N Landers. 'Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS'. In: *The Winnower* 2 (2015), e143518.

[38] Aron Laszka, Mark Felegyhazi and Levente Buttyan. 'A survey of interdependent information security games'. In: *ACM Computing Surveys (CSUR)* 47.2 (2014), pp. 1–38.

[39] Benedikt Lebek et al. 'Information security awareness and behavior: a theory-based literature review'. In: *Management Research Review* (2014).

[40] Kurt Lewin. *Principles of topological psychology*. McGraw-Hill, 1936.

[41] Michele Maasberg, John Warren and Nicole L Beebe. 'The dark side of the insider: detecting the insider threat through examination of dark triad personality traits'. In: *2015 48th Hawaii International Conference on System Sciences*. IEEE. 2015, pp. 3518–3526.

[42] Jennifer Dodorico McDonald. 'Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments'. In: *Enquire* 1.1 (2008), pp. 1–19.

[43] Kenneth O McGraw and Seok P Wong. 'Forming inferences about some intraclass correlation coefficients'. In: *Psychological methods* 1.1 (1996), p. 30.

[44] Rustin D. Meyer. *Taxonomy of Situations and Their Measurement*. Nov. 2015.

[45] James A Mirrlees. 'The theory of moral hazard and unobservable behaviour: Part I'. In: *The Review of Economic Studies* 66.1 (1999), pp. 3–21.

[46] N.A. *Combating the Insider Threat*. Tech. rep. Department of Homeland Security, May 2014, p. 5. URL: https://www.us-cert.gov/sites/default/files/publications/Combating%5C%20the%5C%20Insider%5C%20Threat_0.pdf.

[47] Ngaire Naffine. 'Who are law's persons? From Cheshire cats to responsible subjects'. In: *The Modern Law Review* 66.3 (2003), pp. 346–367.

[48] Tatsuya Nogami and Jiro Takai. 'Effects of anonymity on antisocial behavior committed by individuals'. In: *Psychological Reports* 102.1 (2008), pp. 119–130.

[49] Annika Nübold et al. 'Developing a taxonomy of dark triad triggers at work–A grounded theory study protocol'. In: *Frontiers in psychology* 8 (2017), p. 293.

[50] Nick Nykodym, Robert Taylor and Julia Vilela. 'Criminal profiling and insider cyber crime'. In: *Computer Law & Security Review* 21.5 (2005), pp. 408–414.

[51] Sangchul Park. 'Why information security law has been ineffective in addressing security vulnerabilities: Evidence from California data breach notifications and relevant court and government records'. In: *International Review of Law and Economics* 58 (2019), pp. 132–145.

[52] Elisabeth Paté-Cornell and Louis Anthony Cox Jr. 'Improving risk management: from lame excuses to principled practice'. In: *Risk analysis* 34.7 (2014), pp. 1228–1239.

[53] James Pettigrew and Julie Ryan. 'Making successful security decisions: a qualitative evaluation'. In: *IEEE Security & Privacy* 10.1 (2011), pp. 60–68.

[54] Lisa Rajbhandari and Einar Snekkenes. 'Using the Conflicting Incentives Risk Analysis Method'. In: *Security and Privacy Protection in Information Processing Systems*. Ed. by Lech J. Janczewski, Henry B. Wolfe and Sujeet Shenoi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 315–329.

[55] James Reason. 'The contribution of latent human failures to the breakdown of complex systems'. In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 327.1241 (1990), pp. 475–484.

[56] Natalie M Scala et al. 'Risk and the Five Hard Problems of Cybersecurity'. In: *Risk Analysis* 39.10 (2019), pp. 2119–2126.

[57] Shalom Schwartz. *Computing Scores for the 10 Human values*. [Online; accessed 12-November-2019]. 2016. URL: https://www.europeansocialsurvey.org/docs/methodology/ESS1_human_values_scale.pdf.

[58] Shalom Schwartz. 'Value priorities and behavior: Applying a theory of intergrated value systems'. In: *The psychology of values: The Ontario symposium*. Vol. 8. 2013, pp. 119–144.

[59] Shalom H Schwartz. 'A proposal for measuring value orientations across nations'. In: *Questionnaire Package of the European Social Survey* (2003), pp. 259–290.

[60] Shalom H Schwartz. 'An overview of the Schwartz theory of basic values'. In: *Online readings in Psychology and Culture* 2.1 (2012), pp. 2307–0919.

[61] Philip Selznick. 'Foundations of the theory of organization'. In: *American sociological review* 13.1 (1948), pp. 25–35.

[62] Jordan Shropshire, Merrill Warkentin and Shwadhin Sharma. 'Personality, attitudes, and intentions: Predicting initial adoption of information security behavior'. In: *computers & security* 49 (2015), pp. 177–191.

[63] Patrick E Shrout and Joseph L Fleiss. 'Intraclass correlations: uses in assessing rater reliability'. In: *Psychological bulletin* 86.2 (1979), p. 420.

[64] Simon Singh. *The code book: the evolution of secrecy from Mary Queen of Scots to quantum cryptography*. Anchor, 2000.

[65] Einar Snekkenes. 'Position paper: Privacy risk analysis is about understanding conflicting incentives'. In: *IFIP Working Conference on Policies and Research in Identity Management*. Springer. 2013, pp. 100–103.

[66] Teodor Sommestad and Jonas Hallberg. 'A review of the theory of planned behaviour in the context of information security policy compliance'. In: *IFIP International Information Security Conference*. Springer. 2013, pp. 257–271.

[67] Seth M Spain, Peter Harms and James M LeBreton. 'The dark side of personality at work'. In: *Journal of organizational behavior* 35.S1 (2014), S41–S60.

[68] Cheryl K Stenmark and Michael D Mumford. 'Situational impacts on leader ethical decision-making'. In: *The Leadership Quarterly* 22.5 (2011), pp. 942–955.

[69] Stephen Sutton. 'Predicting and explaining intentions and behavior: How well are we doing?' In: *Journal of applied social psychology* 28.15 (1998), pp. 1317–1338.

[70] Adam Szekeres and Einar Arthur Snekkenes. 'A Taxonomy of Situations within the Context of Risk Analysis'. In: *Proceedings of the 25th Conference of Open Innovations Association FRUCT*. FRUCT Oy Helsinki, Finland. 2019, pp. 306–316.

[71] Adam Szekeres and Einar Arthur Snekkenes. 'Construction of Human Motivational Profiles by Observation for Risk Analysis'. In: *IEEE Access* 8 (2020), pp. 45096–45107.

[72] Adam Szekeres, Pankaj Shivdayal Wasnik and Einar Arthur Snekkenes. 'Using Demographic Features for the Prediction of Basic Human Values Underlying Stakeholder Motivation'. In: *Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 2: ICEIS*. SciTePress, 2019, pp. 377–389.

[73] Rahul Telang and Sunil Wattal. 'An empirical analysis of the impact of software vulnerability announcements on firm stock price'. In: *IEEE Transactions on Software engineering* 33.8 (2007), pp. 544–557.

[74] Maaike A Ten Berge and Boele De Raad. 'Taxonomies of situations from a trait psychological perspective. A review'. In: *European Journal of Personality* 13.5 (1999), pp. 337–360.

[75] Amos Tversky and Daniel Kahneman. 'Rational choice and the framing of decisions'. In: *Multiple criteria decision making and risk analysis using microcomputers*. Springer, 1989, pp. 81–126.

[76] Amos Tversky and Daniel Kahneman. 'The framing of decisions and the psychology of choice'. In: *science* 211.4481 (1981), pp. 453–458.

[77] CR Wilson VanVoorhis and Betsy L Morgan. 'Understanding power and rules of thumb for determining sample sizes'. In: *Tutorials in quantitative methods for psychology* 3.2 (2007), pp. 43–50.

[78] Bas Verplanken and Rob W Holland. 'Motivated decision making: effects of activation and self-centrality of values on choices and behavior'. In: *Journal of personality and social psychology* 82.3 (2002), p. 434.

[79] Stephen Walker. 'Workshop on stochastic error, parameter error and model error'. In: *1996 General insurance convention: 2-5 October 1996*. Springer, 1996, pp. 461–469.

[80] Xiaolong Wang and Wenli Li. 'Understanding the Incentive Mechanism of Penalty for Information Security Policy Compliance Behavior'. In: *Proceedings of the 7th International Conference on Social Science, Education and Humanities Research*. 2018, pp. 19–25.

[81] John B Watson. 'Psychology as the behaviorist views it'. In: *Psychological review* 20.2 (1913), p. 158.

[82] Eva Weishäupl, Emrah Yasasin and Guido Schryen. 'Information security investments: An exploratory multiple case study on decision-making, evaluation and learning'. In: *Computers & Security* 77 (2018), pp. 807–823.

[83] Michael E Whitman and Herbert J Mattord. *Principles of information security*. Cengage Learning, 2011.

[84] Craig S Wright and Tanveer A Zia. 'Rationally opting for the insecure alternative: Negative externalities and the selection of security controls'. In: *Computational Intelligence in Security for Information Systems*. Springer, 2011, pp. 206–213.

[85] Yu Yang, Stephen J Read and Lynn C Miller. 'The concept of situations'. In: *Social and Personality Psychology Compass* 3.6 (2009), pp. 1018–1037.

[86] Robert B Zajonc. 'Social facilitation'. In: *Science* 149.3681 (1965), pp. 269–274.

Chapter 14

Article 6: Representing decision-makers in SGAM-H: the Smart Grid Architecture Model Extended with the Human Layer

Adam Szekeres & Einar Arthur Snekkenes. Representing decision-makers in SGAM-H: the Smart Grid Architecture Model Extended with the Human Layer[1]. Accepted for publication In: *The Seventh International Workshop on Graphical Models for Security*. Springer, Cham. 2020.

Abstract

The safety and security of critical infrastructures is both a technical and a social issue. However, most risk analysis methods focus predominantly on technical aspects and ignore the impact strategic human decisions have on the behavior of systems. Furthermore, the high degree of complexity and lack of historical data for probability estimations in case of new and emerging systems seriously limit the practical utility of traditional risk analysis methods. The Conflicting Incentives Risk Analysis (CIRA) method concentrates on human decision-makers to address these problems. However, the method's applicability is restricted by the fact that humans are not represented in the Smart Grid Architecture Model (SGAM) which is the industry's most well-known model of the Smart Grid ecosystem. Therefore,

[1]This work was partially supported by the project IoTSec – Security in IoT for Smart Grids, with number 248113/O70 part of the IKTPLUSS program funded by the Norwegian Research Council.

the main objective of this paper is to establish a connection between CIRA and SGAM by proposing the SGAM-H, an enhanced version of the original architecture model complemented by the Human Layer. The development and evaluation of the artifact is guided by the Design Science Research methodology. The evaluation presents a working example of applying the CIRA method on a scenario involving intra-organizational risks at a Distribution System Operator. The key benefit of the SGAM-H is that it enables the construction of a common understanding among stakeholders about risks related to key decision-makers, which is a fundamental first step towards forming a more complete picture about potential issues affecting the electric grids of the future.

14.1 Introduction

Nation-wide electrification of industries and societies beginning in the 1880s had tremendous economical and societal benefits [7] and the demand for a stable and reliable supply of electricity has exceeded that for any other forms of energy [28]. A properly functioning power grid represents an indispensable infrastructure for modern societies, which supports all aspects of life. While demand for electricity will keep rising in the future (e.g., due to increasing electrification of the transportation sector, growing populations, etc.) international directives and regulations have been pushing toward a shift from dependency on fossil and nuclear power sources to more eco-friendly and sustainable renewables. Most renewable power sources (e.g., wind, solar) are intermittent in nature which requires a paradigm shift from centralized large-scale generation models to flexible, distributed and small-scale solutions [11]. At the same time economic constraints make the complete reconstruction of the power grid highly unfeasible. The envisaged solution is encompassed in the concept of the Smart Grid (SG), which aims at solving the challenges of the future by relying on the physical infrastructure of the past with enhancements from novel information and communication technologies. Thus the SG represents a highly complex system with real-time sensing and control capabilities using a bidirectional flow of electricity and information, enabled by the addition of internet of things (IoT) devices at various parts of the grid. Several stakeholders are involved in SG-related activities including: legislators, governmental agencies, standardizing bodies, data protection authorities, organizations focusing on the generation, transmission, distribution of electricity, equipment manufacturers, software and security providers, researchers and consumers [8].

Developments in SGs are driven by a combination of political, economic and ecological motives. Misaligned incentives are unavoidable when the number of interacting stakeholders is considered in a system of such complexity (both technically and socially). Misaligned incentives are particularly prevalent in information

systems where those who are responsible for providing security are not the same people who benefit from the protection or suffer when things go wrong. For example, increasing the dependency of critical infrastructures on public information systems (network convergence) can be an efficient short-term cost saving strategy for utility companies, but it increases society's long-term vulnerability, which will ultimately bear the costs [24]. It has been demonstrated that misaligned incentives, negative externalities and moral hazard arise in a variety of settings within the field of information security [1]. The identification and mitigation of such problems is crucial for ensuring the safety and security of societies depending on SGs and other critical infrastructures.

Conflicting Incentives Risk Analysis (CIRA)

The Conflicting Incentives Risk Analysis (CIRA) method focuses on the motivation of individual stakeholders to define risks. The lack of relevant historical data in case of emerging and dynamic systems creates a significant challenge for traditional (i.e., relying on frequentist probability estimations) risk analysis methods [37]. Furthermore, deliberate human actions due to misalignment of incentives is rarely at the center of risk analysis procedures. CIRA defines risk as the misalignment between stakeholder incentives. The analysis focuses on the *Risk owner's* (i.e., person at risk) exposure to the actions or inactions of several other stakeholders (*Strategy owners*) who are in the position to choose courses of actions [32]. CIRA combines quantitative methods to characterize risks attributed to key decision-makers, therefore, aims at overcoming some of the problems associated with qualitative risk scoring methods [15].

Smart Grid Architecture Model (SGAM)

The creation of the Smart Grid Architecture Model (SGAM) was motivated by the need to represent stakeholders, applications and systems that will have to achieve efficient interdependent operations in future SGs. To ensure these goals, developers and standardization bodies of the SG need to have a common understanding or shared model about the systems which will be implemented. To capture the EU-specific requirements the SGAM was designed to tackle the complexity by representing systems in a consistent and comprehensive way. It enables standards gap analysis; visualization and assessment of use cases in a technology-neutral way; comparison of different approaches and road-maps from various viewpoints. Figure 14.1 presents the SGAM, based on [4]. *Domains* represent the energy conversation chain from generation site to customer premises. *Zones* capture the power system management supported by ICT from the level of processes to markets. *Interoperability layers* represent different levels of abstraction from the physical hardware to business perspectives highlighting the interconnectedness and

dependencies between entities.

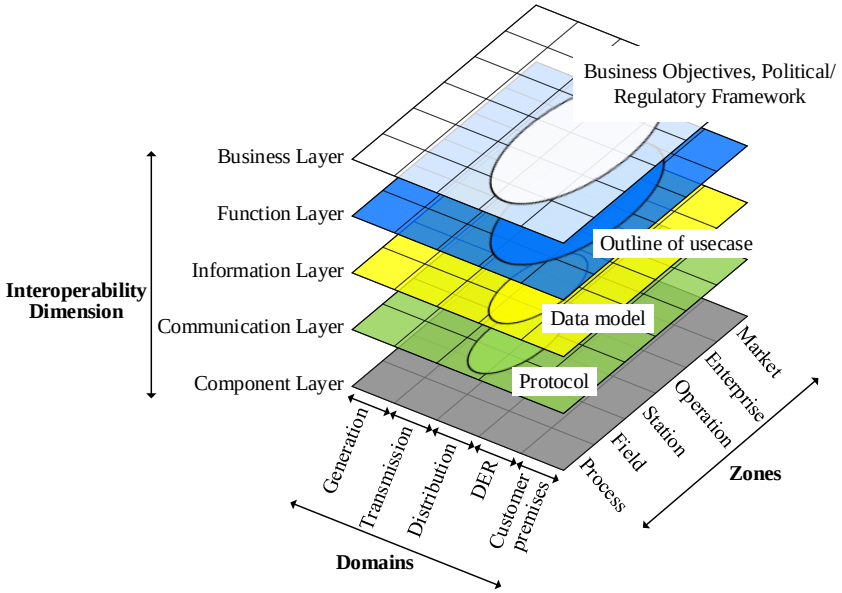


Figure 14.1: The Smart Grid Architecture Model (SGAM) based on [4].

How is it possible to analyse risks arising from human decision-making in a complex system as the SG? Several management failures (management of tree growth, lack of vulnerability and system-health assessment, etc.) contributed to the 2003 Northeast blackout in the US, affecting 55 million people with an estimated economic impact of $6 billion [25]. Organizations responsible for the development and maintenance of the grid need to have the right incentives in place to achieve their goals at a socially optimal level. Are measures in place to protect the privacy of customers despite increased monitoring capabilities enabled by smart meters and other smart home devices [22]? Does information security contribute to the organizational goals or is it perceived as a impediment to smooth operations [48]? Can the SG fulfill the hopes by providing electricity in a safe, reliable and secure way without significantly increasing society's exposure to new threats [19]?

Problem statement and motivation

In order to enable the application of the CIRA method on SG use cases, a connection between the models has to be established. Human decision-makers are not represented in the existing SGAM, which may result in ignoring the impact strategic human decisions have on the grid. The SGAM documentation briefly mentions

human-aspects: *"The concept of an Actor is very general and can cover People (their roles or jobs), systems, databases, organizations, and devices"* [4]. However some critical distinguishing features justify separating human decision-makers from the Actor concept. Human decision-makers:

- are self-determined (i.e., choosing their own goals [10]);
- have unique motivations, which may not be in alignment with organizational/societal objectives (e.g., principal-agent models [47]);
- are in the unique position to control all other objects (e.g., regulations, business goals, components, etc.) within a system.

Ergo, human decision-makers have distinctive and significant impact on every aspect of the system's behavior which requires the explicit integration of human decision-makers into a reference architecture to provide a more comprehensive model. Furthermore, it is necessary to investigate the CIRA method's adequacy for analysing risks in highly complex emerging systems, where the application of traditional risk analysis methods may be infeasible (due to lack of historical data for probability estimations and unmanageable complexity of information systems).

This paper presents an approach for addressing these gaps in the literature. The paper is structured as follows: Section 14.2 provides an overview about modifications to the basic SGAM as well as approaches for modeling humans from a broad range of domains. Section 14.3 describes the Design Science Research Methodology (DSRM) which guided the development and evaluation of the paper's artifact. The artifact is presented and evaluated by a case study throughout Section 14.4. Section 14.5 discusses key findings and Section 14.6 draws conclusions. The paper ends with ideas for further work in Section 14.7.

14.2 Related work

This section is divided into two parts. The first part reviews research work which proposes or implements extensions to the generic SGAM to solve specific tasks. A literature search using the search string ("sgam" extend OR extension) appearing anywhere in the articles was conducted on Google Scholar and articles citing the original publication were screened; other relevant articles were identified among references. Studies describing the application of SGAM were excluded. The second part presents approaches for modeling human behavior across various domains to illustrate design decisions about the models.

Variants of SGAM

The Information System Architecture for e-Mobility (EM-ISA) is an early SGAM variant focusing on electric vehicle (EV) integration into the grid. The model significantly reduces the number of the domains and zones, then proposes the integration of human-machine interfaces into the model to capture interactions between humans (operators) and objects without further specifying human attributes [35]. The Electric Mobility Architecture Model (EMAM) focuses on EV integration as well. In EMAM, the Generation domain is removed and an electric mobility domain is added to the grid plane, while keeping the rest of the original model unchanged. Recognizing the utility of the SGAM for standardisation purposes, two other reference models were developed following similar architecture engineering principles. While the layers of The Smart City Infrastructure Architecture Model (SCIAM) and the Smart Home Architecture Model (SHAM) are the same as those of SGAM, different domains and zones are introduced which may decrease compatibility between models [45]. SGs may differ between countries, therefore it is important to increase compatibility between various implementations. Two state-of-the-art models (the SGAM from EU and the NISTIR 7628 from U.S.) are combined in order to facilitate security analysis from the beginning of the development process [44]. In addition to the previously described variants two more architecture models are described in [43]. The Home and Building Architecture Model (HBAM) utilizes SGAM's layered approach with different zones and domains introduced to capture relevant concepts within scope of smart homes and buildings. The Reference Architecture Model for Industry 4.0 (RAMI 4.0) is regarded as the most sophisticated derivative of the SGAM containing zones and domains relevant for industrial applications and extending the interoperabilty perspectives with an additional layer. Two more reference models have been developed using the SGAM's design principles. The Reference Architecture Model Automotive (RAMA) represents the life-cycle of connected vehicles and the related information technologies and the Maritime Architecture Framework (MAF) models information exchange between various actors in the maritime domain [46].

Approaches for modeling humans

Models in general, are abstract representations of a complex entity or phenomenon capturing its most significant aspects for a pre-specified purpose. Analogies, shared features and other similarities between entities play a key role in modelling activities. For example, pigs and other animals can represent humans in medical experiments due to the high number of shared features (in terms of genetics, physiology and anatomy, etc.) [23]. Investigations in road safety require human models which accurately capture the physical properties of real humans in car crash scenarios [2]. Personas or user archetypes are widely used human models in the software engineer-

ing industry. Personas guide the development process by representing future users and their goals in relation to the product [5]. Realism of human models is becoming increasingly important in virtual environments where representations can replace real humans (in communication context [3]) or simulated agents are required to act realistically (in training context [27]). For behavior prediction, a human model must incorporate psychological constructs that are most likely to govern or influence (i.e., mediate and moderate) the behavior of interest. Models reduce real-world complexity, which enables that only a small set of well-defined parameters are required for predictions. The importance of appropriately modeling humans and human behavior has been recognized in a variety of domains. Human performance and mental load models have been developed to represent operator characteristics and to assist the design of human-machine interfaces in the context of industrial control systems [38]. A variety of human behaviors are of interest to the military, therefore a wide range of human models have been developed (at the individual and group level) to support agent-based behavioral simulations [30]. A key challenge is to find the right balance between the model's complexity and its realism [16]. In the context of information security, humans can be represented by a utility function which is the most suitable level of abstraction for game theoretic simulations [20]. People have great impact on the Earth's overall condition, but humans are not yet explicitly represented in Earth system models used for simulating ecological dynamics. The selection of an appropriate human model relies on the modeler's understanding about the strengths and weaknesses of each model [26].

Summary of related work

The reviewed literature demonstrates the SGAM's acceptance among practitioners and researchers and presents several domain- or task-specific variants inspired by the original model. However, the representation of human decision-makers is lacking, which impedes the efficient application of CIRA on SG scenarios. The broad overview on the literature of human modeling approaches highlights that models should be developed according to relevant design considerations (e.g., specifying the model's content in relation to the behavior of interest, complexity-realism trade off, etc.).

14.3 Methodology

This study is based on the design science research (DSR) paradigm, which provides an organizing framework for the development of purposeful artifacts to solve a specific problem [14]. The DSR methodology defines three cycles which interact with each other during task execution [13]. The *design cycle* represents the core activities (development and evaluation of the artifact in an iterative process) which is embedded in a broader context. The design cycle receives input from two

sources. The *relevance cycle* refers to the interaction between the environment (where problems and needs for a new solution arise) and the design cycle (produces solutions). Artifacts from the design cycle are fed back to the environment through the relevance cycle and the artifacts are applied in the context where they were intended to function. Interaction of the design cycle with the supporting knowledge-base defines the *rigor cycle* which provides the necessary tools, methodologies, theories for the development and evaluation of the artifact. Information flows in both directions between the rigor and design cycles as well, thus new knowledge and experience resulting from the construction of the artifact are recorded in the knowledge-base using the most suitable format (presentation, tutorial, academic paper, etc.).

The relevance cycle serves as a starting point for any DSR activity by specifying the context and problems in the domain (i.e., requirements), that the artifact should solve. Furthermore, it defines evaluation criteria for testing the artifact's utility within the environment. The need to represent human stakeholders within the SG has been arising from interactions with other stakeholders (students, conference and project participants). Difficulty of creating a common understanding among stakeholders about CIRA's applicability and relevance was identified as a major barrier to the method's acceptance and adoption. Thus, a more efficient method of conveying meaning was set as a requirement. The second step focuses on the identification of suitable theories, frameworks to meet requirements. Therefore, the rigor cycle was used for the identification of existing frameworks by reviewing the relevant literature, which resulted in identifying the SGAM as an ideal candidate requiring customization. The development activity within the design cycle was used to extract key concepts from CIRA and to create visual representations of its abstract concepts. An important design consideration was to keep a high degree of compatibly with the original SGAM version, therefore an extension is proposed: the SGAM-H enhanced by a Human Layer and its necessary components. The artifact model was built from scratch in Microsoft Visio, to ensure re-usability and mutability (the Visio-based templates reported in [34] were not available online). The final step within the design cycle is the evaluation of the artifact which is achieved through a hypothetical case study (qualitative, descriptive method) demonstrating how key CIRA concepts are mapped onto the Human Layer and how it conveys meaning. The artifact is evaluated in terms of its efficacy, ease of use, completeness and homomorphism (i.e., correspondence with another model) [31].

14.4 Human Layer

This section presents the Human Layer as an extension of the SGAM, giving rise to the SGAM-H. The Human Layer's basic elements for constructing and

representing the context of risk analysis are introduced. Next, the artifact's efficacy is demonstrated on a hypothetical case study which applies the CIRA method on a SG scenario focusing on risks experienced by the CEO of a Distribution System Operator (DSO). Several aspects of the case study were inspired by media reports [36] and analyses of real-world incidents [25] accompanied by relevant scientific literature [6] in order to increase its realism. Finally, the artifact is evaluated along the previously identified criteria.

Figure 14.2 presents the Human Layer placed on top of the business layer of the original SGAM. This implementation enables the representation of human stakeholders with their relevant attributes on the architecture model and emphasizes the critical role that strategic human decisions can have on various aspects of SGs.

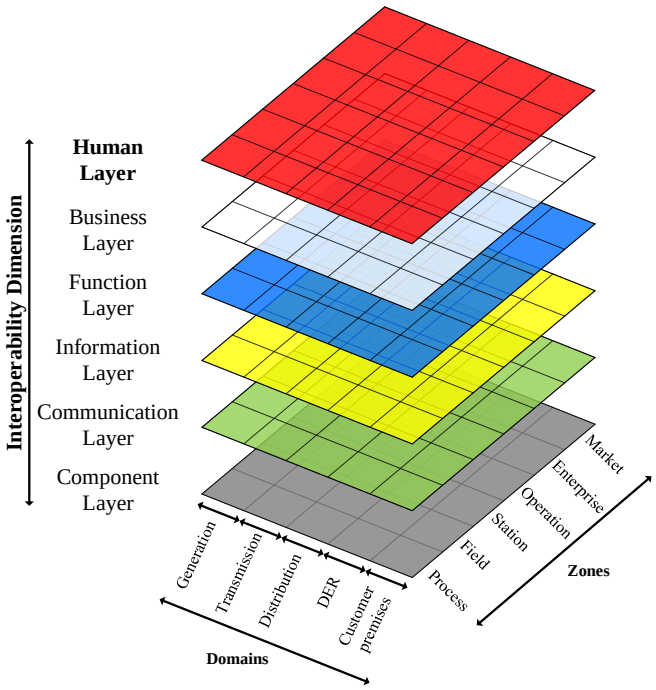


Figure 14.2: SGAM-H including the Human Layer.

Figure 14.3 presents the stakeholder models; components to represent human attributes and other elements of the layer to capture key concepts of CIRA. Two types of stakeholder classes are distinguished by color and related captions: human models in blue represent the risk owner, human models in white represent the class of strategy owners. Post-analysis states are distinguished by a tag above the models to display the risks explicitly (i.e., consequences for the risk owner, incentives

for the strategy owner). The sign (+/-) represents the direction of utility change following strategy execution. Furthermore, incentives are marked with red fill color on the strategy owner figures. The height of the red coloring from the bottom of the figure matches with the magnitude of the incentive (i.e., an incentive of 50 produces a red fill color up to 50% of the figure's height). Strategy owners' profile information is captured in brackets, to record the information used for the construction of motivational profiles before the analysis. Stakeholders are linked to other entities (e.g., physical hardware, organizations, etc.) by dashed lines. Strategies are represented by continuous lines ending in an arrow, directed from the strategy owner to the risk owner.

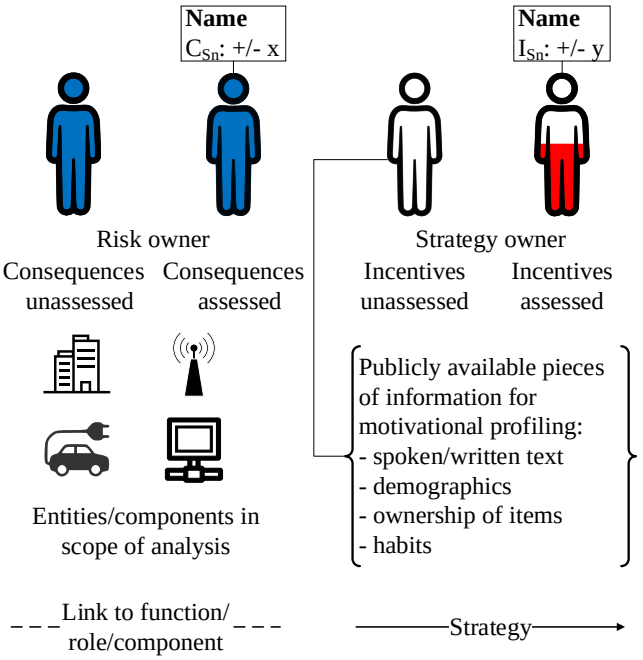


Figure 14.3: Components of the Human Layer.

Case study: DSO risks

This sub-section demonstrates the use of the SGAM-H through a case study in which the CIRA method is applied to a scenario focusing on the risks faced by the organizational leader of a DSO, since the organization has a critical role in the SG ecosystem. Numbering of the subsequent paragraphs follows the steps of the CIRA procedure based on [32].

1. Identification of the risk owner

The risk owner is the CEO of a DSO, who is interested in intra-organizational risks which may interfere with the objectives of the organization.

2. Identification of the risk owner's key utility factors

The key utility factors (UFs) were identified by relying on the Balanced Scorecard (BSC) method, which was designed to aid managers in evaluating and measuring organizational performance through a set of measures linked to organizational objectives [18]. Four perspectives are distinguished by the BSC method: *Financial*, *Customers and stakeholders*, *Learning and growth* and *Internal business processes*. The method enables the development of key performance indicators at various levels (departments, individuals) to achieve better organizational performance. Since utility companies such as DSOs operate as natural monopolies due to high infrastructural costs, their operations differ from purely for-profit organizations. In the not-for-profit sector, the financial perspective is often seen as a constraint rather than an objective, which requires different priorities [21]. Some work has been done to adapt the BSC to the specific needs of utility companies [17, 33]. Table 14.1 presents the risk owner's key utility factors derived from the BSC perspectives.

Table 14.1: Key utility factors of the CEO.

BSC perspectives	**Utility factors**
Financial	Revenue
Customers and stakeholders	Customer privacy
	Contribution to public welfare
Learning and growth	Innovation
Internal business processes	Relationship with regulators

3-5. Identification of strategies that may influence the risk owner's utility factors; Identification of roles and named strategy owners which can execute the strategies

Steps 3-5. of the procedure are summarized in Table 14.2. For each utility factor an appropriate strategy was identified by considering key processes and functions at a DSO. The identification of roles and strategy owners is aided by the organizational chart which allocates the responsibilities and tasks to various roles occupied by actual persons. The scenario description for each person illustrates motivational factors at play regarding the dilemmas they face in a given situation.

Sigurd works as a dispatcher at the organization. He is approached by his best friend who suspects that his wife is cheating on him and asks Sigurd to monitor the detailed electricity consumption of their holiday house which he thinks is used as a hideout by her. He has access to the relevant data, and thinks he can fulfil

Table 14.2: The risk owners' utility factors (UFs); strategies that impact the risk owner's utility factors; roles and individuals.

Affected UFs	Strategy	Role	Person
Customer privacy	Help a friend (S_1)	Dispatcher	Sigurd
Contribution to public welfare	Fix street lights (S_2)	Operations manager	Emma
Innovation	Recruit research applicants (S_3)	Head of R&D	Hanne
Relationship with regulators	Support system integration (S_4)	CISO	Henry

the request without getting into trouble. The legal and financial implications of a privacy breach are of key interest to the risk owner. Emma is responsible for distributing tasks efficiently within her team of technicians working in the field. Citizens are complaining about faulty street lights and dangerously dark streets. She has to decide how to allocate tasks within the team based on existing efficiency measures in place. Hanne works at the R&D department developing new services for customers. Students with novel ideas apply to get work experience at the organization, but she perceives recruitment and training of students as a nuisance since student projects rarely get converted into successful products. She has to decide whether increasing the number of student projects (to fulfill an important societal role) worth lowering her performance indicators. Henry believes that the new agenda to harmonize all data acquisition systems at the organization would create a singularity threat and he believes in security through diversity. He has the final word regarding the new system's implementation in the project.

6. Identification of the strategy owners' utility factors

For each strategy owner two types of utility factors are distinguished. Work-related factors are derived from the BSC method's perspectives. Personal utility factors are represented by basic human values [39]. Table 14.3 presents the key utility factors for each strategy owner.

7. Operationalization of utility factors

To operationalize the utility factors, existing work on DSO-specific KPIs was surveyed [6, 12] as well as relevant regulations (GDPR [9]). KILE (quality-adjusted revenue frames for energy not delivered) represents customers' costs for interruptions, and is a form of revenue reduction due to interruptions, which aims at incentivizing utility companies to maintain operational reliability [29]. Utility factors capturing personal motivations were operationalized in previous work as

Table 14.3: Work-related and personal utility factors for each strategy owner.

Strategy owner	Utility factors						
	Work-related (associated with role)	Personal					
Sigurd	Percentage of successfully located faults and dispatched repair teams within time frame (%)	ST	OC	CO	HE	SE
Emma	Percentage of reconnected electricity customers within time frame (%)					
Hanne	New services ready for market (%)					
Henry	Percentage of resolved cyber-incidents within a time frame (%)					
Note. ST: self-transcendence, OC: openness to change, CO: conservation, HE: hedonism, SE: self-enhancement.							

publicly observable pieces of information, for the construction of motivational profiles [40, 41, 39]. Table 14.4 presents how each utility factor is operationalized.

Table 14.4: Utility factors operationalized.

Role	Type of utility factor	Utility factor	Operationalized as
Risk owner	Professional	Revenue	R = Revenue cap - KILE (CENS) [29]
		Customer's data privacy (%)	CDP = 1 - (privacy-related penalties/privacy breach cap (0.04*annual turnover)) [9]
		Contribution to public welfare (%)	PW = resolved public complaints within 1 month / all complaints in a period
		Innovation (%)	INN = number of established research collaborations with universities / number of applications from students
		Relationship with regulators (%)	REG = number of reports accepted without modification / all reports submitted
Strategy owner		Percentage of successfully located faults and dispatched repair teams within time frame (%)	TDISP = number of successful responses within 30 mins / all trouble calls received
		Percentage of reconnected electricity customers within time frame (%)	TREST = number of successfully reconnected customers within 24 hours / number of customers assigned without electricity supply
		New services ready for market (%)	MARK = new market ready-services / all R&D projects initiated
		Percentage of resolved cyber- incidents within time frame (%)	CYINC = successfully mitigated cyber-incidents within 12 hours / all reported
	Personal	Self-transcendence	Publicly available pieces of information for psychological profiling: text analysis [40], demographic features [41], item ownership and habits [39].
		Openness to change	
		Conservation	
		Hedonism	
		Self-enhancement	

8. Weighing of utility factors

Table 14.5 presents each utility factor's contribution to the person's overall utility. For the purpose of demonstration, the CEO's overall utility is entirely composed of work-related utility factors. Employees on the other hand, derive utility from other factors which are not directly linked to their professional role (i.e., human values). Work-life balance is represented by the global ratio between work-related and personal utility factors. Weights (w) of the personal utility factors capture the relative importance of basic human values for the subject. Thus, weights are inferred from psychological profiles based on various publicly available pieces of information (e.g., demographics [41], texts produced by the subject [40], evidence of past choices reflecting value trade-offs, habits [39]). Various metrics have been used for quantifying the accuracy/uncertainty of the inferred profiles: R^2 - coefficient of determination (range: 0.19-0.39), PI - prediction interval (Mean: 0.077, SD: 0.794), Pearson correlation coefficients between predicted and ground-truth scores (range: 0.34-0.52) [39]. All the weights sum to 1 for each stakeholder.

Table 14.5: Weighing of utility factors.

CEO	w	Sigurd	w	Emma	w	Hanne	w	Henry	w
Revenue	0.300	Percentage of successfully located faults and dispatched repair teams within time frame (%)	0.25	Percentage of reconnected electricity customers within time frame (%)	0.30	New services ready for market (%)	0.35	Percentage of resolved cyber-incidents within time frame (%)	0.40
Customer's data privacy (%)	0.175	Self-transcendence	0.18	Self-transcendence	0.12	Self-transcendence	0.10	Self-transcendence	0.11
Contribution to public welfare (%)	0.175	Openness to change	0.14	Openness to change	0.20	Openness to change	0.20	Openness to change	0.10
Innovation (%)	0.175	Conservation	0.17	Conservation	0.09	Conservation	0.05	Conservation	0.18
Relationship with regulators (%)	0.175	Hedonism	0.16	Hedonism	0.12	Hedonism	0.16	Hedonism	0.06
		Self-enhancement	0.10	Self-enhancement	0.17	Self-enhancement	0.14	Self-enhancement	0.15

9. Determination of each strategy's impact on the utility factors

Each strategy owner's decision-making process is modeled in Table 14.6 with the decisions' impact on the risk owner's utility factors. For simplicity each strategy's influence is limited to a maximum of two utility factors. Real-world choices are

determined by the complex trade-offs between utility factors as perceived by the stakeholders in a choice situation (i.e., dilemma). Personal features (represented by the weights of each utility factor) interact with salient features of the immediate situation (i.e., initial and final values- capturing states as opposed to traits). Decisions are motivated/demotivated by the overall gains/losses expected from the execution of a strategy. The decision-making process is modeled as $C = f(P \times S)$, where C is a choice, P refers to personal features and S captures situational features. The formula may include the accuracies with which an analyst can assess the relevant person-situation interactions. The results of the context establishment are depicted on the SGAM-H in Figure 14.4.

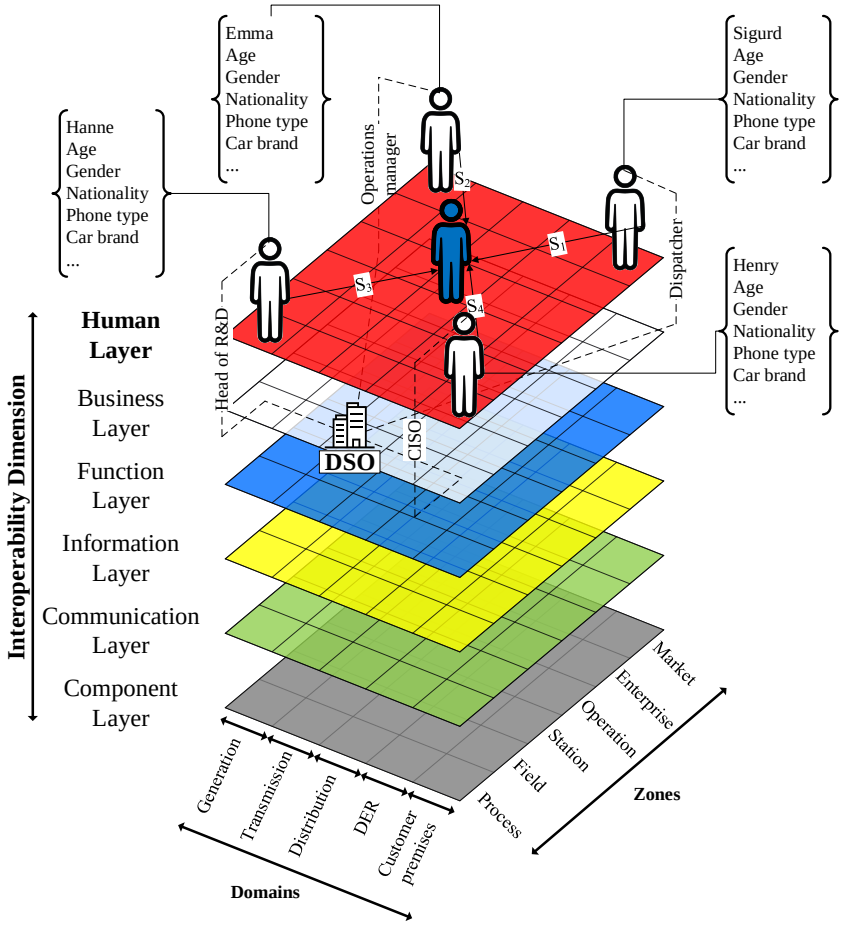


Figure 14.4: Summary of context establishment on the SGAM-H.

Table 14.6: Impact of the strategies on utility factors.

				Final values after strategy execution			
				A	B	C	D
	Utility factors	**Weights**	**Initial Value**	Help a friend (S_1)	Fix street lights (S_2)	Recruit research applicants (S_3)	Support system integration (S_4)
CEO	Revenue	0.3	50	50	48	53	55
	Customer's data privacy (%)	0.175	50	15	50	50	50
	Contribution to public welfare (%)	0.175	50	50	60	50	50
	Innovation (%)	0.175	50	50	50	65	50
	Relationship with regulators (%)	0.175	50	50	50	50	90
Sigurd	Percentage of successfully located faults and dispatched repair teams within time frame (%)	0.25	50	50			
	Self-transcendence	0.18	20	90			
	Openness to change	0.14	50	50			
	Conservation	0.17	50	50			
	Hedonism	0.16	50	50			
	Self-enhancement	0.1	50	50			
Emma	Percentage of reconnected customers within time frame (%)	0.3	90		30		
	Self-transcendence	0.12	50		50		
	Openness to change	0.2	50		50		
	Conservation	0.09	50		50		
	Hedonism	0.12	50		50		
	Self-enhancement	0.17	50		50		
Hanne	New services ready for market (%)	0.35	50			10	
	Self-transcendence	0.1	50			50	
	Openness to change	0.2	50			50	
	Conservation	0.05	50			50	
	Hedonism	0.16	50			20	
	Self-enhancement	0.14	50			50	
Henry	Percentage of resolved cyber-incidents within time frame (%)	0.4	60				30
	Self-transcendence	0.11	50				50
	Openness to change	0.1	50				50
	Conservation	0.18	50				40
	Hedonism	0.06	50				50
	Self-enhancement	0.15	50				50

10. Utility estimation

Each stakeholder's overall utility is calculated in Table 14.7 before and after strategy execution. The weighted sum of each utility factor produces the overall utilities according to the Multi Attribute Utility Theory used in CIRA [32].

Table 14.7: Utility estimation.

	Utility				
	Initial	**Final**			
Stakeholders		Help a friend (S_1)	Fix street lights (S_2)	Recruit research applicants (S_3)	Support system integration (S_4)
CEO	50	43.875	51.15	53.525	58.5
Sigurd	44.6	57.2			
Emma	62		44		
Hanne	50			31.2	
Henry	54				40.2

11. Calculation of incentives

Differences in terms of the overall utilities before and after strategy execution are presented in Table 14.8. Stakeholders prefer options that increase their utility to options that decrease it, therefore options with positive contribution are selected, whereas options which provide disutility are avoided.

Table 14.8: Change in utilities.

Stakeholders	**Change in utilities (incentives)**			
	Help a friend (S_1)	Fix street lights (S_2)	Recruit research applicants (S_3)	Support system integration (S_4)
CEO	-6.125	1.15	3.525	8.5
Sigurd	12.6			
Emma		-18		
Hanne			-18.8	
Henry				-13.8

12. Determination of risks

Risks are expressed and presented to the CEO as incentive-consequence (I-C) pairs in Table 14.9. Incentives represent the strength of motivation for each strategy owner to select/avoid the related option, consequences capture the risk to the risk owner. Risks that are characterized by a positive incentive and a negative consequence are threat risks. Negative incentive and positive consequence pairs represent opportunity risks, which would be desirable for the risk owner but the strategy owner would have to take a loss to provide the benefit. The assessed risks are shown on the Human Layer in Figure 14.5.

Table 14.9: Risks experienced by the CEO.

Strategy	Incentive	Consequence
Help a friend (S_1)	12.6	-6.125
Fix street lights (S_2)	-18	1.15
Recruit research applicants (S_3)	-18.8	3.525
Support system integration (S_4)	-13.8	8.5

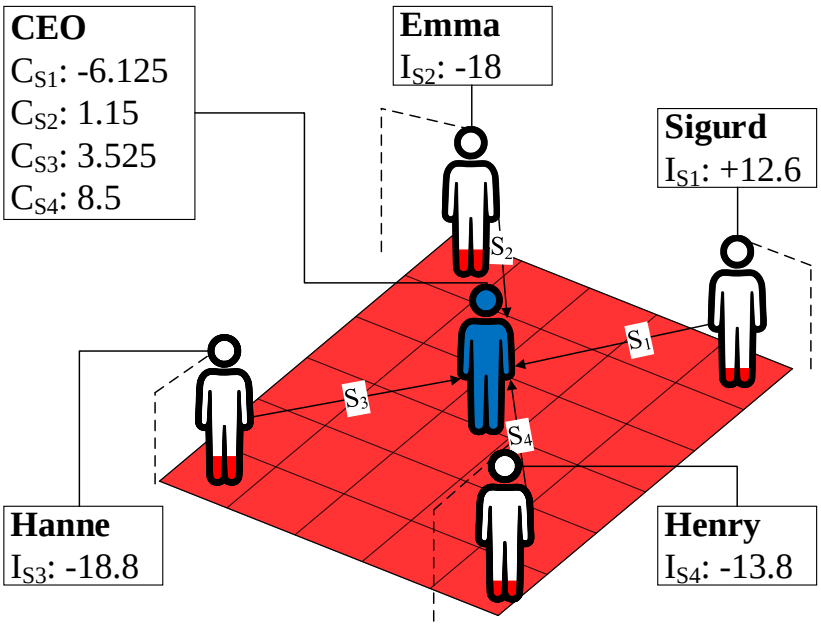


Figure 14.5: Risk representation on the Human Layer.

13. Risk evaluation

The CEO has to subjectively evaluate whether the risks are above or below the acceptability threshold. Risk that are below the acceptance level may not require further action and may only be monitored (e.g., fixing the street lights, recruit students). Risks that are above the threshold require risk treatment. It should be noted that this demonstration relies on crisp numbers, which do not capture

appropriately the accuracies/uncertainties associated with each measurement along the chain of inference. Thus, to draw a more accurate picture for real-world applications it is important to understand how errors propagate. According to [42] the error in a quantity which is derived from other quantities (each measured with some uncertainty) is calculated as:

$$\begin{aligned}(\text{Measured value of})\, x &= x_{\text{best}} \pm \delta x,\\ x_{\text{best}} &= \text{best estimate for}\, x,\\ \delta x &= \text{uncertainty or error in measurement},\\ \frac{\delta x}{x_{\text{best}}} &= \text{fractional uncertainty}.\end{aligned}$$

Since C (choice) is calculated as the product of P and S, the relative error of C can be calculated as the sum of fractional uncertainties in quadrature assuming independent random errors as follows:

$$\frac{\delta C}{C} = \sqrt{\left(\frac{\delta P}{P}\right)^2 + \left(\frac{\delta S}{S}\right)^2}$$

The resulting relative error can be converted into absolute error, and used to compute $C \pm \delta C$ which more accurately captures it's uncertainty.

14. Risk treatment

Strategy 1 and 4, are above the risk acceptance threshold, therefore certain incentive modifications are necessary to make the options more (for opportunity risks) or less (for threat risks) desirable for the strategy owners. A risk mitigation for S_1 would be to increase personal accountability in case of privacy violations to make the option less desirable for the strategy owner. Mitigation of S_4 involves the adjustment of the relevant KPI which focuses exclusively on cyber-incident response times by the inclusion of a cross-departmental rating system linked to bonuses which measures cooperation between departments. This can provide incentives to seek mutually beneficial outcomes. The need for alignment between departments requires novel metrics both at the micro and macro levels within the organization.

Evaluation of the Human Layer

The artifact is qualitatively evaluated across the following criteria by its developers (i.e., internal evaluation by two people): efficacy, ease of use, completeness and homomorphism adhering to the definitions in [31]. A five point grading scale (5-excellent, 4-good, 3-satisfactory, 2-sufficient, 1-unsatisfactory) is used for describing the extent to which the artifact fulfills the evaluation criteria. Efficacy is rated 5 since it successfully establishes a connection between SGAM and CIRA

by representing human stakeholder models, thus addressing the identified gap in the literature. Ease of use is rated 3, since the development and construction of the models from scratch required significant effort initially in terms of time spent (several days). After the basic models have been established and with subsequent reuse of the artifacts (i.e., iterative adjustments and updates applied to the models as the case study was developing which involved the identification of relevant literature, extraction of key concepts and customization of the metrics, etc.) it was possible to reduce the effort significantly (below 1 hour for each iteration). Completeness is rated 5 since it captures all the relevant elements and relationships between elements identified in CIRA. Homomorphism refers to the correspondence with a reference model (i.e., original SGAM) and is rated 4 since the extension does not interfere with the original model's structure but further adjustments may be necessary to ensure full, unambiguous compatibility with SGAM objects.

14.5 Discussion

Critical infrastructures designed and built in the previous century are becoming more autonomous and interconnected by the inclusion of IoT devices. Modernisation is driven by a variety of economical, political and ecological motives. Increasing dependency on ICT gives rise to previously unimaginable risks which may endanger the safety, security and privacy of societies at scale. High levels of complexity and lack of historical data about system behavior represent great practical impediments for traditional risk analysis methods. The CIRA method proposes a solution to these problems by focusing on the behavior of fundamental components of any modern system: key decision-makers. Human decision-makers are not appropriately represented on the most well-established model of the SG (SGAM) which may lead to under-recognition of people's influence on the SG. Consequently, risk analyses may exclusively focus on technical aspects and miss the point, that technology is under the control of human decision-makers with unique motivations. In order to address this imbalance between perspectives, and to enable the creation of a common understanding about the human aspects, this paper proposed the SGAM-H with the Human Layer on top of the SGAM interoperability layers. The extension aimed at keeping compatibility with the original model to a maximum to increase chances of adoption. The extension's efficacy was demonstrated through a case study which applies the CIRA method to a DSO scenario. The case study was inspired by real-world incidents and presented the application of metrics developed for real-world organizations to ensure its realism. The case study presented one threat risk and three opportunity risks to demonstrate the method's applicability. Since the concept of threat risk is more similar to the traditional concept of risk (i.e., an event with negative consequences), the demonstration served the purpose of providing more details about the concept of opportunity risk which has received

relatively less attention previously. The artifact has been evaluated along several criteria, thus completing an iteration within the DSR methodology's design cycle. The evaluation has also uncovered some limitations: lack of formal integration of the decision-maker models (and attributes) into existing SGAM models using the Unified Modeling Language (UML); the case study used for demonstration is hypothetical, since access to real-world organizations is limited; the internal, qualitative evaluation represents a weak form of evaluation.

14.6 Conclusions

The key contributions of this work are as follows: proposal of the SGAM-H augmenting the original SGAM with the Human Layer to create a common understanding among stakeholders operating in the SG ecosystem about the importance of focusing on human-related risks, and to improve risk communication when the CIRA method is applied to SG scenarios. Furthermore, the study contributes by presenting a fully worked-out example of CIRA's application, which may help students and practitioners in better understanding the method's procedures. Recent developments regarding CIRA have been incorporated into the case study (e.g., use of BSC method, operationalization of motivational profiles, differentiation between various aspects of utility, propagation of errors, risk treatment options) and the artifact is evaluated to identify its strengths and weaknesses.

14.7 Further work

This study focuses on intra-organizational risks where the CEO is assumed to have the capability to mitigate the identified risks. However, the connection with the other SGAM-layers ensures that relevant stakeholders can be identified from any layer. Stakeholders from other organizations could be identified and elevated from the business layer to analyze inter-organizational risks. Owners of information or physical assets could be identified and elevated to the Human Layer, where the existing connections between assets are inherited by the stakeholders, enabling the identification and specification of strategies that are at the disposal of the strategy owners. This procedure could be a significant step towards replacing the analyst's intuition for strategy identification (step 3). Development of new tools would be required to increase the usability of the Human Layer (e.g., inclusion of interactive functionality would improve user-experience and risk communication capabilities). Furthermore, scalability could be improved by additional software support to enable the representation of more stakeholders on the Human Layer. Simulation-based analyses could be conducted by a more completely populated SGAM model in which the effects of strategic decisions could propagate through the system to simulate and analyze the reactions of other entities (e.g., customers,

competitors). Finally, the evaluation can be improved by using more rigorous quantitative evaluation methods, independent of the developers of the artifact (external evaluation). Field experiments with practitioners, or students require the creation of training materials, while application to real-world cases and expert evaluations can be useful to assess user acceptance. It should be investigated how the general idea of a Human Layer can be applied to other domains (e.g., e-health, transportation domains, etc.) to improve understanding about deliberate human behavior and information security risks.

Acknowledgements

We would like to thank the four anonymous reviewers whose comments helped to improve the quality of the paper.

References

[1] Ross Anderson and Tyler Moore. 'Information security: where computer science, economics and psychology meet'. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 367.1898 (2009), pp. 2717–2727.

[2] Michel Behr et al. 'A human model for road safety: from geometrical acquisition to model validation with radioss'. In: *Computer Methods in Biomechanics & Biomedical Engineering* 6.4 (2003), pp. 263–273.

[3] Tolga K Capin et al. 'Virtual human representation and communication in VLNet'. In: *IEEE Computer Graphics and Applications* 17.2 (1997), pp. 42–53.

[4] CEN-CENELEC-ETSI Smart Grid Coordination Group. *Smart Grid Reference Architecture*. 2012.

[5] Alan Cooper. *The inmates are running the asylum*. Macmillan, 1996.

[6] Ignacio Delgado and Irene Aguado. *Report on common KPIs D1.4 r2*. Project Demonstration 646531. [Online; accessed 15. Apr. 2020]. Brussels: The UPGRID Consortium, 2016. URL: http://upgrid.eu/wp-content/uploads/2018/01/151104_UPGRID_WP1_D14_KPIs_v14_final.pdf.

[7] Warren D Devine. 'From shafts to wires: Historical perspective on electrification'. In: *The Journal of Economic History* 43.2 (1983), pp. 347–372.

[8] Doina Dragomir, Christoph Nölle and Speranta Stomff. *Stakeholders' Requirements Analysis Report - D3.1*. Project Demonstration 318782. [Online; accessed 15. Apr. 2020]. Brussels: STARGRID project, 2013. URL: http://stargrid.eu/downloads/2014/07/STARGRID_Stakeholders-Report_D3.1_v1.0_2013_10_11.pdf.

[9] European Parliament, Council of the European Union. 'Regulation (EU) 2016/679 of the European Parliament and of the Council (GDPR)'. In: *Official Journal of the European Union* (2016). [Online; accessed 15. Apr. 2020]. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679#d1e40-1-1.

[10] Maryléne Gagné and Edward L Deci. 'Self-determination theory and work motivation'. In: *Journal of Organizational behavior* 26.4 (2005), pp. 331–362.

[11] Vehbi C Gungor et al. 'Smart grid technologies: Communication technologies and standards'. In: *IEEE transactions on Industrial informatics* 7.4 (2011), pp. 529–539.

[12] W. J. Harder. 'Key Performance Indicators for Smart Grids'. MA thesis. 7522 Enschede: University of Twente, July 2017.

[13] Alan R Hevner. 'A three cycle view of design science research'. In: *Scandinavian journal of information systems* 19.2 (2007), p. 4.

[14] Alan R Hevner et al. 'Design science in information systems research'. In: *Management Information Systems Quarterly* 28.1 (2004), pp. 75–106.

[15] Douglas Hubbard and Dylan Evans. 'Problems with scoring methods and ordinal scales in risk assessment'. In: *IBM Journal of Research and Development* 54.3 (2010), pp. 2–1.

[16] Eva Hudlicka, Greg Zacharias and Joseph Psotka. 'Increasing realism of human agents by modeling individual differences: Methodology, architecture, and testbed'. In: *Simulating human agents, American association for artificial intelligence fall 2000 symposium series*. 2000, pp. 53–59.

[17] Jan Henning Jürgensen, Lars Nordström and Patrik Hilber. 'A scorecard approach to track reliability performance of distribution system operators'. In: *23rd International Conference on Electricity Distribution-CIRED Lyon, 15-18 June 2015*. CIRED-Congrés International des Réseaux Electriques de Distribution. 2015.

[18] Robert S Kaplan and David P Norton. 'Putting the balanced scorecard to work'. In: *The economic impact of knowledge* 27.4 (1998), pp. 315–324.

[19] Robert Lee, Michael Assante and Tim Conway. 'Analysis of the cyber attack on the Ukrainian power grid, Defense Use Case'. In: *Electricity Information Sharing and Analysis Center (E-ISAC)* 388 (2016).

[20] Peng Liu, Wanyu Zang and Meng Yu. 'Incentive-based modeling and inference of attacker intent, objectives, and strategies'. In: *ACM Transactions on Information and System Security (TISSEC)* 8.1 (2005), pp. 78–118.

[21] Michael Martello, John G Watson and Michael J Fischer. 'Implementing a balanced scorecard in a not-for-profit organization'. In: *Journal of Business & Economics Research (JBER)* 6.9 (2008), pp. 67–80.

[22] Eoghan McKenna, Ian Richardson and Murray Thomson. 'Smart meter data: Balancing consumer privacy concerns with legitimate applications'. In: *Energy Policy* 41 (2012), pp. 807–814.

[23] François Meurens et al. 'The pig: a model for human infectious diseases'. In: *Trends in microbiology* 20.1 (2012), pp. 50–57.

[24] Tyler Moore. 'The economics of cybersecurity: Principles and policy options'. In: *International Journal of Critical Infrastructure Protection* 3.3-4 (2010), pp. 103–117.

[25] A Muir and J Lopatto. 'Final report on the August 14, 2003 blackout in the United States and Canada: causes and recommendations'. In: *US–Canada Power System Outage Task Force, Canada* (2004).

[26] Finn Müller-Hansen et al. 'Towards representing human behavior and decision making in Earth system models-an overview of techniques and approaches'. In: *Earth System Dynamics* 8 (2017).

[27] Mashrura Musharraf, Faisal Khan and Brian Veitch. 'Validating human behavior representation model of general personnel during offshore emergency situations'. In: *Fire technology* 55.2 (2019), pp. 643–665.

[28] National Research Council. *Electricity in Economic Growth*. Washington, DC: The National Academies Press, 1986. ISBN: 978-0-309-03677-1. DOI: 10.17226/900. URL: https://www.nap.edu/catalog/900/electricity-in-economic-growth.

[29] NVE. 'KILE – kvalitetsjusterte inntektsrammer ved ikke levert energi'. In: (Oct. 2019). [Online; accessed 15. Apr. 2020]. URL: https://www.nve.no/reguleringsmyndigheten/okonomisk-regulering-av-nettselskap/om-den-okonomiske-reguleringen/kile-kvalitetsjusterte-inntektsrammer-ved-ikke-levert-energi/.

[30] Richard W. Pew and Anne S. Mavor, eds. *Representing Human Behavior in Military Simulations: Interim Report*. Washington, DC: The National Academies Press, 1997. ISBN: 978-0-309-05747-9. URL: https://www.nap.edu/catalog/5714/representing-human-behavior-in-military-simulations-interim-report.

[31] Nicolas Prat, Isabelle Comyn-Wattiau and Jacky Akoka. 'A taxonomy of evaluation methods for information systems artifacts'. In: *Journal of Management Information Systems* 32.3 (2015), pp. 229–267.

[32] Lisa Rajbhandari and Einar Snekkenes. 'Using the Conflicting Incentives Risk Analysis Method'. In: *Security and Privacy Protection in Information Processing Systems*. Ed. by Lech J. Janczewski, Henry B. Wolfe and Sujeet Shenoi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 315–329.

[33] Jaime Sánchez-Ortiz, Teresa García-Valderrama and Vanessa Rodríguez-Cornejo. 'Towards a balanced scorecard in regulated companies: a study of the Spanish electricity sector'. In: *The Electricity Journal* 29.9 (2016), pp. 36–43.

[34] Rafael Santodomingo et al. 'The DISCERN tool support for knowledge sharing in large Smart Grid projects'. In: *CIRED Workshop* (2016).

[35] Günther Schuh et al. 'Information system architecture for the interaction of electric vehicles with the power grid'. In: *2013 10th IEEE International Conference on Networking, Sensing and Control (ICNSC)*. IEEE. 2013, pp. 821–825.

[36] Alina Selyukh. 'NSA staff used spy tools on spouses, ex-lovers: watchdog'. In: *U.S* (Sept. 2013). URL: https://www.reuters.com/article/us-usa-surveil-lance-watchdog/nsa-staff-used-spy-tools-on-spouses-ex-lovers-watchdog-idUSBRE98Q14G20130927.

[37] Einar Snekkenes. 'Position paper: Privacy risk analysis is about understanding conflicting incentives'. In: *IFIP Working Conference on Policies and Research in Identity Management*. Springer. 2013, pp. 100–103.

[38] Henk G Stassen, Gunnar Johannsen and Neville Moray. 'Internal representation, internal model, human performance model and mental workload'. In: *Automatica* 26.4 (1988), pp. 811–820.

[39] Adam Szekeres and Einar Arthur Snekkenes. 'Construction of Human Motivational Profiles by Observation for Risk Analysis'. In: *IEEE Access* 8 (2020), pp. 45096–45107.

[40] Adam Szekeres and Einar Arthur Snekkenes. 'Predicting CEO Misbehavior from Observables: Comparative Evaluation of Two Major Personality Models'. In: *E-Business and Telecommunications. ICETE 2018. Communications in Computer and Information Science*. Vol. 1118. Springer, Cham, 2019, pp. 135–158.

[41] Adam Szekeres, Pankaj Shivdayal Wasnik and Einar Arthur Snekkenes. 'Using Demographic Features for the Prediction of Basic Human Values Underlying Stakeholder Motivation'. In: *Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 2: ICEIS*. SciTePress, 2019, pp. 377–389.

[42] John R Taylor. *An introduction to error analysis: The study of uncertainties in physical measurements*. Sausalito, California: University Science Books, 1997.

[43] Mathias Uslar and Dominik Engel. 'Towards generic domain reference designation: How to learn from smart grid interoperability'. In: *DA-Ch Energieinformatik* 1 (2015), pp. 1–6.

[44] Mathias Uslar, Christine Rosinger and Stefanie Schlegel. 'Security by Design for the Smart Grid: Combining the SGAM and NISTIR 7628'. In: *2014 IEEE 38th International Computer Software and Applications Conference Workshops*. IEEE. 2014, pp. 110–115.

[45] Mathias Uslar and Jörn Trefke. 'Applying the Smart Grid Architecture Model SGAM to the EV Domain'. In: *EnviroInfo*. 2014, pp. 821–826.

[46] Mathias Uslar et al. 'Applying the smart grid architecture model for designing and validating system-of-systems in the power and energy domain: A European perspective'. In: *Energies* 12.2 (2019), p. 258.

[47] Richard W Waterman and Kenneth J Meier. 'Principal-agent models: an expansion?' In: *Journal of public administration research and theory* 8.2 (1998), pp. 173–202.

[48] Eva Weishäupl, Emrah Yasasin and Guido Schryen. 'Information security investments: An exploratory multiple case study on decision-making, evaluation and learning'. In: *Computers & Security* 77 (2018), pp. 807–823.