Co-inform

Context Matters,
Your Sources Too

Evaluation Methods

D5.1

Document Summary Information

Project Title	Co-Inform: Co-Creating Misinformation-Resilient Societies		
Project Acronym	Co-Inform	**Proposal Number:**	770302
Type of Action	RIA (Research and Innovation action)		
Start Date	01/04/2018	**Duration:**	36 months
Project URL:	https://coinform.eu		
Deliverable:	D5.1: Evaluation Methods		
Version:	5		
Work Package:	WP5		
Submission Date:	01/04/2019		
Nature:	Report	**Dissemination Level:**	Public
Lead Beneficiary:	Cyprus University of Technology		
Author(s):	Eleni Kyza, Christiana Varda, Dionysis Panos, Evangelos Karapanos, Loukas Konstantinou, Melina Karageorgiou		
Contributions from:	Love Ekenberg (IIASA) Nadejda Komendantova (IIASA) Syed Iftikhar Shah (IHU) Ipek Baris (UKOB) Tracie Farrell (OU) Lara Schibelsky Godoy Piccolo (OU)		

Co-Inform is co-funded by Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020) H2020-SC6-CO-CREATION-2016-2017 (CO-CREATION FOR GROWTH AND INCLUSION).

Revision History

Version	Date	Change Editor	Description
1	28/3/2019	CUT	Initial draft
2	5/4/2019	IIASA, IHU, UKOB, OU	Review
3	10/4/2019	CUT	Revisions, Formatting
4	12/4/2019	CUT	Final Review, Formatting, Proofreading
5	15/4/2019	CUT	Final Review

Disclaimer

Copyright Message

Executive Summary

Rising concerns about the issue of misinformation and the dissemination of misleading information that could potentially cause harm to democratic processes, makes it essential to evaluate how information is accessed, understood and shared on online digital platforms. Co-Inform approaches misinformation through the co-creation of socio-technical solutions to address the challenges of misinformation with three stakeholder groups (citizens, journalists/fact-checkers, policymakers) who will allow the consortium to gain insights on the perceptions, practices, and challenges faced by each group.

D5.1 provides a methodological framework for the evaluation of the effectiveness of the Co-Inform approach and intervention strategies with the goal to promote a misinformation-resilient behaviour for each of the three stakeholder groups. This work is closely connected to the three co-creation pilots which will take place in Austria, Greece, and Sweden, and which provide the primary context for collecting information on stakeholder practices, perceptions, and needs to inform the app development processes happening in WP2, WP3, and WP4.

This deliverable provides a description of the methods that will be used by Co-Inform to evaluate the behaviours and perceptions of misinformation of each of the stakeholder groups: citizens, who are defined as young adults age 18-24 (T5.1); policymakers, a group which includes politicians, researchers, support staff (T5.2); and journalists, a group which includes journalists and fact-checkers (T5.3). In addition, D5.1 describes a decision analytical model (T5.4) which will be used to support the evaluation of policies regarding specific misinformation contexts using data from the work conducted in WPs 1-4.

In order to assess the three stakeholder groups' perceptions, practices and insights associated with Co-Inform tools, we aim to use a variety of data collection tools (e.g. Cognitive Walkthroughs, Heuristic Evaluation, Focus Groups, Interviews, etc.), which are detailed in this deliverable. Evaluation is conceptualized as consisting of both formative and summative evaluation activities which will inform iterative design and evaluation cycles and provide feedback to consortium partners to impact the development of an effective co-created misinformation-assessment tool. Evaluating the effectiveness of the Co-Inform technologies is crucial in order to develop tools and intervention strategies that may persuade misinformation-resilient behaviour in each of the stakeholder groups.

Where necessary, additional material that accompanies these methodologies is provided in the appendix.

Table of Contents

1. Introduction

1.1 Background

Social media platforms have considerably transformed since their inception in the early 2000s. Though social media platforms have democratized media by allowing millions of people to share and produce content of their own, they have also blurred the lines defining the dichotomy of truth vs falsity. Misinformation is defined as *"any false or inaccurate information that is spread either intentionally or unintentionally*" (Antoniadis et al., 2015: 475). The changing media landscape, which has impacted the news media industry, has led to growing concerns about the impact on democratic processes. The Co-Inform project aims to address this issue by bringing together three stakeholder groups that are impacted by misinformation in different ways. For citizens, access to the news is increasingly mediated through social media platforms; journalists rely on the social media platforms to share news, connect with sources and engage with readers; policymakers require access to factual, verifiable information that can enable them to take informed policy decisions. The disparate needs of each stakeholder group highlight the need of identifying and understanding the practices and perceptions on misinformation for each group, in the effort to propose effective solutions.

1.2 Citizens and misinformation

The information-abundant context that prevails in most social media platforms, and the diminishing role of gatekeepers in the contemporary news economy, increases the likelihood of exposure to inaccurate information, while making misinformation detection more challenging. Gualda and Rúas (2019) conducted a survey to evaluate what citizens believed about the information they received, and whether they believed information was withheld from them. Glenski, Weninger and Volkova (2018) found that users who share information from clickbait and conspiracy sources are also likely to share from propaganda sources. Findings from the Reuters Institute for the Study of Journalism provide key insights gained from eight focus groups and survey of online users in the United States, the United Kingdom, Spain and Finland about users' perspectives towards "fake news". The study found that people attribute slight differences between fake news and accurately reported news, and when asked to give examples of misinformation, are more likely to associate it with poor journalism, propaganda and some form of advertising. The study also found that discussions around "fake news" give rise to a general distrust of news media, politicians and platforms and while participants could identify sources that they consistently considered reliable, they tended to disagree on which sources are considered universally reliable for all users (Nielsen & Graves, 2017).

The widespread use of social media to share information amongst peers, has served to decentralize information-sharing from authority sources and has rendered traditional credibility assessment strategies, such as reliance on authority figures or experts, outdated. According to Callister (2000) traditional credibility techniques can work when there is information scarcity, because it allows gatekeepers to produce and filter the information and provides an incentive for upholding credibility standards. Credibility is not inherent in the information or the source,

i.e. it is not a separate property per se, but it is judged by the receiver of the information (Gunther, 1992). Perceptions of credibility can be very situational and may also be impacted by the source of the message, the message itself, as well as the receiver's relationship to the medium (Cronkhite & Liska, 1976; Gunther, 1992). Many users rely on others to make credibility assessments and also rely on cognitive heuristics to evaluate information and sources online, rather than systematically processing information (Flanagin & Metzger, 2007).

Among the challenges of mapping users' credibility evaluation is the noted discrepancy between self-reporting and observed behaviour of credibility evaluation, which may be influenced by factors of social desirability; Flanagin and Metzger (2007) found that participants' self-reported verification methods did not correspond to their observed behaviours, except for experienced users of the web, who were more likely to accurately self-report their credibility behaviours.

Online credibility research has focused on textual information (Morris et al., 2012; Wineburg & McGrew, 2016), fake image detection using algorithmic machine learning approaches (Gupta, Lamba, Kumaraguru & Joshi, 2013; Rath, Gao, Ma & Srivastava, 2017), image authentication using social and cognitive heuristics (Shen, Kasra, Pan, Bassett, Malloch, & O'Brien, 2018) and news and information credibility evaluation using eye-tracking research (van Strien et al., 2016; Sülflow et al., 2019). On Twitter, content alone is not enough to evaluate the truthfulness of a post, and users tend to rely on heuristics such as user names to assess the reliability of posts, according to Morris et al. (2012). Features such as using of non-standard grammar, having a default account image or using a cartoon or avatar as an account image, received low credibility scores by participants; Twitter users who also had an unbalanced ratio accounts followed to number of followers, were also greeted with mistrust.

Closely linked with credibility evaluation, is trustworthiness of information, which can be assessed by taking into consideration variables such as accuracy, objectivity, validity and stability; the constructs of trust and credibility are differentiated by Kelton, Fleischmann, and Wallace (2008) who define trust as dependability and credibility as believability. Kelton et al. also discuss four levels at which trust can be examined: individual (personality characteristic), interpersonal (social tie between people), relational (emergent property of a mutual relationship), and societal (community-based feature, system-based trust). The authors indicate that the interpersonal type of trust is the one that has been mostly investigated.

The nature of the information sought or examined is also important. In a small-scale study, Heath, Motta, and Petre (2006) found differences in how critical the tasks were perceived; for example, in low-criticality tasks participants were willing to use less trustworthy sources than in high-criticality tasks (i.e. looking for a treatment for back pain). In their study, Heath, Motta, and Petre (2006) examined five trust factors: expertise, experience, impartiality, affinity, and track record. The two most prominent factors were expertise and experience. Affinity was more prevalent in situations that allowed for subjective decision making, such as taking a vacation.

1.3 Journalists and misinformation

"*We think the answer rests less on what journalists do - basically, gathering and sharing information, which lots of folks online are doing, too - but how and why they do it. It rests, that is, on ethics.*" (Friend & Singer 2006, p. xv)

The above quote follows the question posed by Friend and Singer, as to where the journalist fits in a world of digital media in which anyone can be a publisher. In an era where everyone can act like a journalist, what differentiates professional journalists from the rest is their level of judgment which is based not only on knowledge and experience but also on their journalism ethics. The debate about the existence and the characteristics of the citizens' journalism phenomenon is already long. Whatever the different approaches, we suggest that citizens' journalism does not necessarily mean rejection of professional journalism values but on the contrary, indicates a need to extend professional journalism values to non-professional (citizen) journalists. As Ward claims, "the globalization of news media requires a radical rethinking of the principles and standards of journalism ethics"(Ward, 2005, p.1). While the spread of technology and globalization have undoubtedly led to the rise of citizens' journalism, the importance of (traditional) professional journalism values, such as the pursuit of truth and accuracy, objectivity and impartiality, are still crucial.

For the last two decades, various authors have drawn attention to the ethical challenges of digital journalism. As Elliot (2008) points out, the Web allows unprecedented access to the opinions of others and to information from credible (and incredible) sources. But, as the author adds, professional journalists, with commitment to the essential shared values of the practice, are necessary to the development and sustenance of democratic process (Elliott, 2008, p.28). In their effort to separate journalists from imitators, Borden and Tew (2007) insist on the moral commitments journalists make, indicating that those get expressed in journalistic performances. According to Lynch, journalists should ask themselves, "*How do we make good decisions in an environment that has neither a long journalistic tradition nor an opportunity for reflection?*" (Lynch, 1998 as cited in Deuze & Yeshua, 2001). Web communication set new rules on defining both professionalism and ethics in journalism. Deuze and Yeshua (2001) base their analysis on the idea that the Web shapes and redefines a number of moral and ethical issues confronting journalists when operating online or making use of online resources.

The above summarize the reasons which led us to dedicate part of the D5.1 data collection to interviews in exploring the journalists' personal and professional values. Subsequently, we seek to find out how journalists perceive challenges related to misinformation and how they handle them. We do believe that there is need for more thorough research and analysis dealing with the ways in which the Web affects ethics and moral decision-making in journalism. Most importantly, we bear in mind that the values of the user lay in front of any app, which aspires to tackle the misinformation problem.

1.4 Policymakers and misinformation

Policymakers are particularly affected by the changing media environment since they require access to accurate and reliable information on which they can base decisions that can impact the wider community. As a stakeholder group, this is a diverse cohort, since it includes

politicians, who hold positions of authority, both practically and virtually, on digital platforms, but also researchers, analysts and assistants who hold positions of expertise, and would, by traditional means of credibility evaluation, be considered as authorities in their relevant disciplines / fields. The potential threat of misinformation to the democratic process makes understanding the attitudes, needs and challenges of this stakeholder group particularly important.

There are, however, limited studies that focus on investigating the attitudes, credibility practices, and proposed solutions as they may stem from the needs of policymakers. As persons of authority, the role of policymakers is often resigned to the concluding sections of studies centering on misinformation, and tend to focus on implications of misinformation for policymakers and decision-making (e.g. Spohr, 2017), interventions needed by policymakers to address the issue of misinformation (e.g. Alemanno, 2018), or the role of opinion leaders' influence (virtually defined with a higher number of followers) in propagating information (Pang & Ng, 2017).

1.5 Alignment with Co-Inform's broader vision

The innovation of the Co-Inform project lies in its potential to create a technological tool that stems from the needs of each of the aforementioned stakeholder groups. The challenge for the data collection and evaluation processes on the Co-Inform project is manifold. Beyond the varying needs of each group, the composite nature of the journalists' group (comprised of journalists and fact-checkers) and the policymakers' group (comprised of politicians, research analysts, decision-makers) is one that is important to consider for each co-creation workshop. Additionally, the three different contexts for each of the workshops (Austria, Greece, Sweden) may make drawing data that is comparable, and that could lead to effective feedback for WP2, WP3 and WP4 particularly difficult to elicit. For these reasons, it is important to ensure that across the three sites, the same evaluation instruments are utilized in a way that is replicable, consistent and which addresses the gaps in the needs, as they are defined by WP2, WP3 and WP4. We provide an overview the context for WP5's data collection framework in Figure 1.

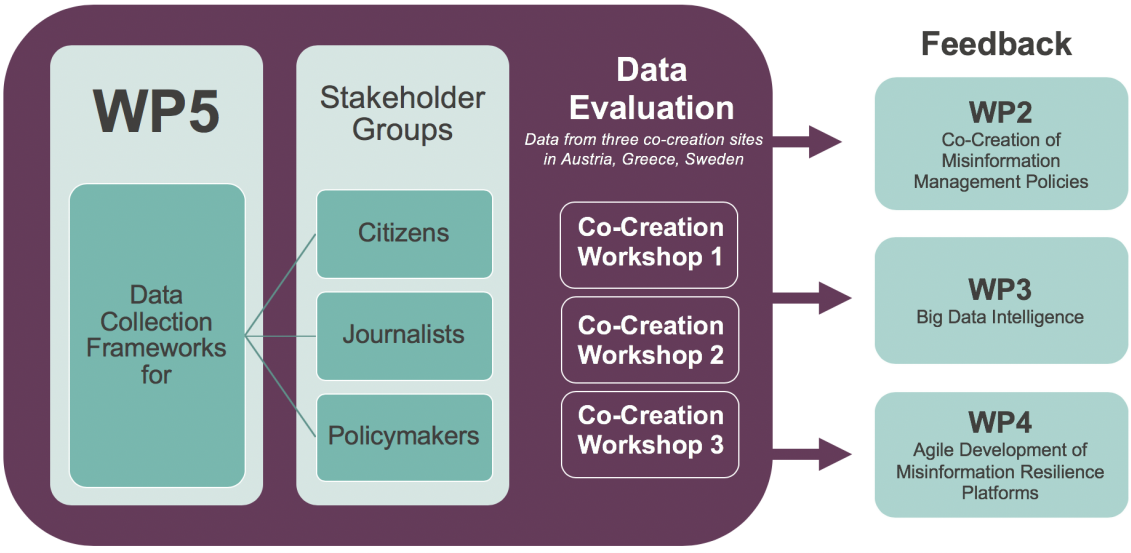

Figure 1. WP5's data collection framework

In this deliverable we describe a framework which will guide the data collection procedures allowing the Co-Inform consortium to elicit formative feedback from the co-creation activities that will take place in the three pilot countries: Austria, Greece, Sweden. The deliverable aims to provide an overview of the evaluation instruments that will be used, in order to gain a greater understanding into each of the stakeholder groups' perceptions, needs, credibility evaluation procedures and challenges.

We begin by addressing the role of WP5 within the wider consortium and continue by providing an overview of the types of evaluation instruments that will be implemented in each of the co-creation sites, according to the changing needs of each of the co-creation workshops. In Section 2 we detail each of the methodological tools analytically, including the intended aims of each instrument, as well as the intended outcomes and how these align with Co-Inform's projected targets for each of the co-creation workshops. Section 3 presents the decision analytical model to guide policymakers' decision making. Section 4 provides a risk assessment and considers a contingency plan for possible issues that might arise in the data collection process amongst partners; this is directly connected to Section 5, which outlines the synergies within the Co-Inform consortium, and specifically WP5's collaborating institutions and their obligations / role in facilitating an optimum data collection process across the duration of the project. Finally, in Section 6 we briefly touch on links with other Co-Inform deliverables regarding the ethical and privacy-related issues that arise from WP5's central role in collecting and evaluating data from each of the co-creation sites.

2. Data Collection Framework

Empirical studies that are focused on misinformation tend to take a quantitative methodological approach. Large-scale surveys are most often used in order to gain insights into participants' perceptions of misinformation (e.g. Gualda & Rúas, 2019), credibility evaluation approaches (e.g. Shen, Kasra and Pan, 2018; Flanagin and Metzger, 2007) and analysis of online content on Twitter to investigate propagation of information or user behaviour (Glenski, Weninger and Volkova, 2018; Gupta, Zhao & Han, 2013; Pang & Ng, 2017; Bastos & Mercea, 2018). Metzger, Flanagin and Medders (2010) deviate from survey and experimental evidence by conducting focus groups to examine assumptions about information credibility.

The multidimensional aspect of misinformation, and the different stakeholder groups that are affected by it to varying extents, makes it particularly important to garner information that is nuanced, and which allows us to explore participants' attitudes, practices, perceptions on trust and credibility evaluation, and challenges faced, through evaluation instruments that promote reflection and discussion, rather than artificial scenarios or removed survey questions. Delving deeper into the issues faced by each of the stakeholder groups (citizens, journalists, policymakers) participating in the co-creation workshops, will allow WP5 to provide effective feedback that stems from the expressed needs of each of the groups.

2.1 Evaluation Objectives

In order to develop the Co-Inform technologies, and proceed to evaluate their effectiveness, it is essential to utilize assessment tools that will enable us to garner specific insights as to the cognitive, behavioural, affective aspects of misinformation for each of the stakeholder groups. The main context for evaluating the Co-Inform approach will be the co-creation pilot sites. Other data may be collected in studies organized by partners to address their ongoing needs (i.e. testing dashboard interfaces) or to gain more in-depth information on issues relating to the project. In addition, larger scale evaluation data will be sought at the end of the project, to test the effectiveness of the developed solution. These data collection efforts will be in addition to the co-creation workshops and their details will be specified at a later time, when the specific context and technologies are known.

As mentioned earlier, one of the main tasks for WP5 is to propose an assessment methodology which will guide the data collection processes at each of the co-creation sites (Austria, Greece and Sweden). The assessment instruments provided will vary according to the focus of each of the co-creation workshops, will be connected to the activities organized by WP1, and will include qualitative and quantitative, formative and summative methodologies that will provide feedback for Co-Inform policies (WP2), and designs and developments (WP3, WP4). Further details regarding the scope of each workshop are provided in Sections 2.2 - 2.4. Given that the workshops will be planned sequentially, planning for the workshops takes on an open-ended, adaptable approach. WP5 will select methodological instruments that also respond to the changing needs and aims of each workshop, as these are co-determined by the consortium through WP1, WP2, WP3 and WP4. As such, we note that the data collection approaches that are provided in this document are indicative, and may be adapted / modified in order to address data evaluation requirements.

Table 1.
WP5 evaluation objectives and deliverables

Objective	Deliverable
Assess the effectiveness of Co-Inform platform interventions in persuading a misinformation-resilient citizen's cognitive and behavioural change, as informed by the outcomes of WP1.	D5.1
Assess the effectiveness of the Co-Inform tools in supporting journalists' practices of misinformation discovery and fact-checking dissemination, as identified in WP1.	D5.1
Assess the effectiveness of the Co-Inform tools in supporting policymakers' practices and formation of informed policy, as these were identified in WP1.	D5.1
Provide regular feedback and design recommendations to WPs 2-4.	D5.2
Investigate how individual characteristics and attitudes, social factors and media design influence when and how citizens assess the credibility of online information and examine their misinformation-related practices as these relate to the Co-Inform tools.	D5.4
Develop and deploy a risk analysis and decision theoretical model for assessing policies and policymaking, using the data collected in WP5.	D5.3

2.2 Data Collection Indicative Approaches

2.2.1 Contextual Inquiry

Contextual Inquiry methods (Whiteside et.al. 1988; Wixon et al., 1990) are user-centred design approaches which aim to shed light to the question of how human users interact with computer systems in their everyday hands-on environment. In the Co-Inform evaluation framework, the contextual inquiry approach will be used prominently during in-depth personal interviews. For instance, researchers will be asked to conduct personal interviews (as conditions allow) with journalists/fact-checkers and policymakers at the participants' workplace, where they can use her/his own working environment and tools that they actually use in everyday professional life.

2.2.2 Behaviour Change (Nudging)

Empirical studies have repeatedly highlighted that misinformative content propagates faster, deeper, and farther than truthful messages. Vosoughi et al. (2018), for instance, used a data set of rumour cascades on Twitter from 2006 to 2017, and found that the top 1% of false news cascades diffused to between 1000 and 100,000 people, whereas the truth rarely diffused to more than 1000 people. A key question raised is: what role does human decision-making play,

and how can technology enable humans to make better decisions? Recent studies have highlighted that cognitive biases in decision making can facilitate the spread, or the consumption of misinformative content. For instance, Vosoughi et al. (2018) found that, contrary to conventional wisdom, the spread of false news could not be attributed to the structure of social media outlets, website platforms and internet bots, but rather to a mere novelty effect. Novelty, as the authors claimed, “attracts human attention, contributes to productive decision-making, and encourages information sharing because novelty updates our understanding of the world”. False news was found to be more novel than true news, suggesting that people were more likely to share novel information.

Badke (2018) claim that humans see only what they expect or want to see, without inspecting news thoroughly. This, he argued, is a product of confirmation bias, the internal tendency of people to seek out information that confirms and verifies what they already believe, instead of examining critically all the pieces of information. According to the theory of cognitive dissonance (Festinger, 1957), whenever a presented piece of news includes information which conflicts with the currently held mental models of people, it immediately induces cognitive dissonance. People are motivated to scale down this dissonance, thus they may avoid or even discount knowledge that contrasts their personal positions. Weeks (2015) argues that emotional experience moderates the influence of partisanship on individuals’ responses to misinformation. Specifically, when individuals experience anger, the influence of partisanship is boosted, making individuals more likely to believe claims that are associated with their political affiliation. On the contrary, anxiety reduces the influence of partisanship and increases the chance of making other political affiliations believable. Schwarz et al. (2016) argue that whenever people come across a new piece of information, they tend to assess its truthfulness by focusing on five criteria. People usually ask themselves about the social consensus of the story, its supporting evidence, its consistency, coherence and credibility. However, instead of evaluating these questions analytically, individuals tend to use mental shortcuts in order to minimize the time and energy spent. This makes them susceptible to errors in decision making.

Given the accumulating knowledge on the cognitive biases that facilitate the spread, or the consumption of misinformative content, designers can leverage this to develop technological interventions that "nudge" individuals towards desirable behaviours. A nudge is defined as “any aspect of the choice architecture that alters people’s behaviour in a predictable way without forbidding any option or significantly changing their economic incentive” (Thaler & Sunstein, 2013). Grounded on empirically proven cognitive biases, leading to systematic deviations from rational decision making, nudges offer the premise of effective, yet unobtrusive behaviour change interventions.

2.2.3 Unified Acceptance and Use of Technology (UTAUT)

UTAUT (Venkatesh, Morris, Davis, & Davis, 2003) comprises of four factors (performance expectancy, effort expectancy, social influence, and facilitating factors) and four moderators (age, gender, experience, and voluntariness) to predict intention to use a technological innovation (Venkatesh, Thong, & Xu, 2016). UTAUT is argued to be a better model for predicting an individual’s technology acceptance as compared to the eight models it is based upon and has been widely used by researchers for evaluation and theory-building. In Co-

Inform, UTAUT can be used to evaluate the Co-Inform app adoption by the broader public. This evaluation study will necessarily take place at the end of the development process, will be conducted online and offline, and, in addition to the co-creation participants who will be invited to participate, the study will be widely promoted to users outside the co-creation workshops who volunteer to participate in this research. The results of this study will help the Co-Inform consortium gauge how the Co-Inform app users perceive the app.

2.3 Data Collection Techniques

We will follow an iterative, mixed-method process for the evaluation of the developed behaviour change interventions. The approach will also investigate the connection of the Co-Inform technologies to the participants' media literacy. As shown in Figure 2, the consortium is currently planning a minimum of three co-creation workshops; however, more workshops may be organized as needed. Due to the nature of the work, it is not possible to fix the number of co-creation workshops in absolute terms at the moment.

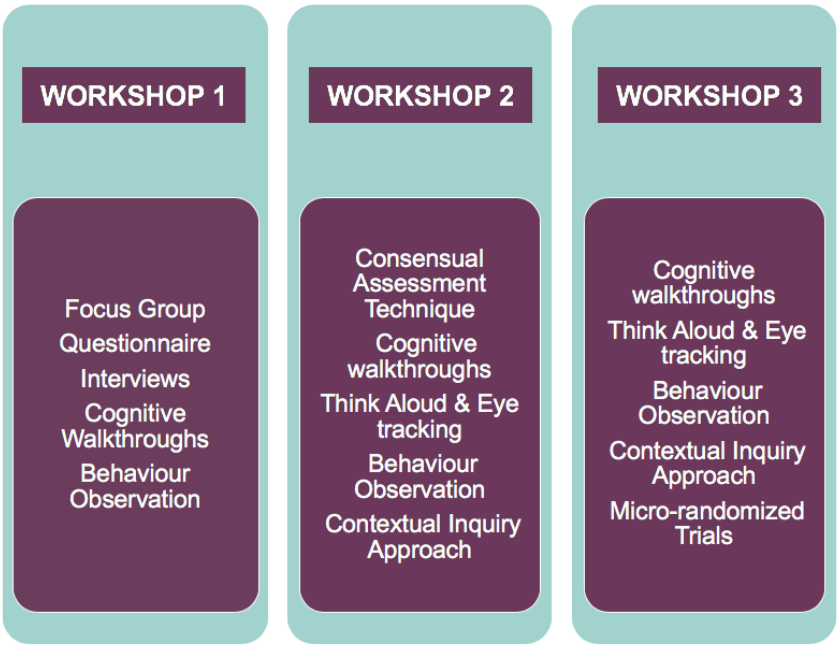

Figure 2. Indicative methodologies that can be used during the co-creation workshops for establishing a baseline and evaluating the impact of the Co-Inform tools.

2.3.1 Consensual Assessment Technique

To support the development of technological interventions that promote desirable behaviours upon citizens, we have developed a design tool, the *Nudge Deck*, which consists of design cards specifying 23 nudging mechanisms tapping to 15 different cognitive biases. The Nudge Deck was formed on the basis of the results of a systematic review of the application of nudging in Human-Computer Interaction literature. We will use the Nudge Deck in design workshops with design students with the goal of producing design solutions that promote desirable behaviours in the context of misinformation. Following the workshop, the design ideas will be evaluated with regards to their creativity and fitness to the context of use, by a panel of experts, using the Consensual Assessment Technique (CAT).

The CAT was first proposed by Amabile in 1982 (see Baer, 2015) as a subjective means to evaluate creativity. The CAT asks expert judges to rate the creativity of a set of stimuli, individually, and in isolation, employing a given rubric (i.e., often an ordinal scale from 1 to 5), without any further justification of their provided scores. The CAT has also been used to measure the "Novelty", "Usefulness", "Effort", "Elaboration" of different solutions, among others. The level of agreement among judges is then estimated, usually, through calculating Cronbach's alpha, and given acceptable agreement among judges the mean or median score for each stimulus is calculated.

2.3.2 Cognitive Walkthroughs

One of our basic research assumptions concerning the Journalists/Fact-Checkers stakeholder group is that misinformation for this particular group is mainly a professional problem, stretching in two different directions: values (professional & personal) and practices/routines. Therefore, the journalists' professional environment is a crucial research field we should focus upon. Using the Cognitive Walkthrough methodology, both directions can be inquired. Cognitive Walkthrough is a well-established methodology (developed in the early '90s but widely used since the mid-00's) mainly used to evaluate the usability of interactive systems and commonly used in HCI (Tching et.al. 2016; Mahatody et.al. 2010; Allendoerfer et.al. 2005). The basic idea behind this method is that users prefer to get accustomed to a system by using it, rather than by reading instructions and manuals. Consequently, this method is at large task-oriented and can provide useful results not only on specific details but also for the overall picture of the problem under investigation.

In our research design, we will use both the journalists' real space working environment (if we are allowed entry) and an in-vitro simulation of this environment. Journalists & fact-checker focus groups participants will be given specific tasks to perform and case-studies to deal with during the pilots' study research stage. Following this stage, researchers will visit participants who have previously consented to an in-depth personal interview. An effort will be made to visit participants at their workplaces and therefore have the opportunity to inquire deeper in their everyday professional routines, apps or tools used, problems they face and choices they make. Participants will be given possible scenarios, information tips, news headlines, social media posts, and news stories and will be asked to follow their usual professional routines in order to verify the validity, the accuracy and the truthfulness of each scenario. Researchers will track down and map the paths they follow in order to accomplish the given tasks and at the same time, they will be able to identify possible problems or obstacles participants face in their effort to validate or discard the given scenario.

On a second level, participants will be given app prototypes (early draft versions) to use in order to accomplish the same tasks. In that sense, the Cognitive Walkthrough methodology will be used in its original purpose – to inquire the usability of an interactive system (namely the prototype app). Results from this procedure along with research results from the Co-Creation Workshops of WP1 will be fed back to WP2, WP3 & WP4, so that app prototypes can be improved, fine-tuned and perfected.

The usability of early working prototypes will be evaluated in a usability evaluation laboratory by Human-Computer Interaction experts, with the use of the cognitive walkthrough technique. Usage scenarios will be developed, in which a user is exposed to online content and engages with Co-Inform's technological interventions; the experts will simulate, in a step-wise fashion, the anticipated user and system responses. Empirical findings will be communicated to WP4 along with a description of the anticipated usability problems as well as directions for the redesign of the platform.

2.3.3 Think Aloud and Eye Tracking techniques

The think-aloud method can be used both during the in-depth personal interviews stage of research. As Charters (2003) states "think-aloud is a research method in which the participants speak aloud any words in their minds as they complete a task" (2003:68). The think-aloud method (known also as a think-aloud protocol) was introduced by Lewis (1982) but it was further developed and established by Ericsson and Simon (1993). Origins of verbal protocols used as research data can be traced way back in time but this kind of method was rejected by behaviourists for a long time until it regained acceptance as valid instruments in the late 70s. Since then the think-aloud method has been widely used in psychology (Güss 2018), education and learning research (Johnstone et. al. 2006; Masood & Thigambaramb 2015), sport and health research (Eccles & Arsal 2017), serious games design (Nawaz, 2015) but also prominently in usability research (Boren & Ramey 2000; Alshamari et.al. 2015). A variation of the method, under the name Talk-Aloud, also exists; in this approach, the research participant is instructed to say aloud only the moves or actions she/he is doing in order to complete the given task but not spontaneous thoughts or connotated words. Supporters of the Talk-Aloud variation suggest that it is more objective than the Think-Aloud as the participant voices only hers/his actions and not their subjective interpretations and thoughts.

In the Co-Inform research framework, the Think-Aloud method will be used in order to supplement other types of data collection, such as the Cognitive Walkthroughs. Participants will be asked to verbally state actions taken in order to verify and validate information or a news story within a given task or a research scenario. It is important to consider that the constant verbalization of their thoughts might distract participants in what they do. As a result, thinking aloud can be accompanied with eye tracking, if this is available at the co-creation partner sites and can be used by the Co-Inform partners.

Once mature versions of the technological prototypes are available, these will be tested in a usability laboratory with users. During the evaluation, participants will be asked to think aloud and in order to inquire into the impact different interventions have on participants' cognitive processes. Eye tracking equipment can also be used with the goal of assessing the extent to which different kinds of information, as well as visual layouts, attract the users' attention.

2.3.4 Behavioral Observation

The behavioral observation method is one of the most widely and long-time used methods of Social Sciences in general. Behaviour scientists of the early '20s developed the first behavioural sampling techniques (Suen & Ari, 1989) with developments and refinements taking place through the years (Altmann, 1974; Alevizos et.al., 1978) until today. The behaviour observation research method is used in a wide variety of scientific fields from

criminology and prevention science (Snyder et.al., 2006) to digital systems (Yang et.al., 2015) and web platform design (Lee & Seo, 2015).

Within the Co-Inform research framework, behaviour observation data collection techniques are mostly task-oriented. In various stages during focus groups, in-depth personal interviews and co-creation workshops, participants will be given tasks to accomplish, dilemmas to decide, scenarios to follow and problems to deal with. Activities will be recorded and transcribed in order to provide diverse but comparable research data.

2.3.5 Micro-randomized trials

In order to evaluate the efficacy of the technological interventions to incur behavior change, we will employ a novel technique called Micro-Randomized Trials (MRTs). Randomized Controlled Trials (RCTs), the gold standard of efficacy assessment, are not well suited for the evaluation of complex technological interventions where multiple intervention components may co-exist and the researchers' interest is in the efficacy of each component. To address this problem, MRTs randomize treatments, from the set of possible treatments, each time a participant interacts with the technology. This way one can study the proximal effects of each intervention component separately, and inquire into which interventions work for whom, and under what conditions (Klasnja et al., 2015). As a result, competent multi-component interventions can be developed since these trials answer whether or not to include time-varying components as part of interventions and in which contexts the effects of the components are most effective. For instance, moments of temptation to smoke might be a turning point towards complete relapse or abstinence. With micro-randomized trials, interventions could be formed which adaptively respond to individuals' actions and are delivered when and where they are most needed. In our pilots, we plan to employ micro-randomized trials and observational studies in order to uncover the proximal effect the technological interventions developed in the context of Co-Inform have on users' behaviours and enable the assessment of their efficacy in real life conditions.

2.3.6 Validated questionnaires

Using validated questionnaires will enable us to collect information factual information relating to participant demographics as well as views and attitudes towards the topic of misinformation. Questionnaires are a useful methodological tool because they can allow us to gain insights through having access to comparable data, across the three co-creation sites. For instance, for the purposes of the first pilot workshop, the EU's Flash Eurobarometer questionnaire 464, which was designed to explore EU citizens' awareness of and attitudes towards the existence of fake news and disinformation online, will be employed. The use of this questionnaire will enable the comparison of data from the first co-creation workshop participants to the general EU population and to the participating countries, according to the published findings of this questionnaire on 'Fake News and Disinformation Online'.

2.3.7 Focus Groups

The focus group data collection method is one of the most appreciated techniques, especially in the Social Sciences. This approach is widely used by researchers for many decades now (Morgan, 1998) with Paul Lazarsfeld and R.K. Merton being credited with formalizing it in the early '40s (Madriz, 2000). A usually small number of participants (e.g. 7-12) form a group; with the guidance of a moderator / facilitator, they have a discussion "focused" on a specific subject or thematic area, providing useful and simultaneously multiple qualitative research data (Wilkinson 2004; Onwuegbuzie et.al., 2009). Therefore, focus groups can provide data that go beyond a simple sum-up of the individual participants' opinions, and which can indicate possible directions of social trends formation capturing at the same time the "group-dynamic" interaction. Within the Co-Inform research framework, moderators will employ a number of projective techniques (indirect data-collection techniques that can bring to surface opinions and attitudes that otherwise would have remained hidden in the direct questions' discussion mode). The following section provides a more detailed description of these Projective Techniques.

2.3.7.1 Focus Groups Projective Techniques

Mock-Up Scenarios

This technique aims to simulate and track down the exact web routines one follows to accomplish her/his professional (journalist/fact-checker, policy maker) or private (citizens) tasks. Ideally, this technique requires every participant to have her/his own workstation with a pre-loaded tracking software or/to an eye-tracking device. If this is not possible, then the participant describes in as much detail as possible (using the think-aloud method, as described above) the paths followed / the sites visited / the thoughts at the time/decisions making.

Each co-creation research team will have prepared in advance three sets of information related to local/national/ regional issues and accordingly three sets of information related to international/global issues. Information sets can be extracted either from misinformation / "fake news" data set (already collected by the Co-Inform partners) or according to researchers' knowledge about highly interesting issues regarding each pilot-study country or specific issue each pilot-study country focus upon. Different sets can be submitted to participants according to the pattern: one accurate information vs two false pieces of information (both on local/regional or international set). Information will be presented to the participants and they will be asked to describe step-by-step the validation/verification methods or paths that they use in order to confirm the accuracy of their professional information.

Example:
"*President Donald Trump said that he will consider seriously all the alarming scientific reports about climate change because he doesn't want China to take the global lead on an issue of paramount importance like that.*" (misinformation - false)

This technique (simulating some possible real-life cases) aims to reveal, on the one hand the daily professional routines of information verification and on the other, to provide research data on the "trust levels" of media professionals (whom they trust most and why?).

Mind-mapping

This technique aims to trigger a dialogue about “values” by asking the media professionals to reach a consensus on prioritizing them. The moderator puts on the table, in random order, A4-Plasticized Boards each one referring to following concepts: Truth / Accuracy / Validity / Profit / Recognition / Facts / Fiction / Sales / Audience. The moderator asks participants to discuss these concepts, as a group, trying to prioritize their choices ranking them from "Most Important" to "Least Important", and also explain and justify them. Through the process of achieving a consensus/agreement about the importance of each concept (value), we will be able to map both their individual opinions and the collective (group-dynamic) final decision.

Spontaneous Response

This technique aims to map the unbiased top-of-mind associations and connotations between notions, thematic areas and concepts under investigation. Different options can be used in implementing this technique. The moderator can ask participants to write on a blank sheet of paper the first three or five words that come up in their minds when listening, for example, to the word "Misinformation" or "Fact Checking". After completing the lists, moderator collects all individual papers and starts to randomly discuss with the whole group what is written in each one. Alternatively, the moderator has prepared in advance blank “thought-bubbles” and distributes them among the participants asking them to fill-in the "thoughts" provoked by reading a specific news-line (one accurate & one false). Another alternative might be for the moderator to prepare a two-column A4 paper with an equal number of pre-given specific words or concepts to each column and ask the group members to associate each element of one column with another one from the opposite column.

It should be noted that this technique is usually used right at the start of the discussion in order to avoid any biases that may possibly arise during the interaction between the group members.

Being the opponent

This is a classic role-playing debate technique. The moderator splits the group members into two different teams assigning one team to defend a specific argument and the other team to defend the opposite. The two teams will start a debate trying to convince the other about their own argument. When the first round of the debate concludes, roles (arguments) will change and proceed to the next round of debate. For example, in the first round, one team has to defend the argument: "Misinformation is nothing new and it’s a phenomenon always existing in societies”, while the other has to defend the argument that “Misinformation is a new phenomenon affecting global society like no other before”. Note that the defense of the team's assigned argument does not necessary coincide with the individual opinions of group members. The aim of this technique is to bring light on the perceptions about the motives or the goals of each opposite argument.

Time traveler technique

The moderator asks the group members to imagine themselves in a future environment where the misinformation (“Fake News”) problem does not exist anymore. What would be the

possible solution that future societies have to adopt in order to solve the problem? What would be the ideal according to their opinion? What would be possible or feasible to be done in the future? How do they project themselves as media professionals working in a future-environment where the misinformation problem will be eliminated? This technique aims to reveal in a clear way the expectations both about the “ideal” and the “possible” solution of the misinformation problem. Data deriving from the “time-traveler” technique will be valuable to track down the limits between the “ideal” and “feasible” (what we aspire to be done and what can actually be done).

Decisions on dilemmas

This technique is similar to the first one (mock-up scenarios) with the difference that this time participants will be asked to take a personal decision under time pressure choosing between two different and opposing options that they have. For example, participants will be given cases resembling real-life professional dilemmas, like: "You’re informed from a rather trusted source that EU Commission will oblige all the EU state-members to downsize their tax-laws by 40% till the end of the year. Source tells you that your main competitor media agency has the same information and decided to publish it right away, aiming to news exclusivity (“be the first”). What would you do: would you fact-check the information (losing valuable time and not be the first to publish it) or would you take the risk and publish it no matter if it is accurate or false?”

The reasoning of decisions taken in dilemmatic circumstances can reveal from the one hand possible dominant professional mentalities and from the other, existing professional value systems.

2.3.8 In-depth Individual Interviews

As a data collection method, individual interviews are well suited for research topics that are complex, not well understood yet, and merit further exploration. Interviews are conducted on the premise of a set of assumptions and understanding about a specific situation, which is not usually connected with casual conversation (Silverman, 1985). Given the complexity of the issue of misinformation and the varying practices of each of the stakeholder groups (citizens, journalists/fact-checkers, policymakers), in-depth interviews are an ideal data collection method because they can provide data that are based on opinions, feelings, emotions and experiences, that can be explored in depth, in a one-to-one setting. This also makes it a suitable methodological tool for exploring sensitive and personal issues in an open and honest matter.

After the conclusion of the co-creation workshops, we believe there will still be ground to explore the above understanding in a more personalized and detailed way, but more than this, the specific attitudes towards misinformation, credibility evaluation practices as well as stakeholders’ perceived contributions to misinformation. The interviewees will be recruited from the co-creation workshop participants and should represent all three stakeholder groups (citizens, journalists/fact-checkers, or policymakers), if possible, in equal numbers. We would like to have at least two to three interviews from each stakeholder group — more if possible.

It should be emphasized that any information gathered during the interviews will be used anonymously and the identity of the interviewee will not be disclosed.

Our aim is to probe deeper into the stakeholders' views and needs, and to identify the challenges in handling misinformation that could be addressed with the Co-Inform app. Compared to the focus groups, this process aims to go deeper and provide answers in a more detailed way.

2.4 Collecting data from the Co-Creation Workshops

2.4.1 Co-creation Workshop 1

The aim of the first co-creation workshop is to gather the needs and recommendations from each stakeholder group, in order to understand where the issue lies. We begin the first pilot with the assumption that each of the stakeholder groups - Citizens, Journalists/Fact-checkers, Policymakers - may have varying understanding of misinformation, its impact and the risks and challenges that are posed. Our aim is to assess stakeholders' views and needs, and to identify the challenges in handling misinformation that could be addressed with the Co-Inform app. Specifically, we seek to understand the views of the stakeholder group, relating to how they perceive misinformation, their daily practices on social media platforms, and the challenges they face in identifying or countering misinformation.

During the first co-creation workshop each co-creation site will conduct three separate one-hour focus groups (one for each stakeholder group) to gauge the individual needs, perceptions and challenges of each group. A Data Collection Framework for each of the stakeholder groups will be provided by WP5, along with a focus group protocol, that should be used to facilitate consistent data collection across each site.

In addition to this, during the first co-creation workshop, we will conduct a survey based on the EU's Flash Eurobarometer questionnaire 464, which was designed to explore EU citizens' awareness of and attitudes towards the existence of fake news and disinformation online. Using this questionnaire will enable us to compare data from our co-creation participants to the general EU population and to the participating countries, according to the published findings of the Eurobarometer questionnaire.

As stated in D1.2 the following data will be collected from the first co-creation workshop:

1. Background information about the stakeholders
2. Views on misinformation (attitudes, impact, trust, ability to recognize it, responsibility):
    - Their level of trust on news sources through different channels
    - Their perception of misinformation, frequency of encountering such news
    - Confidence on identifying misinformation
    - Practices on handling misinformation, and especially how they assess the credibility and validity of information
    - Views on extent of the problem and impact on their countries, trust and democracy

- Views on which institutions should act to combat the problem

3. Input on policies (existing, or suggestions).
4. Recommendations on tools and features they would like to be implemented in these tools.
5.Evaluation of the workshop:
   - Evaluation of stakeholders' mapping.
   - Evaluation of methods and exercises as well as stakeholders' feedback about the workshop.
   - Evaluation of responses and data provided by participants.

In the table below you can find the aims of the first workshop, as stated in D1.2, and how we plan to collect the relevant data.

Table 2. *Data collection during Co-creation workshop 1*

Aim	Method of Data Collection
Background information about the stakeholders and their views on misinformation:	Number and demographics of participants, participant profiles and how they were selected Survey questions 1-9
Level of trust on news sources through different channels	Survey questions 10.1-10.6 Focus Group: Part IV (Practices)
Perceptions of misinformation, frequency of encountering such news	Survey question 11 Focus Group: Part IV (Perceptions)
Confidence in identifying misinformation	Survey question 12 Focus Group: Part IV (Practices)
Practices on handling misinformation, and especially how they assess the credibility and validity of information	Focus Group: Part IV (Practices) Co-creation workshop activities
Views on extent of the problem and impact on their countries, trust and democracy	Survey question 13-14 Focus Group: Part III (Perceptions)
Views on which institutions should act to combat the problem	Survey question 14
Input on policies (existing, or suggestions)	Focus Group: Part VI (Challenges)

Recommendations on tools and features they would like to be implemented in these tools.	Focus Group: Part V (Challenges) Co-Creation Workshop Activities
Evaluation of stakeholders' mapping.	WP1 is responsible for this. WP1 should report back to WP5 about this.
Evaluation of methods and exercises as well as stakeholders' feedback about the workshop.	WP1 is responsible for this --a brief evaluation survey should be shared at the end of the workshop.
Evaluation of responses and data provided by participants.	WP5 is responsible for this – a brief evaluation survey should be shared at the end of the workshop.

The first co-creation workshop is important in setting up the baseline, i.e. understanding the needs and informing the policies that will govern the technological tools that will be developed within the context of the Co-Inform project. For this reason, it is important to understand the complex minutiae of the stakeholders' practices, professional routines and challenges in order to gain an accurate and holistic view of the problem. In-depth individual interviews will ensue each pilot workshop, to facilitate this.

The interviews will take place after the focus groups and the co-creation workshops have been concluded. Before the participants leave the place, researchers are advised to approach them and ask if they would be willing to give an in-depth research personal interview and share their individual views and experiences. If asked, the co-creation coordinators can explain that this is a follow-up activity that will give us the opportunity to gain a better understanding of their answers during the workshop. The interviewees will be recruited from the co-creation workshop participants and should represent all three stakeholder groups (citizens, journalists, or policy makers), in equal numbers if possible. We would like to have at least two to three interviews from each stakeholder group — more if possible. It should be emphasized that any information gathered during the interviews will be used anonymously and the identity of the interviewee will not be disclosed.

Indicative methodologies: Focus Group, Questionnaire, Interviews, Cognitive Walkthroughs, Behaviour Observation.

2.4.2 Co-creation Workshop 2

The second workshop will allow participants from the three stakeholder groups to provide feedback on prototypes suggested by WP3 or/and WP4. Feedback on this early version of the technological solution will also inform WP2 about the policies that could underpin the technological tool that will be developed. Within this workshop, participants may also co-create some parts of the tools, such as the user interface and will also give feedback on strategies that seem most promising in raising awareness and addressing misinformation.

Questionnaires, story-boarding techniques, and individual interviews complete and fill out the collection of methods that will be employed. These methods will help us comprehend users' experience with the nudging interventions and also explore in-depth what each stakeholder group makes of these tools.

Indicative methodologies: Cognitive walkthroughs, Think Aloud, Eye tracking, Behaviour Observation, Contextual Inquiry Approach, Interviews.

2.4.3 Co-creation Workshop 3

During the third workshop the stakeholders will respond to a functioning tool that was developed based on suggestions made in the previous workshops. The focus will be on evaluating the functionality and performance of the technological tool, by taking into consideration each group's stated needs, as these have been defined following the first workshop. Information gathered will inform an improved version of the tool, which will be subsequently evaluated anew.

Indicative methodologies: Cognitive walkthroughs, Think Aloud, Eye tracking, Behaviour Observation, Contextual Inquiry Approach, Interviews.

3. Decision analytical model

Decision making (based on public information) can be affected by misinformation or sometimes even irrational factors and lacks transparent support models for the preparatory, analysis and negotiation stages of democratic decision processes. Generally, policy decisions are to a large extent influenced by temporary hot spots and trends in unstructured data trying to grasp attitudes to societal issues. Therefore, Task T5.4 aims at developing methodologies and tools for supporting complex transparent decision processes on misinformation based on data sets of various kinds. The basic ideas behind the processes are: (i) they must take advantage of transparent decision support models, (ii) the various beliefs and opinions involved must be clearly separated from the actual underlying facts, and (iii) reasonably fair and efficient elicitation procedures must be included.

In the context of the Co-Inform project we can compare policies to deal with misinformation in the area of migration against a set of evaluation criteria and performance indicators. Each policy is evaluated against a set of criteria, which are to be developed in co-creation with stakeholders. The criteria are developed based on the review of scientific literature on migration, as well as on the analysis of policy documents from pilots.

The criteria should include a set of indicators, usually quantitative and qualitative. Data for quantitative indicators can be collected, e.g., from national and international statistical databases, reports and projects. Data for qualitative indicators were collected from surveys, questionnaires, interviews with stakeholders.

Further on, relevant criteria are selected and discussed during the co-creation workshops to see whether the stakeholders agree with the criteria definition, whether the criteria are relevant for them and the pilots and whether stakeholders would recommend any further criteria. A more detailed explanation of the decision analytical model in respect to the Co-Inform goals is provided in the Appendix of this deliverable.

4. Challenges and Risks

A possible challenge that might arise from the research inspection of the tools is the inherent bias due to improper task selection and implementation. For this reason, multiple scenarios will be developed by the researchers to make sure different possibilities are covered. More specific challenges and risks are provided below, along with contingency plans.

- **Consistency across the three co-creation sites:** The three different locations of the co-creation workshops certainly enrich the Co-Inform data and the consortium's understanding of the three stakeholder groups; however, it also adds a challenge for the data collection processes since each location has a different social, political and cultural context. In order to address this, WP5 will provide each research team in charge of each of the co-creation workshops with detailed data collection frameworks for each of the workshops, in order to ensure consistency across the data collection methods for all three co-creation sites. The frameworks will be released in advance, so as to enable a discussion and feedback from participating institutions, taking into account local circumstances that might arise. The WP5 data collection frameworks will provide detailed and thorough instructions that must be followed by all co-creation sites, in order to ensure consistent data collection that will lead to useful insights and feedback.

- **Distribution of subgroups may be varying across the different workshops:** Each site will use similar recruitment methods, but since participation is optional, this may lead to some imbalances in terms of each of the stakeholder groups across each co-creation site. WP5 will be in close communication with WP1 and each of the Austrian, Greek and Swedish teams. In order to moderate the attendance and engagement of each stakeholder group, across each workshop, WP5 will provide relevant reporting forms to be filled out by each co-creation site team, to enable a close monitoring of participants' turnout.

- **Privacy issues:** Participation in the study is optional, and participants will have the option to select what methods of data collection they consent to. We anticipate that there may be instances where participants decline to be video or audio recorded, and it is therefore important to inform participants about the data collection methods in advance of the workshop, to ensure that the proposed data collection procedures are followed. This will also enable the WP5 team to analyze and evaluate the data, in order to provide effective feedback to the relevant technical partners.

- **Language barrier:** Each co-creation site will take place in three different locations, and therefore the language of the communication will vary. As such, it is the responsibility of WP5 to provide data collection frameworks in a manner that allows for Austria, Greece and Sweden to translate any necessary material prior to the workshops. The same challenge exists post-workshop, since all data collected via video, audio, photographs, etc., will need to be translated to English by each of the co-creation research teams, to facilitate sharing among all partners and also allow for WP5 to evaluate and assess the data. All audio / video recordings need to be transcribed and translated verbatim.

- **Varying experience in proposed methodologies**: Each research team will have different experience in data collection procedures, and as such might not always be familiar, or have the relevant experience, with the data collection methodologies proposed by WP5. However, the complexity of observing, monitoring and analyzing the behaviours, attitudes and practices of three very different stakeholder groups prescribes that we use a range of methodologies that will result in rich data that can enable WP5 to provide useful and effective feedback to its relevant counterparts. Therefore, prior to each co-creation workshop, WP5 will be responsible for providing guidance via video conferencing preparatory meetings to the Austrian, Greek and Swedish teams.

- **Data loss**: Issues relating to data safety are provided in detail in D1.1.

5. Roles of partners

Each pilot team should appoint a contact person who will be the point of reference for the specific team and WP5. The role of the contact person is to ensure prompt communication with WP5 and to coordinate the data collection procedures and feedback within the relevant team.

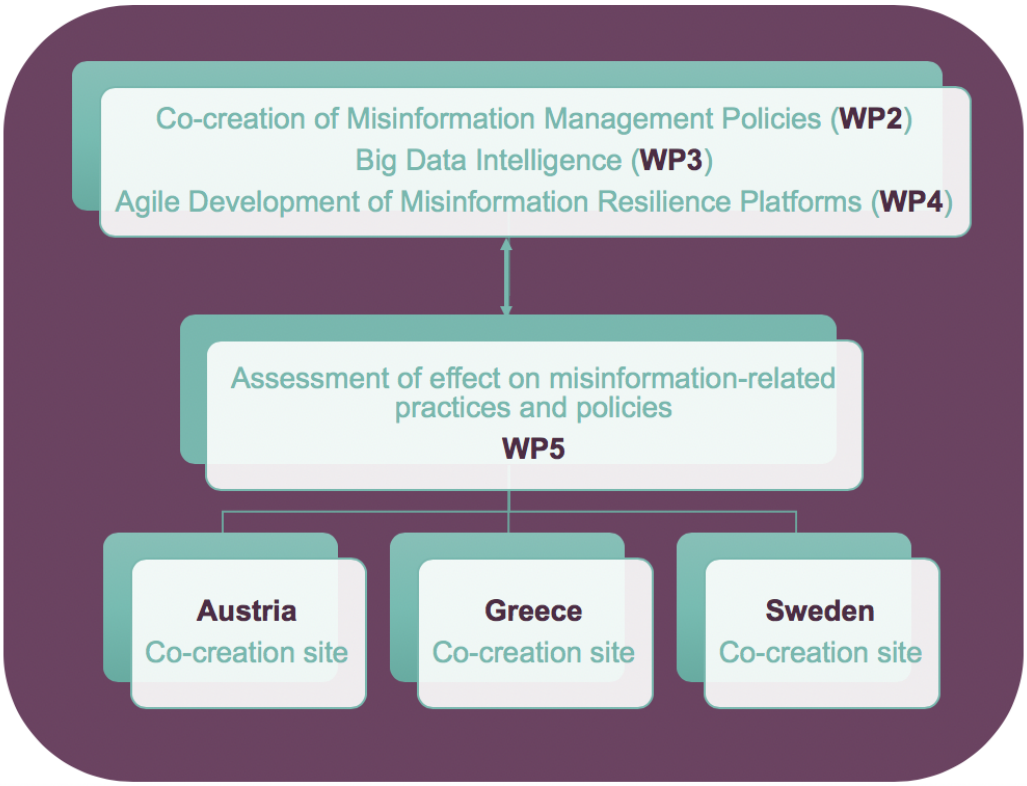

Figure 3. WP5 role and synergies within the Co-Inform project

5.1 Stockholm University (SU)

- Provide detailed description of the WP1 co-creation workshops ahead of time so that the data collection process for WP5 can be situated and connected to the WP1 activities.
- Coordinate with WP5 regarding data collection in connection to WP1 workshops.
- Provide information on the co-creation workshop activities, agenda and participants
- Translate relevant data collection instruments provided by WP5 prior to the workshop to the local language, for example, focus group script or stimulus material
- Administer instruments and collect data as suggested in the evaluation frameworks shared and discussed prior to each of the co-creation workshop.
- Transcribe audio / video data verbatim and translate all verbatim transcripts to English, to allow for evaluation by WP5. Any photographic material will also need translation, where language is depicted.
- Distribute/disseminate instruments and collect data for evaluating the Co-Inform app.

5.2 Cyprus University of Technology

- WP5 lead partner
- Coordinate with WP1 regarding the data collection needs.
- Draft data collection frameworks for each co-creation workshop, while taking into account feedback by co-creation site teams in Austria, Greece and Sweden and the needs of WP2, WP3, WP4.
- Lead guidance and provide virtual assistance for any proposed methodologies to co-creation workshop pilots.
- Provide standardized reporting forms to facilitate consistent data reporting across each site.
- Evaluate, analyse and disseminate insights from data collected at workshops to relevant partners.

5.3 Open University (OU)

- As the leader of WP3, coordinate with WP1 and WP5 to provide feedback on the behaviour analysis needs for each stakeholder, in order to facilitate the development of appropriate evaluation instruments for each workshop.
- Bridge the design of analytic tools in WP3 with WP5 and use evaluation data to revise WP3 algorithms and services.
- Distribute/disseminate instruments and collect data for evaluating the Co-Inform app.

5.4 University of Koblenz and Landau (UKOB)

- As the leader of WP2, coordinate with WP1 and WP5 to provide needs and feedback on misinformation management policies.
- Revise policies based on WP1 and WP5 data.
- Distribute/disseminate instruments and collect data for evaluating the Co-Inform app.

5.5 FCNI (Northern Ireland)

- Provide input regarding the journalist/fact-checking evaluation tools.
- Distribute/disseminate instruments and collect data for evaluating the Co-Inform app.

5.6 ESI

- Provide input regarding the collected data and WP4.
- Distribute/disseminate instruments and collect data for evaluating the Co-Inform app.

5.7 International Hellenic University (IHU)

- Provide information on the co-creation workshop activities, agenda and participants
- Translate relevant data collection instruments provided by WP5 prior to the workshop to the local language (i.e. focus group script or stimulus material)
- Administer instruments, collect data as suggested in the evaluation frameworks shared and discussed prior to each of the co-creation workshop.

- Transcribe audio / video data verbatim and translate all verbatim transcripts to English, to allow for evaluation by WP5. Any photographic material will also need translation, where language is depicted.
- Distribute/disseminate instruments and collect data for evaluating the Co-Inform app.

5.8 IIASA

- Provide information on the co-creation workshop activities, agenda and participants
- Translate relevant data collection instruments provided by WP5 prior to the workshop to the local language (i.e. focus group script or stimulus material)
- Administer instrument, collect data as suggested in the evaluation frameworks shared and discussed prior to each of the co-creation workshop.
- Transcribe audio / video data verbatim and translate all verbatim transcripts to English, to allow for evaluation by WP5. Any photographic material will also need translation, where language is depicted.
- Distribute/disseminate instruments and collect data for evaluating the Co-Inform app.
- Adapt a decision-making tool, based on data from WP1, WP2, WP3, and WP4, to support policymakers' evaluation of options in connection to the Co-Inform objectives.

5.9. SCYTL

- As the leader of WP4, coordinate with WP1 and WP5 to provide feedback on the behaviour analysis needs for each stakeholder, in order to facilitate the development of appropriate evaluation instruments for each workshop.

6. Ethics & Privacy

The Co-Inform consortium respects and values the participants' privacy. All procedures relating to the handling of personal and sensitive data are detailed in the Co-Inform Project Handbook and Quality Assurance Plan. No type of data will be collected unless the participants have been informed and have provided their written consent to participate. Evaluating the data provided in each of the workshops, also means dealing with sensitive and private information, that Co-Inform project participants have consented to share. All relevant precautions will be adhered to when dealing with personal data and all data will be reported anonymously. WP1 is responsible for data storage, collection and processing and D1.1 (Section 6.3) and D1.2 (Section 6) addresses these issues in detail.

7. References

Alemanno, A. (2018). How to Counter Fake News? A Taxonomy of Anti-fake News Approaches. *European Journal of Risk Regulation, 9*(1), 1-5. doi:10.1017/err.2018.12

Alevizos P., Deisi W., Liberman R., Eckman T., Callahan E. (1978). The Behavior Observation Instrument: A Method of Direct Observation for Program Evaluation, *Journal of Applied Behavior Analysis*, 1130 No.2, pp: 243-257, 1978

Allendoerfer K., Aluker S., Panjwani G., Proctor J., Sturtz D., VucovicM., Chen C. (2005). Adapting the Cognitive Walkthrough Method to Assess the Usability of a Knowledge Domain Visualization, EEE Symposium on Information Visualization 2005 October 23-25, Minneapolis, MN, USA, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1532147

Alshammari T., Alhadreti O., Mayhew J.P. (2015). When to Ask Participants to Think Aloud: A Comparative Study of Concurrent and Retrospective Think-Aloud Methods, *International Journal of Human-Computer Interaction (IJHCI)*, Volume (6), Issue (3), pp: 48-64, 2015

Altmann, J. (1974). Observational Study of Behavior: Sampling Methods. *Behaviour*, Vol. 49, No. 3/4, pp:227-267, Brill, 1974

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. Journal of Personality and Social Psychology, 43, 997-1013.

Antoniadis, S., Litou, I., & Kalogeraki, V. (2015, October). A model for identifying misinformation in online social networks. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 473-482). Springer, Cham.

Arant, M. D., & Meyer, P. (1998). Public and traditional journalism: a shift in values? *Journal of Mass Media Ethics*, 13(4), 205-218.

Baer, J., & McKool, S. S. (2009). Assessing creativity using the consensual assessment technique. In *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 65-77). IGI Global.

Badke, W. (2018). Fake News, Confirmation Bias, the Search for Truth, and the Theology Student. *Theological Librarianship*, *11*(2), 4–7. https://doi.org/10.31046/tl.v11i2.519

Bastos, M., & Mercea, D. (2018). The public accountability of social platforms: lessons from a study on bots and trolls in the Brexit campaign. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20180003.

Borden, S. L., & Tew, C. (2007). The role of journalist and the performance of journalism: Ethical lessons from “fake” news (seriously). *Journal of Mass Media Ethics*, 22(4), 300-314.

Boren T.M., Ramey J. (2000). Thinking Aloud: Reconciling Theory and Practice, *IEEE Transactions on Professional Communication*, Vol. 43, No.3, pp: 261-278, Sept. 2000

Callister Jr, T. A. (2000). Media literacy: On-ramp to the literacy of the 21st century or cul-de-sac on the information superhighway. *Advances in Reading/Language Research*, *7*, 403-420.

Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 675-684). ACM.

Charters E. (2003). The Use of Think-aloud Methods in Qualitative Research. An Introduction to Think-aloud Methods. *Brock Education* Vol. 12, No. 2, pp: 68-82, 2003

Deuze, M., & Yeshua, D. (2001). Online journalists face new ethical dilemmas: Lessons from the Netherlands. *Journal of Mass Media Ethics*, 16(4), 273-292.

Eccles D.W. & Arsal G. (2017). The think aloud method: what is it and how do I use it?, Qualitative Research in Sport, Exercise and Health, 9:4, 514-531, https://doi.org/10.1080/2159676X.2017.1331501

Elliott, D. (2008). Essential shared values and 21st century journalism. *The Handbook of Mass Media Ethics* (pp. 42-53). Routledge.

Ericsson K.A. & Simon, H.A. (1993). *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: MIT Press, 1993

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, *58*(2), 203.

Flanagin, A. J., & Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society*, *9*(2), 319-342.

Friend, C., & Singer, J. (2007). *Online Journalism Ethics: Traditions and Transitions. 1.*Armonk, NY: ME Sharpe.

García-Avilés, J. A. (2014). Online newsrooms as communities of practice: Exploring digital journalists' applied ethics. *Journal of Mass Media Ethics*, 29(4), 258-272.

Glenski, M., Weninger, T., & Volkova, S. (2018). Propagation from Deceptive News Sources Who Shares, How Much, How Evenly, and How Quickly? *IEEE Transactions on Computational Social Systems*, 5(4), 1071–1082. https://doi.org/10.1109/TCSS.2018.2881071

Gunther, A. C. (1992). Biased press or biased public? Attitudes toward media coverage of social groups. *Public Opinion Quarterly*, *56*(2), 147-167.

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013, May). Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 729-736). ACM.

Gupta, M., Zhao, P., & Han, J. (2012, April). Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining* (pp. 153-164). Society for Industrial and Applied Mathematics.

Güss, C.D. (2018). What is going on through your mind? Thinking-Aloud as a method in cross-cultural psychology, *Frontiers in Psychology*, 2018; 9: 1292,https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6099082/

Heath, T., Motta, E., & Petre, M. (2006). Person to person trust factors in word of mouth recommendation. Retrieved from: https://oro.open.ac.uk/23640/1/heath-motta-petre-reinvent2006-person-to-person-trust-factors.pdf

Johnstone C.J., Bottsford-Miller N.A., Thompson S.J. (2006). Using the Think Aloud Method (Cognitive Labs) To Evaluate Test Design for Students with Disabilities and English Language Learners, *National Center for Educational Outcomes*, 2006, https://files.eric.ed.gov/fulltext/ED495909.pdf

Kelton, K., Fleischmann, K. R., & Wallace, W. A. (2008). Trust in digital information. *Journal of the American Society for Information Science and Technology*, 59(3), 363-374.

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., & Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S), 1220.

Lee, J. & Seo, D. (2015). Crowdsourcing not all sourced by the crowd: An observation on the behavior of Wikipedia participants, *Technovation*55-56, pp: 14–21. Elsevier.

Lewis, C. H. (1982). *Using the "Thinking Aloud" Method In Cognitive Interface Design* (Technical report) IBM RC-9265.

Madriz, E. (2000). Focus groups in feminist research, in Denzin N.K., Lincoln Y.S., (Eds.), *Handbook of qualitative research* (2nd ed.), pp: 835–850, Sage Publications.

Mahatody T., Sagar M., Kolski C. (2010). State of the Art in Cognitive Walkthrough Method, Its Variants and Evolutions. *International Journal of Human-Computer Interaction*, 26 (8), pp: 741-785.

Masood M. & Thigambaramb, M. (2015). The Usability of Mobile Applications for Pre-schoolers, *Procedia - Social and Behavioral Sciences 197*, pp: 1818–1826.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of communication*, *60*(3), 413-439.

Morgan, D.L. (1998). *The focus group guidebook*. Thousand Oaks, CA: Sage.

McBride, K., & Rosenstiel, T. (Eds.). (2013). *The new ethics of journalism: Principles for the 21st century*. CQ Press.

Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012, February). Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 441-450).

Nawaz A., Skjæret N., Lægdheim Helbostad L., Vereijken B., Boulton E., & Svanaes D.(2015). Usability and acceptability of balance exergames in older adults: A scoping review, *Health Informatics Journal 2016*, Vol. 22(4) pp: 911–931, Sage Publications.

Nielsen, R. K., & Graves, L. (2017). News you don't believe: audience perspectives on fake news. *Reuters Institute for the Study of Journalism, Oxford*.

Onwuegbuzie, A.J., Slate, J.R., Leech, N.L., & Collins, K.M.T. (2009). Mixed data analysis: Advanced integration techniques, *International Journal of Multiple Research Approaches*, 3, pp:13–33.

Pang, N., & Ng, J. (2017). Misinformation in a riot: a two-step flow view. *Online Information Review*, *41*(4), 438-453.

Rath, B., Gao, W., Ma, J., & Srivastava, J. (2017, July). From retweet to believability: Utilizing trust to identify rumor spreaders on Twitter. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 179-186). ACM.

Silverman, D. (1985) *Qualitative Methodology and Sociology*. Aldershot: Gower.

Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy*, *2*(1), 85–95. https://doi.org/10.1353/bsp.2016.0009

Shen, C., Kasra, M., Pan, W., Bassett, G. A., Malloch, Y., & O'Brien, J. F. (2019). Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New Media & Society*, *21*(2), 438-463.

Simons, G. (2017). Fake News: As the Problem or a Symptom of a Deeper Problem? *Media Lens*, 33–44.

Snyder J., Reid J., Stoolmiller, M., Howe, M., Brown, H., Dagne, G., & Cross, W. (2006). The Role of Behavior Observation in Measurement Systems for Randomized Prevention Trials. *Prevention Science*, Vol. 7, No. 1, pp: 43-56.

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, 34(3), 150–160. https://doi.org/10.1177/0266382117722446

Suen, K.H. & Ary D. (1989). *Analyzing Quantitative Behavioral Observation Data*. Psychology Press, Taylor & Francis Group, New York – London.

Sülflow, M., Schäfer, S., & Winter, S. (2019). Selective attention in the news feed: An eye-tracking study on the perception and selection of political news posts on Facebook. *New Media & Society*, *21*(1), 168-190.

Tching, J., Reis, J. & Paio, A. (2016). A Cognitive Walkthrough towards an Interface Model for Shape Grammar Implementations, *Computer Science and Information Technology* 4 (3), pp: 92-119.

Thaler, R. H., & Sunstein, C. (2013). Nudge: improving decisions about health, wealth, and happiness. *Choice Reviews Online*, *46*(02), 46-0977-46-0977. https://doi.org/10.5860/choice.46-0977

Theng, Y. L., Goh, L. Y. Q., Lwin, M. O., & Shou-Boon, S. F. (2013, September). Dispelling Myths and Misinformation Using Social Media: A Three-Countries Comparison Using the Case of Tuberculosis. In *2013 IEEE International Conference on Healthcare Informatics* (pp. 147-152). IEEE.

Traynor, B., Hodson, J., & Wilkes, G. (2016, July). Media selection: A method for understanding user choices among popular social media platforms. In *International Conference on HCI in Business, Government, and Organizations* (pp. 106-117). Springer, Cham.

van Strien, J. L., Kammerer, Y., Brand-Gruwel, S., & Boshuizen, H. P. (2016). How attitude strength biases information processing and evaluation on the web. *Computers in Human Behavior*, *60*, 245-252.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.

Venkatesh, V., Thong, J. Y., & Xu, X. (2016). Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the Association for Information Systems*, *17*(5), 328-376.

Vosoughi, S., Roy, D., & Aral, S. (2018). News Online, *1151*(March), 1146–1151. https://doi.org/10.1126/science.aap9559

Ward, S. J. (2005). Philosophical foundations for global journalism ethics. *Journal of Mass Media Ethics*, 20(1), 3-21.

Weeks, B. E. (2015). Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation. *Journal of Communication*, *65*(4), 699–719.https://doi.org/10.1111/jcom.12164

Wilkinson, S. (2004). Focus group research. In D. Silverman (Ed.), *Qualitative research: Theory, method, and practice*, pp: 177–199, Thousand Oaks, CA: Sage.

Wineburg, S., McGrew, S., Breakstone, J., & Ortega, T. (2016). Evaluating information: The cornerstone of civic online reasoning. *Stanford Digital Repository. Retrieved January*, *8*, 2018.

Yang J., Yuanyuan Q., Zhang X., Haiyang H., Fang L., & Cheng G. (2015). Characterizing User Behavior in Mobile Internet, *Emerging Topics in Computing*, IEEE Transactions.

APPENDIX

Overall Workshop Methodology and Theory Background

Dr. Nadejda Komendantova, research scholar at IIASA and a senior research scholar at ETH Zurich
Prof. Love Ekenberg, senior research scholar at IIASA and a professor in computer and systems science at Stockholm University

The strategy framework is a decision analytical approach to co-creation in a multi-stakeholder/multi-criteria environment, supported by elaborated decision analytical tools and processes:

- a framework for elicitation of stakeholder preferences
- a decision engine for strategy evaluation
- a machinery for risk analysis
- a set of processes for negotiation
- a set of decision rule mechanisms
- processes for combining these items
- various types of implementations of the above

These components apply to decision components, such as:

- agenda settings and overall processes
- stakeholders
- goals
- strategies/policies/sub-strategies/part-policies, etc
- consequences/effects
- qualifications and sometimes quantifications of the components
- negotiation protocols
- decision rules and processes

The process is progressing during four workshops where a set of criteria is thus developed as well as a set of strategy options. The latter will then be benchmarked against a set of evaluation criteria and performance indicators. The problem generated is thus a multi-stakeholder multi-criteria problem under uncertainty.

Multi-criteria decision analysis (MCDA)

1. Introduction

Deployment of changed socio-economical conditions must lead to transitions and transformations of entire sectors [28]. Such transitions are complex processes, which has political, social, economic and technical dimensions and involve a multitude of stake-holders. Therefore, a holistic, inclusive and comprehensive governance approach to such is essential, since unguided significant socio-technical will lead to many frictions and conflicts. Such changes will thus lead to a socio-technological transition processes, which are combined with and emphazised by shifts in technologies, business models, governance structures,

consumption patterns, values and worldviews. Thus, such multi-stakeholder, multi-criteria situations are typical for the planning and decision processes involved herein and a significant issue is of course what methodologies to use.

A multitude of methods for analysing and solving decision problems with multiple criteria have been suggested during the last decades. A common approach is to make preference assessments by specifying a set of attributes that represents the relevant aspects of the possible outcomes of a decision. Value functions are then defined over the alternatives for each attribute and a weight function is defined over the attribute set. One option is to simply define a weight function by fixed numbers on a normalised scale and then define value functions over the alternatives, where these are mapped onto fixed values as well, after which these values are aggregated and the overall score of each alternative is calculated.

One of the problems with the additive model as well as other standard multiple criteria models is that numerically precise information is seldom available, and most decision-makers experience difficulties with entering realistic information when analysing decision problems, and with the elicitation of exact weights, that demands an unreasonably exactness which does not exist. There are other problems, such as that ratio weight procedures are difficult to accurately employ due to response errors. The common lack of reasonably complete information increases this problem significantly. Several attempts have been made to resolve this issue. Methods allowing for less demanding ways of ordering the criteria, such as ordinal rankings or interval approaches for determining criteria weights and values of alternatives, have been suggested, but the evaluation of these models is sometimes quite complicated and difficult for decision makers to accept.

Some main categories of approaches to remedy the precision problem are based on capacities, sets of probability measures, upper and lower probabilities, interval probabilities (and sometimes utilities), evidence and possibility theories, as well as fuzzy measures. The latter category seems to be used only to a limited extent in real-life decision analyses since it usually requires a significant mathematical background on the part of the decision-maker. Another reason is that the computational complexity can be problematic if the fuzzy aggregation mechanisms are not significantly simplified.

For the evaluations herein, we will therefore utilise a method and software for integrated multi-attribute evaluation under risk, subject to incomplete or imperfect information. The software originates from our earlier work on evaluating decision situations using imprecise utilities, probabilities and weights, as well as qualitative estimates between these components derived from convex sets of weight, utility and probability measures. To avoid some aggregation problems when handling set membership functions and similar, we introduce higher-order distributions for better discrimination between the possible outcomes. For the decision structure, we use the common tree formalism but refrain from using precise numbers. To alleviate the problem of overlapping results, we suggest a new evaluation method based on a resulting belief mass over the output intervals, but without trying to introduce further complicating aspects into the decision situation. During the process, we consider the entire range of values as the alternatives presented across all criteria as well how plausible it is that an alternative outranked the remaining ones, and thus provided a robustness measure. Because of the complexity in these calculations, we use the state-of-the-art multi-criteria softwares DecideIT or Decision Wizard for the analysis, which allows for imprecision of the

kinds that exist here [29]. Versions of DecideIT have been successfully used in a variety of decision situations, such as larch-scale energy planning [30], allocation planning [31], demining [32], financial risks [33], gold mining [35] and many others.

Figure 1 shows of the multi-criteria multi-stakeholder tool Decision Wizard, developed for group decisions regarding infrastructure policy making [31] in Swedish municipalities, using the CAR method of [11].

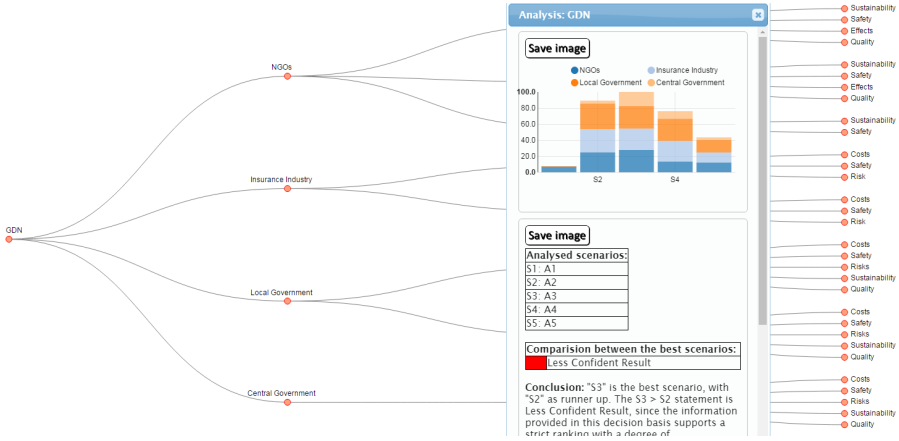

Figure 1. The Group Decision tool Decision Wizard – a simplification of DecideIT

2. Decision Modelling

Typically, a multi-criteria decision situation is modelled like a tree, such in the figure below, where the *w:s* are criteria weights and the *v:s* are values of alternatives under the different criteria.

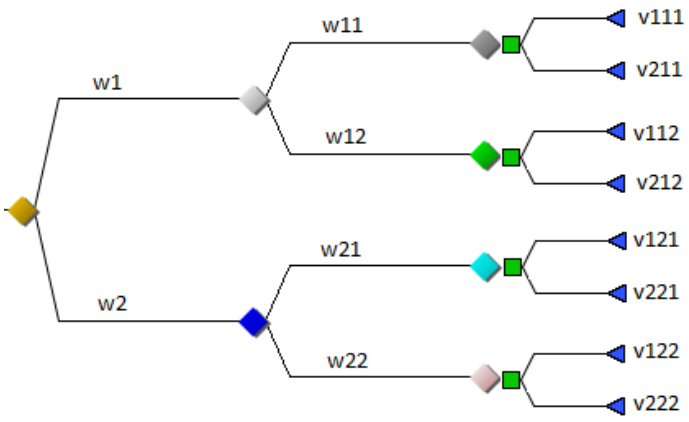

Figure 2. A multi-criteria decision tree.

The normalisation constraint means that the weights are restricted by a the equation $\sum w_j = 1$, where w_j denotes the weight of a criterion G_j and the weight of sub-criterion G_{jk} is denoted by w_{jk}. Denote the value of alternative a_i under sub-criterion G_{jk} by v_{ijk}.

A common value function for evaluating alternatives in the analyses is a weighted average of the components involved. For instance, consider an alternative A_i under two criteria, with the respective weights w_1 and w_2. The overall value of this alternative can be calculated by a weighted average:

$E(A_i) = \sum_{j=1}^{2} w_j \sum_{k=1}^{2} w_{jk}\, v_{ijk}$
Co-InformCo-Inform

This can straightforwardly be generalized to multi-criteria decision trees of arbitrary depth and solved as corresponding multi-linear equations.

One of the problems with most models for criteria ranking is that numerically precise information is seldom available. We have solved this in part by introducing surrogate weights as before. This, however, is only a part of the solution since the elicitation can still be uncertain and the surrogate weights might not be a fully adequate representation of the preferences involved, which of course, is a risk with all kinds of aggregations. To allow for analyses of how robust the problem is to changes of the input data, we will also introduce intervals around the surrogate weights as well as around the values of the options. Thus, in this elicitation problem, the possibly incomplete information is handled by allowing the use of intervals (cf., e.g., [26]), where ranges of possible values are represented by intervals (in combination with pure orderings without the use of surrogate weights at all, if the latter turns out to be inadequate).

There are thus several approaches to elicitation in MCDA problems, and one partitioning of the methods into categories is how they handle imprecision in weights and values, such as fixed numbers, comparative statements, representing orderings or intervals.

Computationally, methods using fixed numbers are very easy to solve, while systems of relational or interval constraints normally require more elaborated optimization techniques. On the other hand, if the model only accepts fixed numbers, we impose constraints that might severely affect the decision quality. If we allow for imprecision in terms of intervals and relations, we usually get a more realistic representation of the problem. These can, for instance, be represented by interval statements, such as $w_i \in [y_i - a_i, y_i + b_i]$, where $0 < a_i \leq 1$ and $0 < b_i, \leq 1$, or comparative statements, such as $w_i \geq w_j$.

Systems of such equations can be solved, and aggregations of decision components in these formats can be optimized, by using the methods from [27]. The disadvantage here is that many decision-makers sometimes perceive these methods difficult to understand and accept, because of complex computations and loss of user transparency.[1]

In this case, the performance of the different alternative options will be estimated. Together with the surrogate weights, they thus provide the decision base for the multi-criteria analysis. Using the weighted aggregation principle, we will combine the multiple criteria and stakeholder preferences with the valuation of the different options under the criteria surrogate weights. This will be further described below.

[1] This should be kept in mind here as always when working with aggregation methods of whatever kind and this should affect how the elicitation mechanisms and software tools that are used.

The results of the process will be (i) a detailed analysis of each option's performance compared with the others, and (ii) a sensitivity analysis to test the robustness of the result.

During the process, we consider the entire range of values as the alternatives presented across all criteria as well how plausible it is that an alternative will outrank the remaining ones, and thus provide a robustness measure.

3. Criteria Ranking

One of the problems with the additive model as well as other Multi-criteria decision aid (MCDA) models is that numerically precise information is seldom available, and most decision-makers experience difficulties with entering realistic information when analysing decision problems. For instance, Barron and Barrett [1] argue that the elicitation of exact weights demands an unreasonably exactness which does not exist. There are other problems, such as that ratio weight procedures are difficult to accurately employ due to response errors [2]. The common lack of reasonably complete information increases this problem significantly. Several attempts have been made to resolve this issue. Methods allowing for less demanding ways of ordering the criteria, such as ordinal rankings or interval approaches for determining criteria weights and values of alternatives, have been suggested, but the evaluation of these models is sometimes quite complicated and difficult for decision makers to accept.

The utilisation of ordinal or imprecise importance information to determine criteria weights is a way of handling this and some authors have suggested surrogate weights as representative numbers assumed to represent the most likely interpretation of the preferences expressed by a decision-maker or a group of decision-makers. The idea is to enable decision-makers to utilise the information they are able to supply and then generate representative weights from some underlying distribution and investigate how well they perform. One such type is derived from ordinal importance information [3,1,4], where decision-makers supplies ordinal information on importance and the information is the subsequently converted into surrogate weights corresponding to and consistent with the extracted ordinal information. Often used such are rank sum (RS) weights, rank reciprocal (RR) weights [5], and centroid (ROC) weights [6]. ROC is a function based on the average of the corners in the polytope defined by the simplex $S_w = w_1 > w_2 > ... > w_N$, $\Sigma w_i = 1$, and $0 \leq w_i$, where w_i are variables representing the criteria weights. The weights then become the centroid (mass point) components of S_w. The ROC weights are then, for the ranking number i among N items to rank, given by Eq. 1.

$$w_i^{\mathrm{ROC}} = \frac{\sum_{j=i}^{N} \frac{1}{j}}{N} \qquad (1)$$

For instance, [1] introduced a process utilising systematic simulations for validating the selection of criteria weights, when generating surrogate weights as well as "true" reference weights. It also investigated how well the result of using surrogate numbers matches the result of using the "true" numbers. This is however heavily dependent on the distribution used for generating the weight vectors.

Still the problem there is to elicit stakeholder information. Different elicitation formalisms have been proposed by which a decision-maker can express preferences. Such formalisms are sometimes based on scoring points, as in point allocation (PA) or direct rating (DR) methods. In PA, the decision-maker is given a point sum, e.g. 100, which they distribute among the criteria. Sometimes, it is pictured as putty with the total mass of 100 being divided and put on the criteria. The more mass, the larger weight on a criterion, the more important it is. When the first N–1 criteria have received their weights, the last criterion's weight is automatically determined as the remaining mass. Thus, in PA, there is N–1 degrees of freedom (DoF) for N criteria. DR, on the other hand, puts no limit on the total number of points to be allocated. The decision-maker allocates as many points as desired to each criterion. The points are subsequently normalized by dividing by the sum of points allocated. When the first N–1 criteria have received their weights, the last criterion's weight still has to be assigned by the decision-maker. Thus, in DR, there are N degrees of freedom for N criteria. Regardless of elicitation method, the assumption is that all elicitation is made relative to a weight distribution held by the decision-maker.

We have earlier investigated various aspects of this in a couple of articles and compared state-of-the-art weight methods, both ordinal (ranking only) [7,8] and cardinal (with the possibilities to express strength) [9-11] in order to devise methods requiring as little cognitive load as possible. We also used these together with ranked values (utilities) and suggested a multi-stakeholder decision method that has been applied in, e.g., [10]. This method fulfils several desired robustness properties and is comparatively stable under reasonable assumptions.

4. Elicitation methods

The crucial issue in all these methods is how to assign surrogate weights while losing as little information as possible and ensuring correctness when assigning the weights. Providing ordinal rankings of criteria seems to avoid some of the difficulties associated with the elicitation of exact numbers. It puts fewer demands on decision-makers and is thus, in a sense, effort-saving. Furthermore, there are techniques for handling ordinal rankings with various success. A limitation of this is naturally that decision-makers usually have more knowledge of the decision situation than a pure criteria ordering, often in the sense that they have an idea regarding importance relation information containing strengths. In such cases, the surrogate weights may be an unnecessarily weak representation, why we have also investigated whether the methods can be extended to accommodate information regarding relational strengths as well, while still preserving the property of being less demanding and more practically useful than other types of methods.

One well-known class of methods is the SMART family, where [12,13] proposed a method to assess criteria weights. The criteria are ranked and then 10 points are assigned to the weight of the least important criterion (w_N). Then, the remaining weights (w_{N-1} through w_1) are given points according to the decision-maker's preferences. The overall value $E(a_j)$ of alternative a_j is then a weighted average of the values v_{ij} associated with a_j (Eq. 2):

$$E(a_j) = \frac{\sum_{i=1}^{N} w_i v_{ij}}{\sum_{i=1}^{N} w_i} \qquad (2)$$

The most utilised processes for converting ordinal input to cardinal use automated procedures and yield exact numeric weights. For instance, [14] proposed the SMARTER method for eliciting ordinal information on importance before converting it to numbers, thus relaxing information input demands on the decision-maker. An initial analysis is carried out where the weights are ordered such as $w_1 > w_2 > ... > w_N$ and then subsequently transformed to numerical weights using ROC weights and then SMARTER continues in the same manner as the ordinary SMART method.

The most well-known ratio scoring methods is the Analytic Hierarchy Process (AHP). The basic idea in AHP [15,16] is to evaluate a set of alternatives under a criteria tree by pairwise comparisons. The process requires the same pairwise comparisons regardless of scale type. For each criterion, the decision-maker should first find the ordering of the alternatives from best to worst. Next, he or she should find the strength of the ordering by considering pairwise ratios (pairwise relations) between the alternatives using the integers 1, 3, 5, 7, and 9 to express their relative strengths, indicating that one alternative is equally good as another (strength = 1) or three, five, seven, or nine times as good. It is also allowed to use the even integers 2, 4, 6, and 8 as intermediate values, but using only odd integers is more common.

As an, as we have claimed in [11], a better alternative to these, we have suggested the CAR method and shown that this is more robust and efficient than methods from the SMART family and AHP. We will also combine this method with a SWING model variety from [29]. We will develop this further below.

5. The Workshop Setup

There will be up to four interrelated workshops with a progressive process:

- During the first workshop, the alternative strategies as well as the criteria structure is formed, i.e., the focus is at the goal and strategy parts. The elicitation is done by stakeholders during workshops, wherafter these are categorized by a moderator in dialogue with the concerned parties. The resulting structure van then be a basis for refinement. The details of the first workshop are described in section 5.1 below.
- The second workshop will focus on strategy refinement or policy formation as well as criteria elaboration. It also explains what will happen next during the preference elicitation amd evaluation phases.
- The preference elicitation will take third workshop, where criteria weights and alternative valuation is focussed. During this, the alternatives will be tentatively analysed. The details of the third workshop is described in section *The CAR Method* below.
- Final evaluations, sensitivity, stability analyses and refinements will be done during the final workshop. The details of the fourth workshop is described in section *Evaluations under Strong Uncertainty* below.

5.1 Workshop 1

The workshops with stakeholders will last an entire day and include several sessions. The first session starts with the introduction overall process, during which the organizers presented the

workshop and its objectives as well as the goals of the workshops and the agenda. The participants introduce themselves and their organizations.

During the second session, the overall goals are discussed. Participants had an opportunity to, for instance, describe how they see social and economic aspects. Then they write their choices on the different coloured cards and put them on a flipchart. Furthermore, they explain their choices.

5.2 Workshop 2

The first session focus on the refinement of criteria and possible sub-criteria. First, the criteria and their definitions are presented to the participants. Each criterion is discussed to make sure that participants understand its definition. Participants also have a chance to provide suggestions on how the definition of criteria could be changed and to add further criteria.

During the second session, alternative options are discussed followed by a discussion of positive and negative sides of the options. Participants also have a chance to suggest further options, which were not originally included in the list of discussed technologies.

During the third session, preference rankings are introduced to be followed up during Workshop 3.

5.3 Workshop 3

The first session is on criteria ranking during silent negotiation given the alternative options from Workshop 2. The silent ranking is a tool for collective ranking but in silence by avoiding any discussion.

The following rules applies to the session: At the beginning, the set of cards is displaced on the table in a random order. Then the moderator explains the ranking and the rules and ask participants to order cards in three rounds of silent negotiations. The three rounds are followed by a discussion to identify lines of conflicting opinions. During the first three rounds, participants make eight moves during the first round, five moves during the second round, three moves during the third round and finally, after the open discussion, two moves in the fourth and final round. The order how participants are putting the cards was identified by the lottery.
The second session is on silent negotiation and white cards. The moderator introduces the blank cards and explains that they show the relative difference in importance for different criteria. The greater the difference in importance between two criteria, the more blank cards should be positioned in-between these criteria. Altogether, there are three rounds of silent negotiations. The first round has three moves, the second one has two moves and is followed by the open discussion. The final round has one move.

During the third session, the entire decision structure is analysed and tentatively discussed.

5.4 Workshop 4

The final workshop aims to re-rank the criteria again and to discuss the results. The procedure during the workshops was similar to that during workshop 3, except that we rather have two

rounds of rankings when participants have a chance to see and discuss our results between the rankings and then rank the criteria again. Participants also have to explain to other participants the rankings against the background of the final results. The participants discuss procedural and output justice. This discussion focused on the following questions:

- Access to information: How high is the need for information about the different alternatives?
- Meaningful participation in decision-making: How high is the need for participation concerning the different alternatives?
- Benefit sharing: How high is the need to share a reasonable amount of benefits with immigrants?
- Compensation of adverse impacts: How high is the need to claim the right to compensation?

To discuss these questions, the participants form two groups and try to reach a compromise on grouping four criteria in a ranking according to what they believe is important. Later, the results of the ranking are discussed, and participants provide their arguments why they find some criteria to be more important than others.

6. The CAR Method

Simos proposed a simple procedure, using a set of cards, trying to indirectly determine numerical values for criteria weights [23-24]. The Simos method is, however, a bit different from the methods discussed above. It is a relatively simple method for easily expressing criteria hierarchies while introducing some cardinality if needed. It has been widely applied and has been well-received by real decision-makers. When this method is used, a group of decision-makers are provided with a set of coloured cards with the criteria names written on them. They are also given a set of blank cards. Then, they are asked to rank, the coloured cards from the least important to the most important, where criteria of equal importance are grouped together. Furthermore, the decision-makers are asked to place the blank cards in-between the coloured cards to express preference strengths. Then, the surrogate numbers can be computed. A constant value difference, ‘u’, between two consecutive cards is assumed here. A blank card between two consecutive coloured cards signifies a difference of 2·u, and two white cards represent a difference of 3·u, etc.

However, one problem with the Simos method is that it is not robust when the preferences are changed [36] and that it has some other contra-intuitive features, such as that it only picks one of the weight vectors satisfying the model, while there can, of course, be an infinite number of them. Furthermore, because the weights are determined differently depending on the number of cards in the subsets of equally ranked cards, the differences between the weights also change in an uncontrolled way when the cards are reordered. This is why [37] suggested a revised version, where there is a more robust proportionality when these blank cards are used. It is accomplished by requesting the decision-makers to state how many times more important the most important criterion or criteria group is—compared to the least important. This addition seemingly solves some problems but introduces the complication that the decision-maker has to reliably and correctly estimate a proportional factor ‘z’ between the largest and the smallest criteria weights.

We, therefore, use a variant of the Simos method for elicitation purposes and kept the card ranking part while changing the evaluation significantly compared to the Simos method and its revisions. At that point, the participants already know the criteria well from the previous sections of the workshops. The key challenge in our workshops is to elicit a collective ranking. Most methods for ranking and weighting deal with individuals, we have to do it as a group effort. This is the main reason to opt for the card-ranking through a silent negotiation, not the calculation behind it.

Each criterion is written on a coloured card and arranged horizontally on a table. Then each of the participants successively rank the cards from the least important to the most important by moving the cards to a vertical arrangement, where the highest-ranked criterion is put on top and so forth. If two criteria are considered to be of equal importance, they are put on the same level. This process goes on for four rounds, where the number of moves for each round are 8, 5, 3 and 2. Furthermore, the first and third round are concluded by an open discussion before the following round. The ranking procedure laste120 minutes or until a final ranking is achieved that the participants find acceptable.

It is true that the decreasing number can be disputed and is a weak point of the method since it induces / forces the participants to act strategically in relation to the information they got during the process. So when this method is used, the potential conflicts must come to the open and be dealt with. In some cases, by working with a set of final ranking in the evaluations, where it turns out whether the differences are of importance or not. After the first ordinal ranking is finalized, the participants are asked to introduce preference strengths in the ranking by introducing the blank cards during three additional rounds (with three, two and one move). The number of white cards (i.e. The strength of the rankings between criteria) is also interpreted verbally:

Table 1: Blank cards	
Equal level of cards	Equally important
No blank card	Slightly more important
One blank card	More important (clearly more important)
Two blank cards	Much more important
Three blank cards	Extremely more important

While being more cognitively demanding than ordinal weights, these are still less demanding than, for example, AHP weight ratios or point scores like. In an analogous manner as for ordinal rankings, the decision-maker statements can by using these be converted into weights.

The final rankings of the workshops and its rationales are then presented to the other participants during an introductory presentation round.

7. Preference strengths

In analogy with the ordinal weight functions above, counterparts using the concept of preference strength can straightforwardly be derived.

1. Assign an ordinal number to each importance scale position, starting with the most important position as number 1.
2. Let the total number of importance scale positions be Q. Each criterion i has the position $p(i) \in \{1,\ldots,Q\}$ on this importance scale, such that for every two adjacent criteria c_i and c_{i+1}, whenever $c_i >_{s_i} c_{i+1}$, $s_i = |\, p(i+1) - p(i)\,|$. The position $p(i)$ then denotes the importance as stated by the decision-maker. Thus, Q is equal to $\Sigma s_i + 1$, where $i = 1,\ldots,N-1$ for N criteria.

Then the cardinal counterparts to the ordinal ranking methods above can be found by using the results from [10], where the ordinal SR weights were given by Eq. 3

$$w_i^{\mathrm{SR}} = \frac{{}^{1}/_{i} + \frac{N+1-i}{N}}{\sum_{j=1}^{N} w_j^{\mathrm{SR}}} \qquad (3)$$

and using steps 1–3 above, the corresponding preference strength SR weights (CSR, Eq. 4) are obtained as

$$w_i^{\mathrm{CSR}} = \frac{{}^{1}/_{p(i)} + \frac{Q+1-p(i)}{Q}}{\sum_{j=1}^{N}\left({}^{1}/_{p(j)} + \frac{Q+1-p(j)}{Q}\right)} \qquad (4)$$

Using the idea of importance steps, ordinal weight methods in general are easily generalised to their respective counterparts. In the same manner, values (or utilities) can be ranked, either ordinally (ranking only) or cardinally (additionally expressing strength).

Already in [11], we combined cardinal weights with cardinal values into the CAR method and assessed the method by simulations as well as a large number of user cases. The CAR method was found to outperform SMART and AHP on terms of performance and the ease of use (the cognitive load), but some of the users still wanted a method with even less cognitive load so we tried to satisfy this while still preserving reasonable requirements of correctness. The CAR method follows the three-step procedure below [11].

First, the values of the alternatives under each criterion are elicited in a way similar to the weights described above:

1A. For each criterion in turn, rank the alternatives from the worst to the best outcome.

1B. Enter the strength of the ordering. The strength indicates how strong the separation is between two ordered alternatives. Similar to weights, the strength is expressed in the notation with '$>_i$' symbols.

Second, the weights are elicited with a swing-like procedure in accordance with the discussion above.

2A. For each criterion in turn, rank the importance of the criteria from the least to the most important.

2B. Enter the strength of the ordering. The strength indicates how strong the separation is between two ordered criteria. The strength is expressed in the notation with '$>_i$' symbols.

Third, the usual weighted overall value is calculated by multiplying the centroids of the weight simplex with the centroid of the alternative value simplex.
As described in [25], the same description can be used to introduce the three candidate methods called C+O, O+C, and O+O depending on whether a cardinal or ordinal procedure is used for the representation of weights and values respectively. In the original CAR method all the steps 1A, 1B, 2A, 2B, and 3 was performed in that order. The steps in the three candidate methods that we suggest are performed as follows: In O+C, step 1B is omitted, resulting in the sequence 1A, 2A, 2B, and 3 in order. In C+O, step 2B is omitted instead, resulting in the sequence 1A, 1B, 2A, and 3 in order. Finally, in O+O, both steps 1B and 2B are omitted, resulting in the sequence 1A, 2A, and 3 in order.

We will in the next section compare these with combinations of the other ordinal and cardinal methods in search for a method with less cognitive load but still performing better than SMART and AHP. This is, to our knowledge, the first time ordinal and cardinal ranking methods (and combinations thereof) have been compared systematically in this way.

Now we turn our attention to the general evaluation of the entire decision problem.

8. Evaluations under Strong Uncertainty

We will use the evaluation method from [26]. In the type of multi-criteria decision problems we consider, we hold that strong uncertainty exists if the decision is also made under risk, with uncertain consequences for at least one criterion, in combination with imprecise or incomplete information with respect to probabilities, weights, and consequences or alternative values. Decision evaluation under strong uncertainty and computational means for evaluating these models should both be capable of embracing the uncertainty in the evaluation rules and methods and provide evaluation results reflecting the effects of uncertainty for the subsequent discrimination between alternatives.

We will call our representation of a combined decision problem a multi-frame. Such a frame collects all information necessary for the model in one structure. One part of this is the concept of a graph.

Definition: A graph is a structure $\langle V,E\rangle$ where V is a set of nodes and E is a set of node pairs. A tree is a connected graph without cycles. A rooted tree is a tree with a dedicated node as a root. The root is at level 0. The adjacent nodes, except for the nodes at level i-1, to a node at level i is at level i+1. A node at level i is a leaf if it has no adjacent nodes at level i+1. A node at level i+1 that is adjacent to a node at level i is a child of the latter. A (sub-)tree is symmetrical if all nodes at level i have the same number of adjacent nodes at level i+1. The depth of the tree is max (n | there exists a node at level n).

Definition: A criteria-consequence tree $T = \langle C \cup A \cup N \cup \{r\}, E\rangle$ is a tree where
r is the root,
A is the set of nodes at level 1,

C is the set of leaves, and
N is the set of intermediary nodes in the tree except those in A.

In a multi-frame, represented as a multi-tree, user statements can either be range constraints or comparative statements (see below); they are translated into inequalities and collected together in a value constraint set. For probability and weight statements, the same is done into a node constraint set. We denote the values of the consequences c_i and c_j by v_i and v_j respectively. User statements have the following forms for real numbers a_1, a_2, b_1, b_2, d_1 and d_2:

- *Range constraints*: v_i is between a_1 and a_2, denoted $v_i \in [a_1, a_2]$ and translated into $v_i > a_1$ and $v_i < a_2$.
- *Comparisons*: v_i is larger than v_j by an amount ranging from d_1 to d_2, denoted $v_i - v_j \in [d_1, d_2]$ and translated into $v_i - v_j > d_1$ and $v_i - v_j < d_2$.

A constraint set is said to be consistent if it can be assigned at least one real number to each variable so that all inequalities are simultaneously satisfied. Consequently, we get potential sets of functions with an infinite number of instantiations.

Definition: Given a criteria-consequence tree T, let N be a constraint set in the variables $\{ n_{...i...j...} \}$. Substitute the intermediary node labels $x_{...i...j...}$ with $n_{...i...j...}$. N is a node constraint set for T if, for all sets $\{n_{...i1},\ldots,n_{...im}\}$ of all sub-nodes of nodes $n_{...i}$ that are not leaves, the statements $n_{...ij} \in [0,1]$ and $\sum_j n_{...ij} = 1$, $j \in [1,\ldots,m]$ are in N.

A probability node constraint set relative to a criteria-consequence tree then characterizes a set of discrete probability distributions. Weight and value constraint sets are analogously defined. Weight and probability node constraint sets also contain the usual normalization constraints ($\sum_j x_{ij} = 1$) requiring the probabilities and weights to total one.

Definition: A multi-frame is a structure $\langle T,N \rangle$, where T is a criteria-consequence tree and N is a set of all constraint sets relative to T.

The probability, value and weight constraint sets thus consist of linear inequalities. A minimal requirement is that it is consistent—i.e. there must exist some vector of variable assignments that simultaneously satisfies each inequality in the system.

Definition: Given a consistent constraint set X in the variables $\{x_i\}$, ${}^{X}\max(x_i) =_{def} \sup(a \mid \{x_i > a\} \cup X$ is consistent. Similarly, ${}^{X}\min(x_i) =_{def} \inf(a \mid \{x_i < a\} \cup X$ is consistent. Furthermore, given a function *f*, ${}^{X}\text{argmax}(f(x))$ is a solution vector that is a solution to ${}^{X}\max(f(x))$, and ${}^{X}\text{argmin}(f(x))$ is a solution vector that is a solution to ${}^{X}\min(f(x))$.

The set of orthogonal projections of the solution set is the orthogonal hull, consisting of all consistent variable assignments for each variable in a constraint set.

Definition: Given a consistent constraint set X in $\{x_i\}_{i \in [1,\ldots n]}$, the set of pairs $\langle {}^{X}\min(x_i), {}^{X}\max(x_i) \rangle$ is the orthogonal hull of the set.

The orthogonal hull is the upper and lower probabilities (weights, values) if X consists of probabilities (weights, values). The hull intervals are calculated by first finding a consistent point. Thereafter, the minimum and maximum of each variable are found by solving linear programming problems. Because of convexity, the intervals between the extremal points are feasible—i.e. the entire orthogonal hull has been determined.

9. Introducing Second-Order Beliefs

We will first extend the representation to obtain a more granulated representation of a decision problem. Often when we specify an interval, we probably do not believe in all values in the intervals equally: we may, for example, believe less in the values closer to the borders of the intervals. Additional values are nevertheless added to cover everything that we perceive as possible in uncertain situations. These additions give rise to belief distributions indicating the different strengths with which we believe in the different values. Distributions over classes of weight, probability and value measures have been developed into various models, such as second-order probability theory.

In the extended model, we introduce a focal point to each of the intervals used as parameters for belief distributions for probabilities, values and criteria weights. We can then operate on these distributions using additive and multiplicative combination rules for random variables. The detailed theory of belief distributions in this sense is described in [38-42].
To make the method more concrete, we introduce the unit cube as all tuples $(x_1, \ldots, x_n)$ in $[0,1]^n$. A second-order distribution over a unit cube *B* is a positive distribution *F* defined on *B* such that

$$\int_B F(x)\, dV_B(x) = 1,$$

where V_B is the n-dimensional Lebesgue measure on *B*.

We will use second-order joint probability distributions as measures of beliefs. Different distributions are utilized for weights, probabilities and values because of the normalization constraints for probabilities and weights. Natural candidates are then the Dirichlet distribution for weights and probabilities and two- or three-point distributions for values. In brief, the Dirichlet distribution is a parameterized family of continuous multivariate probability distributions. It has a probability density function given by a function of those parameters, such that $\alpha_1,\ldots,\alpha_k > 0$ depends on a beta function and the product of the parameters x_i.

More precisely, the probability density function of the Dirichlet distribution is defined as

$$f_{dir}(p,\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} {p_1}^{\alpha_i - 1} {p_2}^{\alpha_2 - 1} \ldots {p_k}^{\alpha_k - 1}$$

on a set $\{p = (p_1,\ldots p_k) \mid p_1,\ldots,p_k \geq 0, \Sigma p_i = 1\}$ where $(\alpha_1,\ldots, \alpha_k)$ is a parameter vector in which each $\alpha_i > 0$ and $\Gamma(\alpha_i)$ is the Gamma function.

The Dirichlet distribution is a multivariate generalization of the beta distribution and the marginal distributions of Dirichlet are thus beta distributions. The beta distribution is a family of continuous probability distributions defined on [0, 1] and parameterized by two parameters, α and β, defining the shape of the distribution.

For instance, if the distribution is uniform, the resulting marginal distribution (over an axis) is a polynomial of degree n−2, where n is the dimension of a cube B. Let all $\alpha_i = 1$, then the Dirichlet distribution is uniform with the marginal distribution

$$f(x_i) = \int_{B_i^-} dV_{B_i^-}(x) = (n-1)(1-x_i)^{n-2}$$

However, for our purposes, we need a bounded Dirichlet distribution operating on a user-specified $[a_i, b_i]$ range instead of the general interval [0,1]. Bounded beta distributions are then derived—the so-called four-parameter beta distributions, also defined only on the user-specified range. We then define a probability or weight belief distribution as a three-point bounded Dirichlet distribution $f_3(a_i, c_i, b_i)$ where c_i is the most likely probability or weight and a_i and b_i are the boundaries of the belief with $a_i < c_i < b_i$.

For values, the generalization to a trapezoid from a triangle is analogous. We will utilize either a two-point distribution (uniform, trapezoidal) or a three-point distribution (triangular). When there is a large uncertainty regarding the underlying belief distribution in values and we have no reason to make any more specific assumptions, a two-point distribution modelling upper and lower bounds (the uniform or trapezoid distributions) is preferred. On the other hand, when the modal outcome can be estimated, the beliefs are more congenially represented by three-point distributions. The Beta and Erlang belief distributions are widely used in many models and generally give results similar to triangular distributions; we use the latter as a representative for the class of three-point distributions. This is thus a description when there are only limited sample data, particularly in cases where the variable relationships are known as well as the minimum, maximum and modal values. The PERT distribution is a classic example and the mean value of these three-point value belief distributions $f_3(a_i, c_i, b_i)$ is $\mu(\lambda) = (a_i + b_i + \lambda c_i)/(\lambda + 2)$, with special cases $\lambda = 1$ for a triangular distribution and $\lambda = 0$ for a two-point uniform or trapezoid distribution.[2] Because triangular distributions are less centre-weighted than other three-point distributions, the risk of underestimation is less, which is why there are no particular reasons to use any other distribution for practical purposes.

10. The Evaluation Model

We will use a generalization of the ordinary expected value for the evaluation—i.e. the resulting distribution over the generalized expected utility is

$E(A_i) =$,

[2] Beta-PERT usually has $\lambda = 4$ and Erlang-PERT has $\lambda = 3$. However, higher values of λ tend to underestimate the uncertainties involved.

given the distributions over the random variables *p* and *v*. There are only two operations of relevance here, multiplication and addition.

Let G be a distribution over the two cubes A and B. Assume that G has a positive support on the feasible distributions at level *i* in a general decision tree, as well as on the feasible probability distributions of the children of a node x_{ij} and assume that *f*(*x*) and *g*(*y*) are the marginal distributions of G(*z*) on A and B, respectively. Then the cumulative multiplied distribution of the two belief distributions is

$$\mathrm{H(z)} = \iint_{\Gamma_x} f(x)g(y)dxdy = \int_0^1 \int_0^{z/x} f(x)g(y)dxdy = \int_z^1 f(x)G(z/x)dx$$

where G is a primitive function to g, $\Gamma_z = \{(x,y) \mid x \cdot y \leq z\}$, and $0 \leq z \leq 1$.
Let *h(z)* be the corresponding density function. Then

$$\mathrm{h(z)} = \frac{d}{dz}\int_z^1 f(x)G(z/x)dx = \int_z^1 f\frac{f(x)g(z/x)}{x}dx.$$

The addition of the products is the standard convolution of two densities restricted to the cubes. The distribution h on a sum $z = x + y$ associated with the belief distributions *f*(*x*) and *g*(*y*) is therefore given by

$$\mathrm{h(z)} = \frac{d}{dz}\int_0^z f(x)g(z-x)dx.$$

Then we can obtain the combined distribution over the generalized expected utility.

As in most of risk and decision theory, we assume that a large number of events will occur and a large number of decisions will be made. This way, the expected value becomes a reasonable decision rule and, at the same time, the belief distributions over the expected values tend to normal distributions or similar. Note that even when assuming that the expectations are estimated a large number of times and consequently can be approximated by a normal distribution, there are three particular observations to be made:

The resulting distributions will be normal only when the original distributions are symmetrical, which of course is not usually the case for beta and triangular distributions. The result then will instead be skew-normal.

Even if the original distributions are symmetrical, the non-linear multiplication operator breaks the symmetry. The result then will again be skew-normal.
To obtain a resulting normal (or skew-normal) distribution, both the original distributions and their aggregations must allow for long tails. This is not generally the case in our models since our estimates have lower and upper limits.

We therefore use a truncated skew-normal distribution, generalizing the normal distribution by allowing for non-zero skewness and truncated tails. This is accomplished by introducing a shape parameter α, where the standard normal distribution has $\alpha = 0$, and where $\alpha = 1$ yields the distribution of the maximum of two independent standard normal variates. We can then conveniently represent truncated (skew-)normal distributions as probability distributions of (skew-)normally distributed random variables that are bounded. The skewness of the distribution increases along with the absolute value of α, and when $|\alpha| \to \infty$, we obtain folded

normal or half-normal distributions. Distributions are right-skewed when $\alpha > 0$ and left-skewed when $\alpha < 0$. Assume that a distribution X has a normal distribution within the interval (a, b). Then X, $a < X < b$, has a truncated normal distribution and its probability density function is given by a four-parameter expression that tends to normality as the intervals are widened.

11. Results

The analyses will be supported by the tool decideIT. Figure 3 shows an example of one of the result windows from the tool.

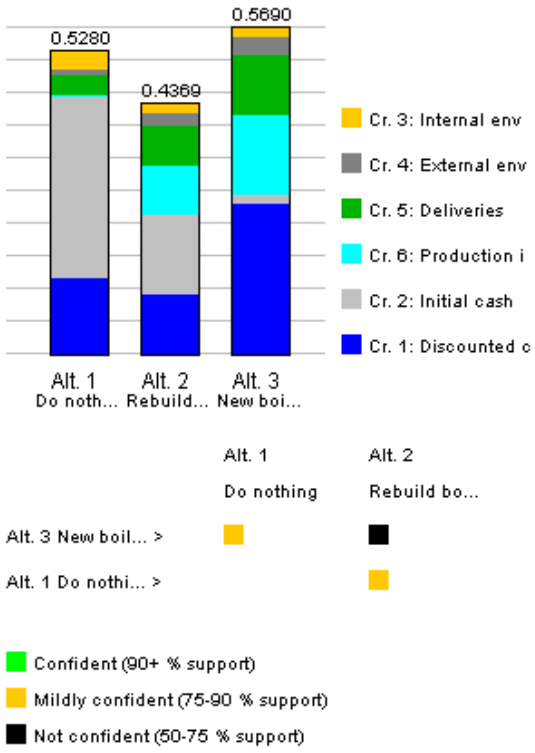

Figure 3. Main decision evaluation result.

The empirical data collected during the workshops allowed us to develop the following sets of results:

- Stakeholders' visions about the economic, societal and environmental future of immigration.
- Perceptions of risks and benefits of different options.
- Rankings of different criteria.
- Trade-offs of technologies, including results based on a modelling of criteria ranking.
- Individual evaluations during the following up survey.

References

1. Barron, F. and Barrett, B., Decision Quality Using Ranked Attribute Weights. Management Sci. 42(11), 1515–1523 (1996b).
2. Jia, J., Fischer G.W. and Dyer, J., Attribute weighting methods and decision quality in the presence of response error: a simulation study, J. Behavioral Decision Making 11(2), 85–105 (1998).

3. Barron, F. and Barrett, B., The Efficacy of SMARTER: Simple Multi–Attribute Rating Technique Extended to Ranking. Acta Psych. 93(1–3), 23–36 (1996a).
4. Katsikopoulos, K. and Fasolo, B., New Tools for Decision Analysis. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 36(5), 960–967 (2006).
5. Stillwell, W., Seaver, D. and Edwards, W., A Comparison of Weight Approximation Techniques in Multiattribute Utility Decision Making. Org. Behavior and Human Performance 28(1), 62–77 (1981).
6. Barron, F.H., Selecting a Best Multiattribute Alternative with Partial Information About Attribute Weights. Acta Psych. 80(1–3), 91–103 (1992).
7. Danielson, M. and Ekenberg, L., and He, Y., Augmenting Ordinal Methods of Attribute Weight Approximation, Decision Analysis, Vol. 11(1), pp. 21–26, 2014.
8. Danielson, M. and Ekenberg, L., Rank Ordering Methods for Multi-Criteria Decisions, Proc. 14th Group Decision and Negotiation – GDN 2014, Springer (2014).
9. Danielson, M. and Ekenberg, L., Trade-offs for Ordinal Ranking Methods in Multi-Criteria Decisions, proceedings of GDN 2016, Springer.
10. Danielson, M. and Ekenberg, L., A robustness study of state-of-the-art surrogate weights for MCDM, Group Decision and Negotiation, 7, 2016, doi: 10.1007/s10726-016-9494-6.
11. Danielson, M. and Ekenberg, L., The Car Method for using Preference Strength in Multi-Criteria Decision Making, Group Decision and Negotiation, 25(4), pp.775–797, 2016, doi: 10.1007/s10726-015-9460-8.
12. Edwards, W., Social utilities. Engineering Economist, Summer Symposium Series 6, 119–129, 1971.
13. Edwards, W., How to Use Multiattribute Utility Measurement for Social Decisionmaking. IEEE Transactions on Systems, Man and Cybernetics 7(5), 326–340, 1977.
14. Edwards, W. and Barron, F., SMARTS and SMARTER: Improved Simple Methods for Multiattribute Utility Measurement. Organizational Behavior and Human Decision Processes 60, 306–325, 1994.
15. Saaty, T.L., A Scaling Method for Priorities in Hierarchical Structures, Journal of Mathematical Psychology 15, 234–281, 1977.
16. Saaty, T.L., The Analytic Hierarchy Process, McGraw-Hill: New York, 1980.
17. Ahn, B.S. and Park, K.S., Comparing methods for multiattribute decision making with ordinal weights, Computers & Operations Research 35 (5), 1660–1670, 2008.
18. Barron, F. and Barrett, B., The Efficacy of SMARTER: Simple Multi-Attribute Rating Technique Extended to Ranking. Acta Psychologica 93(1–3), 23–36, 1996a.
19. Barron, F. and Barrett, B., Decision Quality Using Ranked Attribute Weights. Management Science 42(11), 1515–1523, 1996b.
20. Butler, J., Jia, J. and Dyer, J.: Simulation Techniques for the Sensitivity Analysis of Multi-Criteria Decision Models. European Journal of Operational Research 103, 531-546, 1997.
21. Danielson, M. and Ekenberg, L.: Computing Upper and Lower Bounds in Interval Decision Trees, European J. Operational Res. 181(2), 808–816 (2007).
22. Danielson, M., Ekenberg, L., Larsson, A. and Riabacke, M., Weighting Under Ambiguous Preferences and Imprecise Differences in a Cardinal Rank Ordering Process, Int. J. Comp. Int. Syst. (2013).

23. Simos, J., L'evaluation environnementale: Un processus cognitif neegociee. Theese de doctorat, DGF-EPFL, Lausanne. (1990).
24. Simos, J., Evaluer l'impact sur l'environnement: Une approche originale par l'analyse multicriteere et la negociation. Presses Polytechniques et Universitaires Romandes, Lausanne. (1990).
25. Danielson, M. and Ekenberg, L.: Automatic Criteria Weight Generation for Multi-Criteria Decision Making under Uncertainty, to appear in A. de Almeida, L. Ekenberg, P. Scarf, E. Zio, M.J. Zuo, Multicriteria Decision Models and Optimization for Risk, Reliability, and Maintenance Decision Analysis - Recent Advances, Springer 2019.
26. M. Danielson, L. Ekenberg and A. Larsson, Evaluating Multi-Criteria Decisions Under Strong Uncertainty, to appear in A. de Almeida, L. Ekenberg, P. Scarf, E. Zio, M.J. Zuo, Multicriteria Decision Models and Optimization for Risk, Reliability, and Maintenance Decision Analysis - Recent Advances, Springer 2019.
27. M. Danielson and L. Ekenberg, Computing Upper and Lower Bounds in Interval Decision Trees, *European Journal of Operational Research* 181, pp. 808–816, 2007.
28. N. Komendantova, E. Rovenskaya, L. Ekenberg, N. Strelkovski, S. Sizov, E. Sedighi, A. Stepanova, N. Karabashov, N. Atakanov, U. Chekirbaev, Z Zheenaliev and F. Santiago, Connecting Regional Development, Regional Integration and Value Added Creation, Strategic elements for Industrial Development of Kyrgyzstan, United Nations Industrial Development Organization, 2018a.
29. M. Danielson and L. Ekenberg, An Improvement to Swing Techniques for Elicitation in MCDM Methods, Knowledge-Based Systems, 2019, https://doi.org/10.1016/j.knosys.2019.01.001.
30. N. Komendantova, L. Ekenberg, L. Marashdeh, A. Al-Salaymeh, M. Danielson and J. Linnerooth-Bayer, Are Energy Security Concerns Dominating Environmental Concerns? Evidence from Stakeholder Participation Processes on Energy Transition in Jordan, Climate, 2018b.
31. A. Larsson, T. Fasth, M. Wärnhjelm, L. Ekenberg and M. Danielson, Policy Analysis on the Fly with an Online Multi-Criteria Cardinal Ranking Tool, Journal of Multi-Criteria Decision Analysis, 2018:1–12. https://doi.org/10.1002/mcda.1634.
32. L. Ekenberg, T. Fasth, and A. Larsson, Hazards and Quality Control in Humanitarian Demining, International Journal of Quality & Reliability Management 35(4), pp. 897–913. 2018, doi:10.1108/IJQRM-01-2016-0012.
33. M. Danielson and L. Ekenberg, Efficient and Sustainable Risk Management in Large Project Portfolios, proceedings of BIR 2018 (17th International Conference on Perspectives in Business Informatics Research), Springer, 2018.
34. L. Ekenberg, K. Hansson, M. Danielson, G. Cars, et al, Deliberation, Representation and Equity: Research Approaches, Tools and Algorithms for Participatory Processes, 384p, ISBN 978-1-78374-304-9, Open Book Publishers, 2017.
35. A. Mihai, A. Marincea, and L. Ekenberg, A MCDM Analysis of the Roşia Montană Gold Mining Project, *Sustainability* Vol. 2015(7), pp. 7261–7288, doi:10.3390/su7067261, 2015.
36. Scharlig, A., (1996). Pratiquer Electre et PROMETHEE Un complement à decider sur plusieurs critères. Collection Diriger L'Entreprise, Lausanne: Presses Polytechniques et Universitaires Romandes, 1996.
37. Figueira, J., and Roy, B., (2002). Determining the weights of criteria in the ELECTRE type methods with a revised Simos' procedure. European Journal of Operational Research, 139, 317–326, 2002.

38. Ekenberg L, Thorbiörnson J. Second-order decision analysis. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2001;9(1):13–38.
39. Ekenberg L, Thorbiörnson J, Baidya T. Value differences using second order distributions. International Journal of Approximate Reasoning 2005;38(1):81–97.
40. Danielson M, Ekenberg L, Larsson A. Belief distribution in decision trees. International Journal of Approximate Reasoning 2007;46(2):387–407.
41. Danielson M, Ekenberg L, Larsson A, Sundgren D. Second-order risk constraints in decision analysis. Axioms 2014;3:31–45.
42. Sundgren D, Danielson M, Ekenberg L. Warp effects on calculating interval probabilities. International Journal of Approximate Reasoning 2009;50(9):1360–1368.