Developing a monitoring process for IPC Acute Food Insecurity analyses

A case study on Human-Centered AI for humanitarian decision-making

Marijn Roelvink

TU Delft

Cover: Photo from the Virtual Irrigation Academy in Malawi (Modified)
URL: via.farm/stories_malawi

Developing a monitoring process for IPC Acute Food Insecurity analyses

A case study on Human-Centered AI for humanitarian decision-making

by

Marijn Roelvink

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday June 11, 2024 at 15:30.

Thesis committee:	Dr. ir. C.C.S Liem,	TU Delft, supervisor
	Prof. dr. T. Comes,	TU Delft, supervisor
	T. Baar,	IPC
	Prof. dr. ir. B. Taebi,	TU Delft
	Dr. ir. W.P. Brinkman,	TU Delft

Abstract

Due to climate change, man-made conflicts, and rising inflation, a growing number of people around the world are struggling to have consistent access to safe and nutritious food. This phenomenon is known as food insecurity (FI). Therefore, we take in this thesis the first steps towards developing a monitoring process for assessing FI using Human-Centered AI (HCAI). We developed this process for, and in collaboration with, the Integrated Food Security Phase Classification (IPC). The IPC is an organization that helps countries classify levels of food insecurity in their regions to inform humanitarian decision-making. During our research process, we found that any form of HCAI for the IPC would need to be informed by input from their domain experts, and we concluded that we could not start implementing HCAI until we found a way to formalize their input in a way that was robust and suited their technical capabilities. To this end, we ran an experiment with 18 IPC experts in Malawi to see whether they could quantify their assumptions by setting thresholds for food security drivers. The results are encouraging but show that there is still much to be done to bridge the gap between domain knowledge and technical expertise. We also show in this thesis that there is a lack of real-life case studies on HCAI development and share therefore our lessons learned from our real-world HCAI case study on FI monitoring. In this way, we hope to promote the development of a practice-informed methodology for HCAI.

Preface

As I reach the end of my thesis project and, with that, the end of my student time here in Delft, there are many people whom I would like to thank for their help and for making the journey enjoyable. First and foremost, I would like to thank my supervisors Cynthia Liem, Tina Comes, and Thomas Baar.

I am grateful to Cynthia Liem for agreeing to be my thesis supervisor, as she was a role model for me long before we started working together, and my impression of her has only grown since. I really enjoyed our open discussions about our many ideas, insights, and frustrations on various societal and technical topics. I thank Tina Comes for taking a chance on me a year ago when I knocked on her door with more thoughts than words to articulate them. She connected me with Thomas and the IPC, and her sharp questions and TPM expertise brought the thesis to a next level. She encouraged me to set my own boundaries and priorities and made sure that I did not become overwhelmed by the research process. Lastly, I appreciate the many hours Thomas Baar has dedicated to discussing the research with me, for his check-ins on how I was doing, and for guiding me through the IPC's bureaucratic processes. To all three of my supervisors, I am sincerely grateful for the trust they placed in me during the thesis project, for their constructive feedback, and for the creative freedom they granted me. It was truly empowering, and I look forward to working with them in the future.

I also want to give special thanks to all the other people at the IPC who have helped me in various ways to make this thesis possible. The level of open discussion and self-reflection I encountered there was very inspiring and makes me excited to continue this research.

Lastly, I would like to thank the following people:

- My thesis buddies Francis Behnen, Jeroen Nelen, and Khalid El Haji, for joining me in the daily thesis process and the many coffee breaks that made everything manageable.
- My friends Beryl van Gelderen, Vera Hoveling, Maaike Mol, Daniela Patrova, Evelien Scheffers, and David Allaart, for providing feedback on my thesis and being excellent company.
- Marta Gavioli, for her excellent tips and the literature she provided on how to successfully complete a master's thesis (or PhD).
- My old roommates (Selma and Arwen) and new roommates (Marijn, Jeroen, Bram, and Arun) for enduring all my thesis rants and helping me to relax.
- My dad, sister, brother, and Ruth, for their quality advice and the many glasses of good wine.
- And my study buddies and many other friends from the Bende van Ellende, Lijst Bèta, Vocalzz, JC Mulan, Computer Science, PRIME(CH), and other places, who made this period of my life worthwhile.

I also bid a grudgeful thanks to the letter "E" on my keyboard, who, though unreliable, unsound, and unstable, has survived the whole thesis process and allowed me to finish this piece of literature.

Marijn Roelvink
Delft, 2024

Contents

Chapter 1

Introduction

1.1 Motivation

We are living in an age of crises. Food Insecurity (FI) is a prominent one among them. According to the Global Report on Food Crises [13], over 258 million people face acute food insecurity, requiring urgent food, nutrition, and livelihood assistance. These numbers have been growing in the past years due to climate change, conflicts, and soaring food prices due to COVID-19 and the war in Ukraine.

Assessing Food Insecurity with the IPC

One of the most important practices for assessing the level of acute food insecurity in a country is the approach developed by the Integrated Food Security Phase Classification (IPC) [17]. The IPC assessments have a great impact as they inform humanitarian organizations where they should send their aid. Moreover, given the scarcity of humanitarian funds, it is essential that these assessments are as accurate as possible: underestimating the problem results in unnecessary deaths and hardships while overestimating it diverts limited humanitarian resources away from areas with greater needs. As of 2022, only 56% of the humanitarian funding needs were met [8].

The IPC utilizes an evidence- and consensus-based analysis method to classify the severity and magnitude of food insecurity within a country and identify the underlying factors driving it. This process relies on experts from different organizations with extensive knowledge of the country's context and food security dynamics. This not only ensures that the results are backed up by causal factors that can be addressed but also that the outcomes will be supported by relevant stakeholders.

However, as building technical consensus takes a considerable amount of resources in terms of time and data gathering, the IPC process is only carried out once or twice a year at country level. This leaves a lot of room for FI situations to change in the meantime, making it hard for organizations to act on what the current situation demands, especially in dynamic situations such as the outbreak of a conflict or an unpredicted drought. The IPC does make projections of what the food insecurity situation will look like 3 to 12 months into the future, but these projections are based on assumptions about FI drivers that may not remain valid throughout the time in between assessments.

Can AI help them?

Given the increasing urgency of this problem and the rising availability of new data sources, it could be very beneficial to start using AI to monitor Food Insecurity in between IPC analyses so the IPC can be alerted when their projections are starting to be invalid. Using this information, they can update their predictions and start to serve more as an early warning system. Our first goal for this thesis is therefore to develop an AI system that can assist the IPC with decision-making on projection updates by assessing whether their projection assumptions are still correct.

Well, maybe watch out with that..

However, caution is warranted when making AI solutions for such political and high-stakes decision-making processes. While the automation and new insights from AI have provided many benefits to the world, the list of cases where its irresponsible use in high-stakes decisions has led to harmful consequences is growing rapidly [1], [22], with the most harm often being felt by groups that were already vulnerable and/or marginalized. In the case of AI for humanitarian decision-making processes such as the IPC, this risk is therefore even greater as their decisions are primarily concerned with the most vulnerable groups in different countries.

Problems in the aforementioned cases were often caused by a combination of biased data, a lack of transparency in the algorithm, and, as an effect, a lack of accountability in the decision-making process [49]. In other words, the algorithms were trained to optimize a mathematical function that did not reflect all relevant values of its institutional context [1] and were designed to give recommendations rather than arguments, in this way promoting automation bias [44] and diluting the autonomy of its users.

While it would be easy to conclude from these cases that AI just shouldn't be used in such high-stakes decision-making processes, we believe that that would be a wasted opportunity, especially in the context of humanitarian decision-making, where every amount of money saved means that that is money that can go to actually helping affected communities. That is why we aimed to investigate how we *can* develop AI that has the institutional values of its context embedded in its design, that enhances rather than replaces its users' reasoning power, and that takes the expertise into account of people who represent vulnerable populations susceptible to harm.

So how can we make this "responsible" AI?

In order to develop such AI, we see two questions that need to be answered: "What" are the requirements that AI needs to adhere to in order to be responsible and trustworthy, and "How" do we design for those requirements, as well as for the needs of the organization?

Based on the literature study we have done in chapter 2, it turns out that the answer to the "What" question is: it is messy. The terms trustworthy, ethical, and responsible AI have been used interchangeably. [40], and organizations that have tried to define requirements for them often appear to work with a seemingly arbitrary subset of high-level requirements that are not properly problematized.

For the "How" question, we found a very promising solution in the concept of Human-Centered AI (HCAI). This is a vision on AI decision-support systems that truly *support*, rather than replace humans and that adhere to our human values [32]. However, this turned out not to be a fully developed concept yet. While there have been papers on frameworks and guidelines for HCAI since 2019, few of those have been used in practice. A reason for this lack of adoption is that all the frameworks that have been proposed in these papers have not been accompanied by any form of case studies or other methods to evaluate whether they are usable in practice and whether they actually help achieve the goals underlying HCAI. As an effect, many of these frameworks either function on a level too high to give practical guidance, or only repeat standard practices in the industry (such as in [42]), without evaluating whether those approaches are actually sufficient for achieving HCAI.

A second research goal

As we do see much potential in the concept of HCAI, this research gap of missing real-life evaluations has led us to a second research goal: to bring the field of HCAI forward by sharing what we have learned from our research on developing an HCAI monitoring system for the IPC. This has resulted in a double set of research goals:

1. Develop an HCAI system that can help the IPC in monitoring their projections in between analyses.
2. Use this process as a case study to help further the field of HCAI by identifying pitfalls to avoid, useful practices to promote, and research gaps to further consider.

1.2 Research questions

Given that our research operates on the intersection of two different domains, (IPC/Food Insecurity monitoring and Human-Centered AI), we have disambiguated our research questions by mapping them each to a part of the Venn diagram of these domains. See Figure 1.1. As HCAI can be involved in a wide range of decision-making processes, we have narrowed the scope of our research on HCAI to the context of humanitarian decision-making. We have chosen this scope based on its shared characteristics with our case study of high-impact and political decision-making and its responsibility concerning vulnerable communities. This has given us three main questions: the (IPC) question of how to develop the monitoring process, the (HCAI+IPC) question of how to do that such that it adheres to Human-Centered AI, and the (HCAI) question: what did we learn from it for HCAI for humanitarian decision-making? Below one can find the full versions of these questions, together with subquestions to answer them.

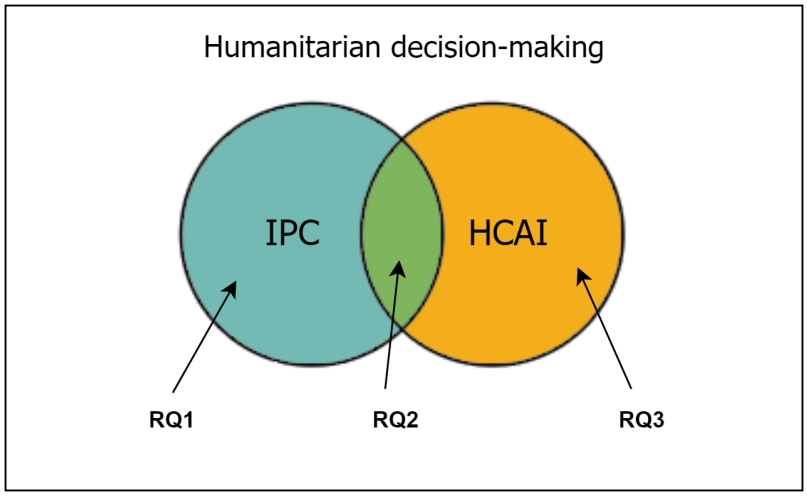


Figure 1.1: Research questions mapped to research fields

RQ1: How can we design an HCAI-supported monitoring process for Acute Food Insecurity (AFI) analyses that explicitly takes the validity of ongoing assumptions into account?

1. How does the IPC analyze Food Insecurity and what computational models are currently available for predicting it?
2. What does the process in the IPC analysis cycle currently look like for developing and monitoring assumptions?
3. How can we change this process to make it produce assumptions that are more formally measurable and verifiable?
4. How can we design a monitoring process that, using HCAI, assesses the validity of AFI analyses based on these assumptions?

RQ2: How can we create a design process for answering RQ1 that fulfills the requirements for HCAI?

1. What are the requirements for HCAI in terms of the product and its development process according to literature?
2. How can we translate these requirements into our design process?

RQ3: Based on our case study, what lessons can we learn about the development of HCAI systems for humanitarian decision-making?

1.3 Contributions

As we are doing research in both Food Insecurity modeling as well as in HCAI, we group our contributions below along these two lines of research, and indicate to which practitioners and researchers we hope to add value.

IPC/Food Insecurity domain

- A practitioner-informed overview of the IPC analysis cycle in practice, extending the current literature and documentation available on the IPC process. This can help policy-makers at the IPC and other FI researchers aiming to improve the IPC to better tailor their research to the IPC's current practices and needs
- A detailed problem analysis of the IPC monitoring process and a high-level plan to improve it. As this plan identifies different solutions that need to be developed in the coming years, this can help steer the efforts of the IPC and FI modelers/researchers.
- A set of design requirements and corresponding values that any monitoring support system for the IPC should adhere to, based on literature and half a year of discussions with people from the IPC on the subject.
- A set of specific criteria that FI machine learning models should meet in order to be usable by the IPC for FI monitoring, and an inventory of how current FI machine learning models perform against these criteria.
- A proof of concept for setting thresholds on projection assumptions as a way to make them more measurable. This provides a basis for further implementation of a data-driven monitoring process.

Human-Centered AI domain

- A literature study on the history of Human-Centred AI explaining its research gap in relation to real-life evaluation. We hope this will steer future HCAI efforts more toward case study-based research.
- A reflection on the requirements for responsible and trustworthy AI: opening up a conversation about how these requirements relate to each other by mapping them to a value hierarchy [35].
- A set of tips and guidelines based on our case study for other researchers and workers in the humanitarian sector aiming to start working with HCAI.

1.4 Thesis structure

The rest of the thesis is structured as follows:

- **Chapter 2** delves deeper into the concept of HCAI and explains our chosen methodology, Human-Centered Design (HCD), and how it can be useful for developing HCAI.
- **Chapter 3** kickstarts the case study part of this thesis by giving a bird's eye view of our case study on FI monitoring and describing how our HCD research plan evolved as the project progressed.
- **Chapter 4** explains the analytical framework on which IPC Acute Food Insecurity assessments are based and provides an analysis of the suitability of current state-of-the-art FI Machine Learning models for our purposes.
- **Chapter 5** dives into our research on the IPC and the interviews we have done with IPC experts to get a better understanding of the context and ends with a detailed overview of the IPC analysis process in practice.
- **Chapter 6** gives, based on the outcomes of Chapters 4 and 5, an analysis of the IPC's problem in monitoring FI and presents our plan on solving it, which is based on better quantifying IPC analysis assumptions.
- **Chapter 7** describes the experiment we have run to see whether our plan of Chapter 6 was feasible by testing whether IPC analysts are able to quantify their projection assumptions using thresholds.
- **Chapter 8** reflects on the case study we described in Chapters 3 to 7 and shares the lessons we have learned about developing HCAI for humanitarian decision-making.
- **Chapter 9** wraps up this thesis by summarizing the answers to our research questions as posed in Section 1.2 and giving pointers for future research.

Chapter 2

Literature Study I: Human-Centered AI

In this chapter, we discuss the field of HCAI and related terms such as responsible AI and trustworthy AI with the intention to answer our RQ2.1: "*What are requirements for HCAI and how can we design for it?*". Moreover, we back up our argument for the need for real-life case studies on doing HCAI, by explaining its history and in doing so, revealing its persistent shortcomings in relation to intersectional case studies and evaluations.

Therefore, we will start this chapter by explaining HCAI and its history, then dive into what requirements different literature has named for HCAI, and finally end with an overview of the design framework that we will use to structure our case study.

2.1 About Human-Centered AI

2.1.1 What do we mean by it?

There have been different definitions for what HCAI is, and encompasses. Some of the more often-used definitions are the one of Shneiderman [43]: "HCAI focuses on amplifying, augmenting, and enhancing human performance in ways that make systems reliable, safe, and trustworthy." and the one of Riedl [36] "Human-centered artificial intelligence is a perspective on Artificial Intelligence (AI) and Machine Learning (ML) that intelligent systems must be designed with awareness that they are part of a larger system consisting of human stakeholders, such as users, operators, clients, and other people in close proximity.". As the first definition lacks any mention of usefulness for humans and the second definition might not make it immediately clear for most readers what the practical implications are of "designing with awareness", we have chosen for a more to-the-point definition derived from the pivotal 2019 paper on HCAI from Xu:

> An HCAI solution is a solution *"encompassing not just ethics and technology, but also explainable, comprehensible, useful, and usable AI"*. [58]

The previously different definitions also show that there have been different visions throughout its history on what HCAI is and who should work on it. In the next section, we will go through this history to explain its development and why we need to start putting it into practice:

2.1.2 A promising field, in theory

The movement around HCAI started around 2016 when the first centre for HCAI was opened by UC Berkeley [39]. In the following years, they were followed as well by Stanford and MIT in 2019 [58]. While the

aim was already to "reorient AI research towards provably beneficial systems, over which humans can retain control even as they approach or exceed human-level decision-making capabilities" [39], their main approach was through a technological lens, looking at issues such as: How can we ensure human control and ethical alignment in our AI systems[39] and: How can we develop novel technologies "inspired by the depth and versatility of human intelligence?" [26]. There was, for example, no mention of explainability of accountability in these reports.

Then in 2019, the Human-Computer Interaction (HCI) community started to get involved. In the paper "Toward human-centered AI" [58], HCI researchers proposed a new framework for doing H(C)AI, which included the two principles as discussed by the research institutes above but had a third new component for doing HCAI: human factors design (roughly put: a specific form of human-centered design). With this component, they argued for a human-centered design approach that should result in AI that is "explainable, comprehensible, useful, and usable." With this new definition, they also called for the HCI community to get more involved with the current AI research in order to start developing this new form of HCAI, as this was not currently happening. They saw themselves as key players in changing the AI landscape: "Thus, a new version of User-Centered Design practice, HAI, has again fallen on the shoulders of HCI professionals, promising many great opportunities for the HCI community.".

In the following years, more papers started being published on the topic, with frameworks, theories, new definitions, and design guidelines [36], [43], [4]. However, the number of practically applied HCAI papers remained lacking: in the paper "Six Human-Centered Artificial Intelligence Grand Challenges" in 2023 [32], they concluded that apart from some experiments, widespread adoption of HCAI remained forthcoming and reiterated the grand role for HCI: *"in the age of AI, HCI can lead the way in providing a much-needed human-centered approach to AI.".*

Xu, from the original paper of 2019, saw this problem of lacking practice as well in 2023 and, together with others, published an extensive paper to kickstart the adoption into practice called "An HCAI Methodological Framework: Putting It Into Action to Enable Human-Centered AI*" [59]. They reasoned that the lack of practical HCAI applications was due to a lack of "comprehensive HCAI methodologies to guide the implementation of HCAI in practice" and decided, therefore, to put yet another HCAI framework forward.

2.1.3 But let us also put it in practice

While we agree that no good HCAI methodology exists yet for developing HCAI in practice, we would argue that any good HCAI methodology can only be developed if it is evaluated with - and developed from - real-life case studies. Only by applying and trying out methods in actual contexts can one discover which methods actually work and what information, considerations and steps are needed during implementation. As Wu et al. state in the introduction of their 2023 paper, back in the 1980s, User-centered design also only started evolving and accelerating once it started to be used and informed by its practitioners.

2.2 Requirements for HCAI

In our definition of HCAI, one can find different dimensions in requirements that any AI needs to adhere to, which can be mapped roughly to different discussions and research fields surrounding AI. When talking about ethical, technological, and explainable aspects of AI, one quickly encounters terms such as "trustworthy" and "responsible AI". In terms of what is needed to make AI useful and usable, it is more a question of design methodology.

While researching the requirements for trustworthy and responsible AI, we found a need to start relating them to each other. These requirements are often presented as isolated concepts, but when mapping them to their corresponding purposes and values, one discovers very quickly that most of those requirements are very related to each other and, in many cases, provide dilemmas and trade-offs: When one moves more towards one value, this can often have a consequence that other values get strayed from. A good example of this is transparency vs. privacy. When trying to be more transparent about decision-making or data use, this could

mean that the privacy of other people involved needs to be breached. This tension can be found in many different decision-making processes.

Therefore, in the next section, we will first explore the discussions around trustworthy and responsible AI and try to come to some sort of holistic overview of their requirements by mapping them to a value hierarchy as defined by van der Poel [35]. In the section following that, we will discuss our chosen design methodology, Human-Centered Design, to develop AI that is useful and usable and we will show through our value hierarchy how it can help to meet some of the norms and values identified in the previous section.

2.3 A value hierarchy for HCAI

2.3.1 About value hierarchies

Van der Poel developed the value hierarchy in his research on translating values into design requirements. Such a hierarchy puts design requirements, norms, and values into perspective by relating them to each other through "for the sake of" relations. At the top of this hierarchy, one can find the values (see Figure 2.1. Values are here things we see as good and that should be strived for; they are the ultimate reasons for our norms and requirements. Norms form the next layer in the hierarchy and are used by Van der Poel as a term to denote any kind of "prescriptions for, and restrictions on, action" [35]. In the last layer are the design requirements. These are the most concrete requirements of what the object of design should adhere to. As we are talking about the more general concept of Human-Centered AI, our design requirements are still quite vague, as further specification is not possible without a specific context.

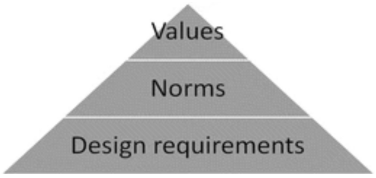


Figure 2.1: Value hierarchy pyramid as presented in [35]

2.3.2 About Trustworthy AI and Responsible AI

The distinction between Trustworthy and Responsible AI is not immediately clear when looking at literature, as Schiff et al. in [40] point out that the terms trustworthy, ethical, and responsible AI have been used interchangeably. However, to create some clarity in this thesis, we have decided to simply choose two definitions from the literature that make sense to us:

- **Trustworthy AI** | tr∧st wur:zi ei-ai | *noun*: An Artificial Intelligence system that meets different ethical and institutional requirements and has become worthy of trust. [14].
- **Responsible AI** | ris'pon:suh:bel ei-ai | *noun*: Artificial Intelligence that is developed in a process that "truly considers all humankind when determining the purpose of the system". [7].

As one can see in the definitions we chose, the requirements for Responsible AI go a bit further than Trustworthy AI, as the AI system does not just have to be ethically acceptable but actively seeks to be societally beneficial. This vision on Responsible AI is explained in [7] and resonates with the related concept and requirements of Responsible Reseach and Innovation (RRI) as described by Stilgoe et al. in 2013 in "Developing a framework for responsible innovation" [45].

Given this more stringent requirement for Responsible AI, it is not a given that all HCAI systems should adhere to those extra requirements, as commercial parties are, for example, not expected to be model citizens as long as they do not break any laws. However, given our context of humanitarian decision-making, we do think it is relevant to add some of the requirements for Responsible AI to our HCAI requirements, given their public values.

2.3.3 The hierarchy

As can be seen in the definitions below, most of our requirements stem from documents and papers on Trustworthy AI, as some of the most influential requirements documents, such as the EU guidelines on Trustworthy AI [14] have chosen this term as their objective. According to [20], the most overlapping requirements from publications on Trustworthy AI are transparency/explainability, justice and fairness, non-maleficence/societal and environmental well-being, responsibility/accountability, and privacy. These terms already bring our need forward to differentiate them, as some of those requirements are rather values, while some are definitely more norms or design requirements. In Figure 2.2 we have placed these terms, as well as other important terms from Trustworthy and Responsible AI and related them with each other through "for the sake of" relationships. In the sections below, one can find an explanation for each of the values, norms and design requirements.

Values

Before we discuss the values in this section, it is important to note that those may not be universal for everyone. While (western) philosophers have been active in trying to find more universal judgments on what can be classified as values and which are intrinsic or extrinsic values [55], because of the subjective nature of what we deem as "good", it differs between people, cultures and institutions on what they classify as values. As such, the whole diagram rather functions as a way to start and structure our conversation, than as some form of absolute truth.

- **Privacy:** The right to privacy is the right to be let alone in one's private and family life. [9]. This covers a wide range of aspects, such as a person's intimacy, identity, name, gender, honour, dignity, appearance, feelings and sexual orientation. Many have argued that privacy is an intrinsic value [30].
- **Societal and environmental well-being:** The golden rule: any innovation should be developed in the end to enhance and promote the well-being of (preferably) all human-beings, including future generations [14].
- **Fairness:** To deal fairly is to show no bias towards some people or individuals: people should be treated equally in a situation if their characteristics that are relevant to the situation are similar. []
- **Justice:** A core definition in ancient Roman law describes it as "the constant and perpetual will to render to each his due". This inhabits both the requirement of fairness, as well as a broader societal perspective that all people should have the same liberties and opportunities, as long as those liberties do not harm others. [28]. While it is evident that this is an instrumental value, philosophers such as Socrates and Kant have also given arguments for its classification as intrinsic value.
- **Safety:** This refers to the "No Harm" principle from the EU guidelines for trustworthy AI [14]. Any trustworthy AI system should be developed and governed such that hazards and conditions leading to physical, psychological or material harm are controlled.

Norms

- **Data governance:** As defined by [14], data governance as norm covers two important aspects. The first aspect is the protection of personal data: the AI system should adhere to the rules as set out by the GDPR and should prevent data leakage and unlawful use of personal data. The other aspect sets boundaries and guidelines for the data that is used to inform the AI system. This not only refers to the quality and integrity of the data that is used, but also whether the data is relevant and fair for the intended use case.
- **Technical robustness and safety:** Originating as well from the EU guidelines [14], technical robustness and safety describes a norm for a preventative approach to AI development where risks are actively anticipated and minimized or mitigated. This means it should be protected for adversarial attacks, as well as for unexpected or unintentional harm. The EU specifies this further as that an AI system should be accurate, reliable and reproducible, and have general safeguards implemented for

unexpected cases. This form of active anticipation also aligns with the RRI anticipation dimension as described by Stilgoe [45].

- **Do no harm:** This norm is named as a principle for trustworthy AI by the EU guidelines [14] It means that AI should "neither cause nor exacerbate harm or otherwise adversely affect human beings" [14]. This means AI systems must be technically robust, and that special care should be taken to prevent harm to vulnerable persons or groups by including them in the development process. Do no harm is also an already widely established norm within the humanitarian sector that applies to all facets of their operations.
- **Economy of resources:** As we live in a world of limited resources, these must be wisely spent. This holds double in the context of humanitarian help, where every dollar that goes into bureaucracy is a dollar not spent on precious aid for human beings. Therefore, a good AI system should promote an efficient use of time and money.
- **Acceptance:** This term is included as a principle for trustworthy AI by [20]. It represents the willingness of a user to use the system and works as a mechanism to ensure that an AI tool actually meets the needs and requirements of its intended use-case.
- **Inclusion:** Inclusion, as defined by the Responsible Research and Innovation (RRI) framework by [45], is a norm for the development process. It stipulates an active effort to include the relevant stakeholders, values and considerations throughout the innovation process.
- **Preventing biases:** This is named in most papers as a norm for trustworthy or responsible AI, as it is a strong requirement to ensure fairness in AI systems. This means careful scrutiny and mitigation of any bias in the training data, as well as using diverse test sets to assure no bias can be found in the actual machine learning model. While Shneiderman in [43] advocates for a dedicated bias testing leader in every development team, we see also here the added value of inclusion as a preventative method against biases by including a diverse group of stakeholders in the development process.
- **Accountability:** To ensure justice, there needs to be some form of accountability in the decision-making process. To be accountable, one must be able to bear responsibility for a decision and give a satisfactory reasoning for them. One can, however, only be morally responsible for a decision if one satisfies two conditions: they had to have adequate control over the decision, e.g., have freedom of action, and they had to have awareness of the consequence and moral significance of the decision [38].
- **Human agency and oversight:** This term is used as one of the guiding principles in the EU guidelines on trustworthy AI [14]. All AI systems should support human autonomy and decision-making. This covers the idea from HCAI that AI should augment rather than replace human decision-making.

Design requirements

We have a rather short list of design requirements, as including all the specific system requirements (such as accuracy, precision or usability), and all the specific explainable AI requirements (such as algorithmic transparency, explanation effectiveness and scrutability) would create quite a large graph and would be impossible to make fully complete without further months of research. We have therefore compiled all the technical robustness requirements under the term "Good software engineering practices" and all the data quality and privacy requirements under "Responsible data management". Two design/method requirements that we would like to further specify here however, are transparency and explainability as they are specifically named as main principles by [20]. For Human-Centered Design, we will dedicate a separate section afterwards as it will be the basis for the further methodology of this thesis.

- **Transparency:** Transparency is a term used within different domains and has different definitions depending on which scope it is applied to [23]. For clarity, we have made a distinction between transparency of the whole process, and algorithmic transparency which is a specific attribute of an AI. We refer here to transparency of the whole process, and define its scope as an overarching requirement to be transparent in all phases of the AI life-cycle and its practical use within the institutional context.

As transparency in itself is not an absolute value or norm (do you need to be transparent to everyone?), we have put it here as a design requirement to emphasize its function as an instrument rather than an important principle in itself.

- **Explainability:** In [20], explainability is described as a need to be able "to communicate the reasoning for the AI system's decisions to different stakeholders". For this, transparent AI systems are needed, as well as transparency about the training data and the metrics that are used. As explainability needs to be specified in order to be relevant (what do we need to explain to whom?), we also see this as a design requirement to achieve human oversight and accountability.

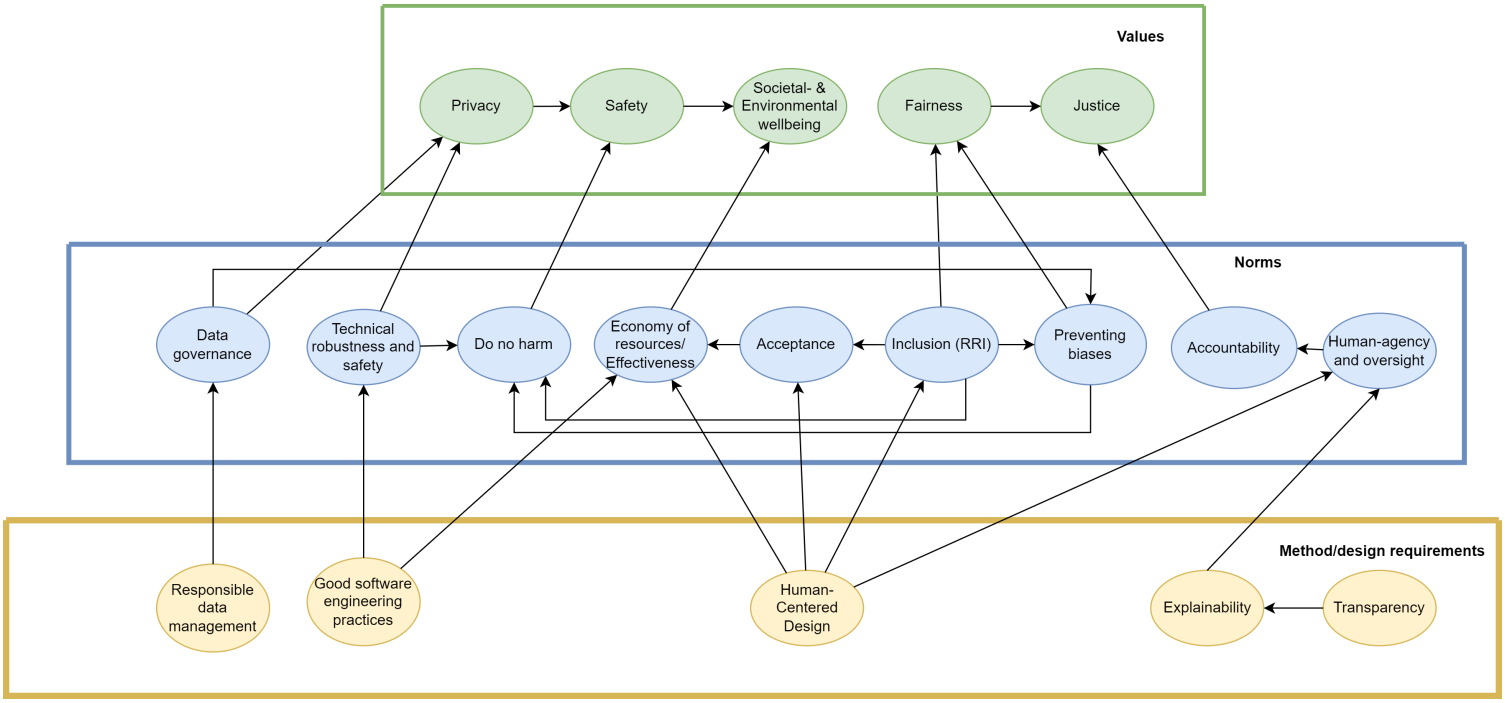


Figure 2.2: Value hierarchy for trustworthy AI.

2.3.4 Limitations

The value hierarchy we present here has some limitations that are important to address. Some of those limitations are solvable when extra work is put into it, but some will remain inherent in this format.

The first limitation pertains to the terms and clusters used in the hierarchy. As we have tried to, where possible, directly copy terms from literature to retain clear reference to these sources, some terms might be overlapping and encompass too many different aspects to make the hierarchy very clear. "Data governance" could for example in the future be better split into the different aspects it entails, as it now also partly overlaps with "preventing biases". But in order to systematically create such an overview, one would need to encode all the different documents on responsible and trustworthy AI, which is an undertaking outside of the scope of this thesis.
Another limitation in our hierarchy is the manner in which we have drawn relations between norms and values. It is important to state here that the relations we draw between different norms and values are not presented as an absolute truth, and there are many possible remarks to make about it. Many more norms and values could for example have been related to each other than the connections we have determined here. We have chosen however to be restrictive in the lines we draw, to give more focus to the relations we see as important to consider (and to prevent, in all honesty, a complete spaghetti of a diagram). Depending on one's point of view and background, this prioritization could be done completely differently. One could

for example question our choice for not drawing a line between Fairness and Justice and Human Wellbeing. Our argument is that, theoretically, a world might be possible that is unfair but where people are not per se unhappy, such as is demonstrated in Aldous Huxley's book "Brave New World". Others might point out that in practice, it is highly unlikely to obtain general Human Wellbeing without Fairness and Justice and argue for the importance of that relation. This shows that the connections are more important as a tool to start having a conversation within an organization on where the priorities lie and how they view the relationship between their requirements, norms and values rather than to find one epistemic true version of them.

2.4 Human-Centered Design

Human-Centered Design (HCD), is a long-standing innovation approach that bases its design decisions on interaction with the intended users and that has contributed to many successful business cases in the past decades [15]. This was also the reason we chose to start with it: due to its maturity as a framework and widespread use, it gives us the right methods and backbone to do HCAI in practice, as it is informed by people how know what is needed in practice.

Its standard approach consists of three [15] or four [24] main phases describing the main steps any human-centered design project needs to undertake. To make it work for our project, we have combined concepts from both the IDEO.org field guide to Human-Centered Design [15] and the guide from the Harvard Online Business School [24] to come to the following definition of the phases:

1. **Clarify:** This phase is about understanding the problem and identifying the users' needs. Activities in this phase centered around framing the design challenge, gathering information on the problem, and interviewing different stakeholders.

2. **Ideate:** The ideation phase is about generating ideas, identifying opportunities and refining solutions based on our gained insights. This phase does not only consist of brainstorming, but also of defining design requirements and soliciting feedback from stakeholders. When done well, it is characterized by multiple cycles of ideating based on the feedback received from the stakeholders.

3. **Develop:** In the developing phase we prototype and evaluate our chosen solution. In this phase, it is best to do multiple rounds of prototyping. There should looked sharply at the developed prototype/solution to determine whether it is desirable, feasible and viable.

As one can see, each of those phases are marked by different goals and different (design) activities. The activities used to fill in each of the phases can be mixed and matched to accommodate the specific needs of the project in case.

Through HCD's system of getting user input during each phase of the development process and their focus on centering the product around the users needs, its method helps in achieving different norms identified in the value hierarchy. When done well, it helps to promote human agency by aligning the AI tool to the user's explanation and information needs, it helps inclusion by including different stakeholders into the design process, it creates acceptance by focusing on creating a solution that is actually needed and wanted by the people who are expected to use it, and it helps with economy of resources by creating products that are actually needed, instead of something that is afterwards put on the shelf because it didn't serve the right purpose.

In the next chapter, we will describe how we created our research plan based on Human-Centered Design, and explain how the rest of the thesis is structured according to this plan.

Chapter 3

Our case study: a bird's eye view

In this chapter and the following four chapters, we will describe our case study with the IPC, which answers RQ1: "How can we design an HCAI-supported monitoring process for Acute Food Insecurity (AFI) analyses that explicitly takes the validity of ongoing assumptions into account?" and RQ2.2: "How can we translate our identified HCAI requirements into our design process? However, before we delve into each part of the case study, we will give a bird's eye view of our research process and compare it to our original plan. In this way, the reader can understand how the following chapters came to be and how our understanding has evolved since the conception of this project.

After the plan was made to look into data-driven risk factor monitoring for the IPC, our research started in April 2023 with a wider exploration of the organisation, their analysis process and their monitoring practices. These explorations were initially based on going through the extensive documentation of the IPC, such as the IPC Manual [19] and their self-learning courses on Acute Food Insecurity [18]. However, these sources were not sufficient in giving us a full picture of how the analysis process works in practice, or how they used its outcomes for risk factor monitoring afterwards. Our other first source in this initial exploration was, therefore, our contact person at the IPC: a product development manager who supported us from their side. During the whole process, we have had many interviews with him to get a better insight into the workings of the IPC, the problems that needed to be solved in relation to monitoring and any other questions that came up during our research process.

Based on that initial exploration, as well as the insights gained from our literature study on HCAI and HCD, we developed a first set of research questions and an HCD-based research plan to answer them. Our objective: develop a new monitoring system for the IPC using computational modelling methods to assess the validity of its underlying assumptions. The research plan (see Figure 3.1) consisted of four phases as described in Section 2.4.

In the first phase, "Clarify", one can see that next to doing a literature study on Food Insecurity theories and models, the main activity for this phase was interviewing IPC facilitators. IPC facilitators are people with a high level of IPC training who are responsible for guiding IPC analyses. There was a high need for these interviews, as we found out during our initial exploration that there were many unknowns about how IPC procedures/prescriptions were implemented in practice: the manual might describe the steps that needed to be taken and the decisions that had to be made, but in many cases didn't specify how, by whom and when these decisions were made. Moreover, these specifics were also not always known by our contact person, as he might have an inkling of how this had worked in specific cases he was involved with, but could not say up to what point this could be considered general practice. Therefore, in Chapter 5, we describe how we planned and executed these interviews and present a resulting overview of how the IPC analysis cycle works in practice, how assumptions and risk factors are developed within this cycle, and what their monitoring practices are.

In the next phase, "Ideate", we used the inputs gained from the Clarify phase to develop a high-level plan

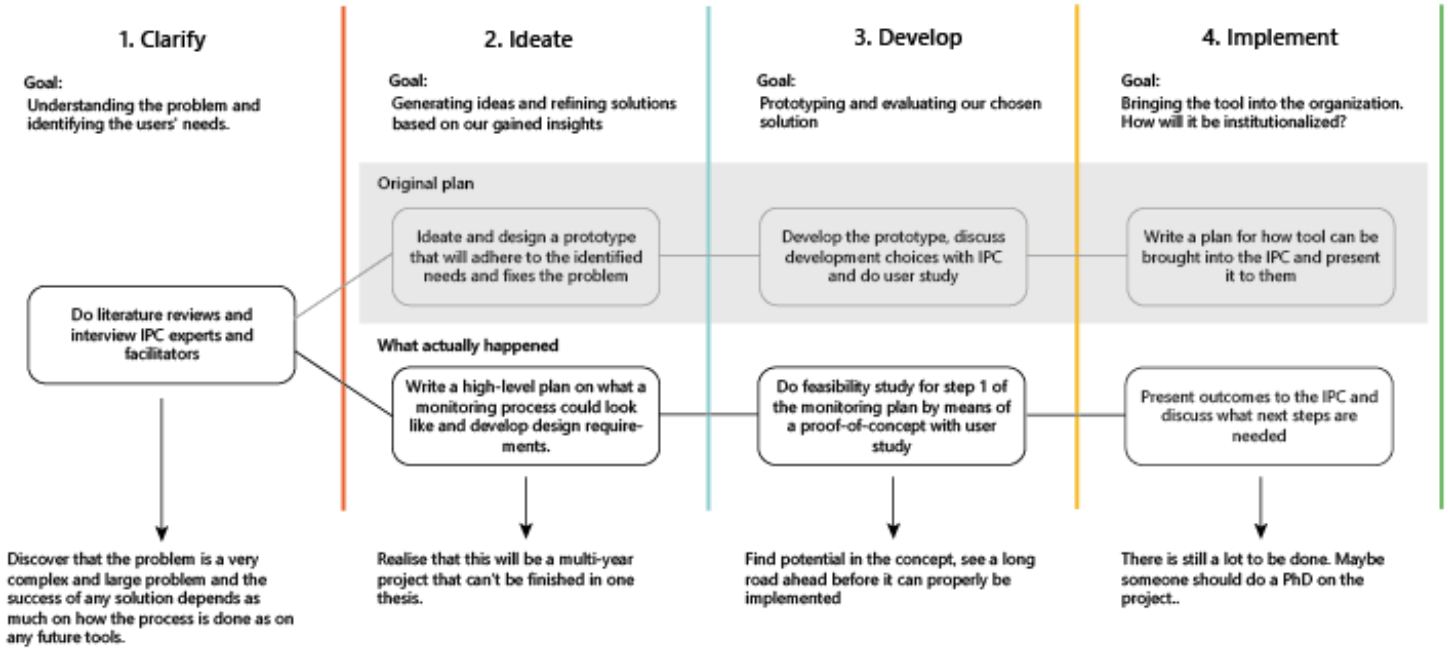


Figure 3.1: Original research plan compared to the evolved research process.

for how the IPC might implement a more structured and data-driven monitoring system. The results of this phase can be found in Chapter 6. One will find there that the outputs of this phase are much more high-level and process-based than we had originally planned. This came from the realization that the problem was not just a technological problem but a process/human-factors problem as well, and a very complicated one at that. That is why, when we had developed this plan, we decided that it would not be useful at this point to continue with our original plan of developing some data-driven monitoring system immediately.

Instead, we chose during the "Develop" phase to start investigating whether the first step of our proposed plan was even possible: letting analysts quantify analysis assumptions using thresholds. In this phase, we have, therefore, developed a prototype for quantifying projection assumptions and tested it with IPC analysts as proof of concept. This phase and experiment are described in Chapter 7.

In conclusion, our original research plan was, while still trying to take all the organization's needs into account, solely focused on creating one technological product that would be used as a monitoring tool. What we learned in the coming chapters is that it is, in this case, not possible to ignore the organizational and human aspects of the solution, as is in many sociotechnical systems. With this we mean that it is not just enough to take human needs into account in the technological design; you need to design the surrounding process and determine the boundary conditions for your system as well.

Chapter 4

Literature Study II: Food Insecurity Theory and Modeling

In this chapter we will dive deeper in to RQ1.1 of our research: "How does the IPC analyse Food Insecurity and what computational models are currently available for predicting it?". In the first sections, we will discuss the theory behind food insecurity, and we will end the chapter with an overview of previous approaches to machine learning for food insecurity and their suitability for our monitoring purpose.

4.1 Food Insecurity

Food Insecurity can be present in different degrees. According to the Food and Agriculture Organization of the United Nations (FAO) (one of the partners of the IPC), when a person is food insecure, it means they lack consistent access to the amount of safe and nutritious food they would need to have normal development and a healthy life [10]. This does not mean that they are necessarily physically hungry; when people are forced to limit their diet or engage in unsustainable coping strategies such as selling household items, this can be an important indicator for food insecurity as well.
The IPC uses the analytical framework in Figure 4.1 to analyze and classify the level of food insecurity for a given context. This analytical framework consists of:

- **Contributing factors:** forces that determine the level of food security (the combination of causal factors and food security dimensions)
- and **Food security outcomes:** expected manifestations of food (in)security (the combination of second-level and first-level outcomes).

In the following sections, we will explain each element of this framework and give the reader a broader understanding of the dynamics of Food Security.

4.1.1 The dimensions of food insecurity

Food security is dependent on four sequential dimensions of food security [19]: food availability, accessibility, utilization and stability. These dimensions can all function as a bottleneck in the nutrition pipeline of households: food availability is the most obvious cause for food insecurity, but even if there is food available, without accessibility or proper utilization there will still be food insecurity. In the following paragraph we will explain each dimension in further detail.

1. **Food availability:** *Is there food in the fields or in the market stalls?*
   The first factor impacting food security is whether food is actually physically present to purchase or harvest. This dimension is dependent on food production, availability of food reserves, availability of imports, transportation to markets, and wild foods.

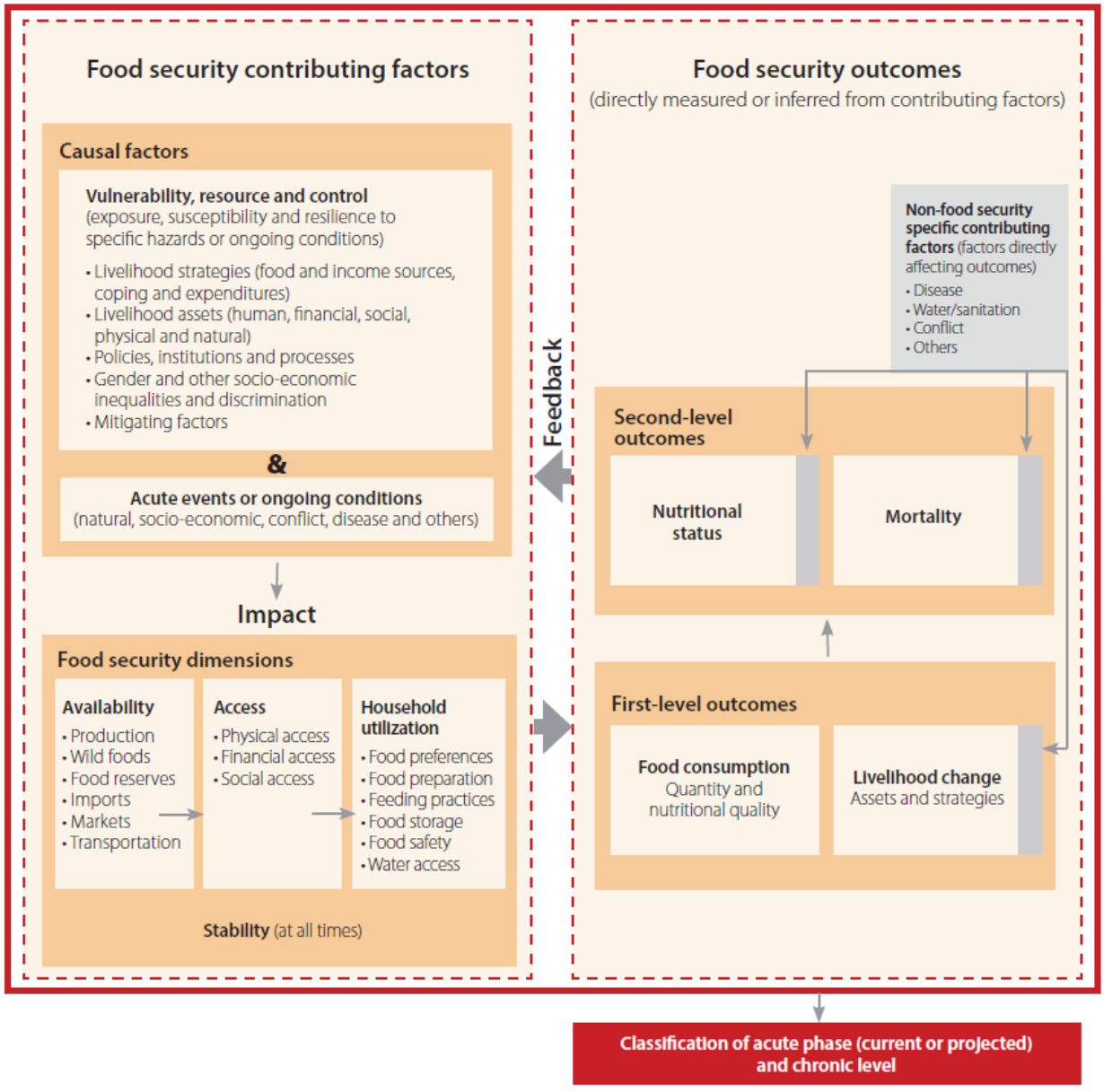


Figure 4.1: The IPC Food Security Analytical Framework, taken from the IPC Manual [19]

2. **Food accessibility:** *Can I reach the markets and actually buy the food there?*
   The next limiting factor is whether households are able to acquire enough food. This dimension is an important addition to the availability dimension, as the presence of food does not imply that all households in a region are able to access said food. This dimension is determined by three factors:

   - *Physical access:* What is their distance to markets? Or are households able to produce or gather their own food?
   - *Financial access:* Do households have enough purchasing power to acquire food at the markets?
   - *Social access:* Do households have a social network through which they can obtain food in times of scarcity?

3. **Food utilization:** *Is the food adequately stored and prepared?*
   The next factor impacting food security is whether households are optimally using their resources to obtain an adequate diet. This can be affected by: food preferences, food preparation, storage availability and access to potable water. Inadequate storage can, for example, lead to high post-harvest losses, resulting once more in food insecurity.

4. **Stability over time:** *Are people food secure for the coming period?*
   An overarching factor for all these dimensions to take into account is whether the system is stable. Instability is both a driver and a consequence of food insecurity [53]. That food insecurity develops when there is conflict or economic problems, is evident. However, it is also a driver for major societal instability. There have been numerous examples, from food riots toppling the governments in Madagascar and Haiti to the social unrest leading to the Arabic Spring, that were influenced by growing food insecurity. A survey from the UN World Food Program (WFP) [53] identifies three main sources for food insecurity and instability: agricultural resource competition (e.g. competition for land and water), market failure (e.g. food price spikes, price uncertainty and price volatility) and extreme weather (e.g. droughts and floods).

4.1.2 Causal factors

In discussing the food insecurity dimensions, we already touched on some causal factors for food insecurity, also called drivers in IPC context. How much a population suffers from food insecurity is not driven by natural, political or economic threats alone but also by the vulnerability of the population. A drought is, for example, differently experienced by a farmer whose main source of nutrition is their own grown crops than by someone who works in the city as a doctor and can easily get food from other sources. Therefore, the IPC considers causal factors (or drivers) of food insecurity as an interaction between hazards (often referred to as "acute events or ongoing conditions") and vulnerabilities [19].

While the IPC does not give an explicit definition of hazards in the IPC manual, in their self-learning course on acute food insecurity [18], they define hazards as "phenomena that have happened or may happen in the future" and include "acute events or ongoing conditions which can be natural or human-made". Here, they name "droughts, sharp price spikes, war" and other events that can impact acute food insecurity as examples of hazards. However, according to the EU expert working group on disaster risk reduction terminology [31], hazards do not cover economic disturbances or political conflicts. Therefore, we will work with the term "shock" to cover a wider range of hazardous events. We follow the risk categorization [51] of the World Economic Forum to cover this broader range:

- Economic events: energy or food shortages, sharp price increases, prolonged recessions, debt crises, etc
- Environmental events: natural disasters, extreme weather, human-made environmental damage, etc.
- Geopolitical events: political instability, interstate conflicts, terrorist attacks, economic warfare, riots, etc.
- Societal events: chronic diseases and epidemics, lack of public infrastructure, societal polarization, involuntary migration, rising unemployment, etc.

- Technological events: breakdown of critical information infrastructure

The IPC measures the potential impact of a shock through its severity, magnitude and occurrence or probability of occurring. How much a shock actually impacts a population in terms of food insecurity depends on their vulnerability, which comprises, according to the IPC manual of three factors: exposure, susceptibility and resilience. However, the manual subsequently does not define these terms, as they state that their main method of doing vulnerability analysis is through analysing livelihood strategies, livelihood assets and the effects of policies, institutions, gender and other mitigating factors affect their ability to successfully respond to shocks.

To better understand how these two approaches relate to each other, we will first define exposure, susceptibility, and resilience before moving on to the IPC's view on vulnerability analysis.

- **Exposure:** Following [29], "exposure can be defined as all the elements at risk from the shock under analysis", specifically according to [5]: "human beings, their livelihoods, and assets".

- **Susceptibility:** This term relates to how social, institutional and environmental conditions make some people or communities experience higher impacts of a shock, according to the UN Office for Disaster Risk Reduction (UNDRR). [47] [1]

  The level of susceptibility of a population depends on many different variables. These can be clustered according to the UNDRR [47] in four dimensions[2]:

  1. Physical factors: poor construction of buildings and infrastructure
  2. Social factors: poverty and inequality, marginalisation, social exclusion and discrimination by gender, social status, disability and age (amongst other factors) psychological factors
  3. Economic factors: the uninsured informal sector, vulnerable rural livelihoods, dependence on single industries, globalisation of business and supply chains
  4. Environmental factors: poor environmental management, overconsumption of natural resources, decline of risk regulating ecosystem services, climate change

  As those factors are often correlated with each other for certain socio-economic groups, this means that the impacts of hazards and shocks are often felt the most by the same groups. The most important category here is people living in poverty, as research has shown that they tend to be the group that suffers the most when shocks occur [57]. Moreover, on an individual level, we see that women, children, the elderly, the disabled, migrants and displaced populations are also much more often susceptible to food insecurity.

- **Resilience:** This refers to the ability of the population to adapt to a shock and recover from it in a timely manner [31]. It is not only dependent on the capacity of a system in terms of strengths and resources to cope with a shock, but also how they are effectively used by social, institutional and informational services. Resilient systems are characterised, among other factors, by a high level of diversity in terms of access to assets and economic opportunities, a level of redundancy where some areas can fail without this leading to a total collapse of the system, and social cohesion and support within communities [46].

The IPC's primary focus is on vulnerability in the context of Food Insecurity. Consequently, their analysis of resilience and susceptibility is constrained to this domain. Furthermore, as susceptibility and resilience are closely related to each other (some experts consider them two sides of the same coin [46]), the IPC analyses these terms together on a granular level by examining three characteristics of the target population in relation to food insecurity:

[1] The UNDRR uses the term "vulnerability" to describe susceptibility and sees exposure as a separate factor from vulnerability, but in order to align with the framework of the IPC, we will adhere to their terminology instead.

[2] This list or set of dimensions is, however, not exhaustive. In the literature review of [29], they found 10 dimensions and many more factors to consider.

1. Their livelihood strategies: how do they obtain food and income? What are their common coping strategies should their normal way of obtaining food be obstructed?
2. Their livelihood assets: what resources can they turn to in times of scarcity? These assets can come from different sources [18]:
    - Human: literacy, professional skills, educational level
    - Social: supporting network, extended family
    - Financial: savings, access to credit and loans
    - Physical: tools and equipment, housing, livestock, stores
    - Natural: rivers, lakes, pastures, wild foods
3. Their societal and institutional context: How do policies and governance positively or negatively impact their ability to cope? This is an important factor, as a governments ability to properly respond to a shock can significantly influence the amount of food insecurity and instability endured by their people [53]. Moreover, whether there is high inequality, marginalisation, social exclusion or discrimination happening, has also a high impact on whether vulnerable groups can properly adapt to shocks.

This detailed level of inspection is made possible by the extensive primary survey performed before each IPC analysis. In this survey, evidence is gathered not only on the direct level of food insecurity but also on the contributing factors.

4.1.3 First-level Food Security outcomes

The actual manifestation of food insecurity at the household and individual level is determined by the IPC essentially through two types of primary outcomes: the food consumption levels of the households and the degree of livelihood changes happening. These outcomes are measured during the primary survey by querying people on their food insecurity levels.

Food consumption is the most important indicator for FI and can be lacking in two ways. The most straightforward method of examining it, is to see whether people have enough energy intake each day to sustain themselves and their way of living. However, especially for chronic food insecurity, the diversity of food intake (micronutrient balance) is also important [19]: a diet of solely potato's might give enough calories, but will lead to serious vitamin deficits and other health problems if not complemented with other food sources. Food consumption can be measured with different indicators. Three indicators that are often used are:

- **Food Consumption Score (FCS):** This score is based on a household's self-reported consumption frequency of 8 different food groups over the course of a week. It stems from the WFP and is collected in all assessments and monitoring activities. [19]
- **reduced Coping Strategies Index (rCSI):** Information for the rCSI indicator is collected by asking households with which frequency they have used 5 different food-based coping strategies in the past 7 days. The IPC specifies that this indicator is more useful in onset crises when households are starting to respond to shocks, rather than in longer emergencies when some of their coping mechanisms have likely already been exhausted. [19].
- **Household Hunger Scale (HHS):** The HHS is based on self-reported information from households on whether they have experienced problems in accessing food in the last 30 days.

Livelihood change is another significant factor in assessing the degree of food security. When people are eating enough but obtain this food by selling their livelihood assets such as cattle, this means that the worsening of their situation is imminent as they are surviving through **unsustainable coping strategies**. Unsustainable coping strategies do not only relate to the selling or consuming of livelihood assets, also activities that are detrimental for one's health or the discontinuation of education can be seen as unsustainable coping strategies. Livelihood change is often measured by the **Livelihood Coping Strategies (LCS)**

indicator, which assesses a household's experience with livelihood stress and asset depletion due to a lack of food or money in the past 30 days. The questions for this indicator need to be adapted to each local context [19].

4.1.4 Second-level Food Security outcomes

When inadequate food consumption, combined with negative livelihood changes and other non-food-security-specific contributing factors, persists, it can lead to second-level outcomes in the form of **malnutrition and mortality**. These outcomes are often sequential and contain some time lag, as lower energy intake will first lead to malnutrition, and then eventually, combined with disease, will lead to higher mortality rates. However, as these factors do not necessarily have to be the direct effect of food insecurity as they can be caused by other non-food security contributing factors as well, they are rather used to support classifications than as driving factors. [18].

4.1.5 Convergence of evidence

In the IPC analysis process, they use this analytical framework to gather data on all the different elements, which are then examined and compared in a consensus process. They call this process "convergence of evidence", as they combine all the pieces of information to get a clear picture of the level of food security in a specific area. When this process is completed, they should be able to answer the following questions about food insecurity in a specific area:

- **How severe** is the situation?
- **When** will populations be acutely food-insecure?
- **Where** are the most acutely food-insecure people located?
- **How many** people are acutely food-insecure?
- **Why** are people acutely food-insecure?
- **Who** are those most acutely food-insecure?

4.2 Food Insecurity and computational models

4.2.1 Criteria

We have inventoried which models are available for our monitoring purposes. To this end, we developed several criteria based on our initial exploration to which a usable food security model should adhere to:

- **Now/soon-casting:** Predictions are only relevant for now-casting FS and predicting it for maximally 6 months in the future, else, it is not relevant to use for our monitoring purposes as its predictions will coincide with the next IPC analysis. Also, we want to monitor new developments that were not foreseen during the analysis.
- **Data availability:** For a model to be useful, it should run on data that is consistently accessible during monitoring periods. Preferrably, only data is used that is publicly available or otherwise at least accessible for the IPC. This means that data that stems, for example, from Living Standards Measurement Surveys (LSMS) or that is based on daily outcome indicators is not considered as accessible as it is rather the exception than a rule that such data is available for a country.
- **Data quality:** The model should be trained on target data based on primary outcome data. By that, we mean that the target data should be based on surveys that have actually been done in the field to assess the level of (different aspects of) food security in different households. This means that food security indicators such as rCSI, HDDS and FCS (see Section 4.1.3) are acceptable target data, as well

as actual classifications done by the IPC, as those are always based on outcome indicators next to data on contributing factors. However, we are making this point about primary data, because 50% of the algorithms found in the papers below base their algorithms on FEWSNET IPC data, which is, while in the same format, a very different kind of indicator that is not based on primary data. FEWSNET is a famine early warning system funded by the U.S. Agency for International Development (USAID) that assesses food security through scenario development based on contributing factors. Because they are not dependent on the time- and resource-consuming practice of gathering primary data, they are able to do their food security analyses four times a year, which makes them a more attractive option to many modelers through the larger amount of data they can provide. However, as their outcomes are not based on data collected in the field, it is not a given that their outcomes reflect the actual state of food security during those periods, which is why the IPC does not count their data as valid target data.

- **Spatial resolution:** As Food Insecurity can differ quite substantially between different regions in a country, it is important that a model can give results on some form of subnational scale in order to give relevant and somewhat trustworthy results.
- **Model performance:** Naturally, a model should also have a high level of accuracy and precision in order to be useful. However, comparing the results of these different models is quite hard due to their different contexts, and training and target data. Therefore, we have left the consideration of their performance out of the scope of this thesis and first focused on whether any model was even suitable based on its other characteristics.
- **Explainability & interpretability:** The IPC needs to be able to account for all their decisions surrounding food insecurity. Given the high-impact and political nature of IPC decisions, we put here an extra stringent note on what we define as *explainable*: models that can have explanations that are both *sound* and *complete*, or in other words, models that are inherently interpretable. This means that we do not see any form of black box model as an explainable model. As Rudin namely explains in her seminal Nature article "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead" [37], post-hoc explanation models for black-box algorithms cannot perfectly explain the algorithms (else, they could just replace the original black-box algorithms). They will therefore always give explanations that either do not cover the full reasoning of the black box (are incomplete) or will give a full reasoning by approximating the black box but are therefore not guaranteed to be truthful (are unsound). Both types of explanations can mislead the users and give a wrong impression of what the model is doing and therefore not make any decisions made with these black boxes truly explainable.
- **Alignment in practice:** preferrably, these models are based on the same theoretical framework as IPC bases their classifications on
- **Possibility to incorporate expert input:** As we saw in Section 4.1.2, Food Insecurity can in different countries be driven by different phenomena, depending on the vulnerabilities of their population and the shocks they experience. It is therefore important that a model can be tweaked to the context of a country (or even subnational regions), which would be done ideally by the local experts of that country who have the best understanding of the food security there. Moreover, as explained for example by [50], given the scarcity of food security outcome data, allowing some form of expert input can also help make the model more robust/stable to make up for the limited usable data points, as well as make it possible to account for unprecedented events of which there is no historical data yet, such as COVID19 used to be. An important consideration here however, is whether it is best to use one large flexible model that can be tweaked to different countries, or choose for each country a simpler more specific model that is tailored to their type of food insecurity. Therefore, we will also note for each model, where applicable, for which type of contexts they function.

4.2.2 Comparison

To compare the different Food Insecurity machine learning models that might be useful for developing our monitoring system, we created a table highlighting whether they satisfied our different criteria. Some of the

columns of this table are taken from an as of now yet unpublished literature survey of Mélissande Machefer on machine learning for Food Insecurity. We used this paper as well to find the relevant food insecurity models in the first place. The short summary version of our table can be found in Table 4.1.

Findings:

As one can see in the right-most column, none of the models we analyzed managed to meet our requirements. This would even remain the case if we were to remove our criterion for expert input. The main reasons for this unsuitability are that:

- More than half of the models are not interpretable due to their use of black box models;
- Half of them are based on FEWSNET data and do therefore not meet our requirements for using primary outcome data as target;
- Not all of them have accessible training and target data.

Some models that did show promising directions, though, were the models of Wang et al. [50] and Krishnamurthy et al. [21]. Wang et al. used Bayesian priors to incorporate expert input. Given the versatile ways in which priors can be used to adapt Bayesian models, we found this a very promising approach for incorporating expert input. However, given the complexity of probabilistic theory, special consideration should be given to how to elicit those priors from domain experts.

Krishnamurthy et al. stood out through their careful consideration of how to model droughts and showed that, when applied to the right areas, droughts in combination with food prices were already quite good predictors. What we like about this model is that it is simple and context-driven. E.g. it chose to target a specific form of food insecurity (namely drought-driven FI) and made a simple model with still quite some predictive power. We foresee that such a model could be more easily used as an extra information source during monitoring, where the expert input for the model does not lie necessarily in changing the model, as much as deciding whether the model is applicable and usable for a specific context.

Nonetheless, this short inventory did show that there was quite some work to do for us if we wanted to develop our HCAI driven monitoring system.

Paper	**Predicts**	**Trained on primary outcome data**	**Model**	**Inter-pretable**	**Subnational spatial resolution**	**Accessible data**	**Expert input**	**Context**	**Fulfils criteria**
Lentz et al, 2019 [25]	rCSI, HDDS, FCS	Yes	LASSO	Yes	Yes	Partly	No	1 country	No
Andree et al, 2020 [2]	FEWSNET IPC: P3+	No	Random Forest	No	Yes	Yes	No	Globally	No
Deleglise et al, 2021 [6]	FCS, HDDS	Yes	CNN, LSTM, RF	No	Yes	Yes	No	1 country	No
Westerveld et al, 2021 [52]	FEWSNET IPC: state transitions	No	XGBoost	No	Yes	Yes	No	1 country	No
Zhou et al, 2021 [60]	FCS, rCSI	Yes	Random Forest and Gradient Boosting Trees	No	Yes	No	No	3 countries	No
Krishnamurthy et al, 2022 [21]	FEWSNET IPC phase changes of ≥ 5 months	No	Autocorrelation + sigmoid	Yes	Yes	Yes	No	Regions with drought-driven food insecurity	No
Martini et al, 2022 [27]	FCS, rCSI	Yes	XGBoost	No	Yes	Yes	No	globally	No
Wang et al, 2022 [50]	FEWSNET IPC: P1,P2,P3+	No	Panel vector-autoregression	Yes	No	Yes	Yes	globally	No
Foini et al, 2023 [12]	FCS (on a daily basis)	Yes	Gradient boosted regression trees	No	Yes	No	No	6 countries	No
Penson et al, 2024 [33]	FEWSNET IPC: P4+	No	Linear classifier and GLM	Yes	Yes	Yes	No	1 country	No

Table 4.1: Comparison of different Food Insecurity machine learning models

Chapter 5

Monitoring assumptions in the IPC analysis cycle: theory and practice

As indicated in the previous chapter, in this chapter, we will discuss the interviews we conducted in the "Clarify" phase of our research and present an overview of how the IPC analysis cycle works in practice, informed by those interviews.

5.1 Developing the interview guide

Before approaching our interviewees, we went through a careful preparation process to ensure that we would get the most out of each expert interview. This meant gathering all the information about the IPC that we could find in advance and creating a fine-tuned interview guide. With this interview guide, we were able to structure and prioritize the information we wanted to know and ensure that we would have a coherent picture of the IPC process afterward. In the following sections, we will take a closer look at the challenges and considerations that brought the interview guide to its current form.

5.1.1 Understanding the process

As discussed in Chapter 3, there were many things unclear to us at first about how the IPC analysis works. This made the development of an interview guide initially quite difficult, as it is hard to ask purposeful questions about a process if one doesn't understand the process itself yet. The interview guide had to go through several iterations therefore before we reached a well-targeted and cohesive conversation line. The difficulty in understanding the analysis process and developing the interview guide lay in several factors.

The first factor was **jargon**: the IPC methodology is highly analytical, domain-specific, and procedural, resulting in an analysis process filled with IPC-specific terms and phenomena. Therefore, it took quite some time for us as outsiders to make sense of what was specifically meant by different terms and to know how to phrase the questions in the same proper language so the facilitators would understand. Sometimes, this jargon confusion would also be exacerbated when people from the IPC didn't realize in discussions that some of their terms were not known by the outsider parties yet.

The second factor was the nature of the process: it's simply a **highly complex system**. There are many different actors, many different steps, and many different factors that need to be determined at different times throughout the whole process. Each time we thought we understood the main dynamics of the process and created an overview or a new version of the interview guide, we would learn through feedback from our IPC contactperson that there was another subprocess, actor or other situation that was not taken into account yet. However, as researchers/developers/designers we do recommend to keep creating these overviews, as it makes it much easier for your stakeholders to point out whether you have still misconceptions or gaps

in your knowledge.

The complexity of the system also leads us to the last reason why it was initially hard to structure our interview guide: the process was not only complex but could work **differently in practice** between different countries and analyses. That is why, inspired by the sequence diagrams [54] used in software engineering, we decided to first let interviewees create a diagram for one of the recent analyses they had performed, before we would dive deeper into the different elements and decisions within the process. To save time, we created a basic diagram describing the analysis as far as our knowledge went, which interviewees could subsequently adapt to fit with their experiences. See Figure A.1 in Appendix A for an example. In Appendix A one can find the interview guide as well.

5.1.2 Finding values to consider

Given our aim to develop an HCAI-aligned system, and that we found in our literature study in Section 2.2 that values had to be related to each other and prioritized, we also wanted to get more information on which values had the highest priority for the IPC in decision-making. Our first identified values came from our discussions with people from the IPC, who could tell us a lot about general requirements for the IPC and the humanitarian sector, such as accountability, economy of resources, and safety.

What we found out during these conversations was that it was easier to *identify* important norms and values than to know how to *weigh* them against each other. The only more practical ground rule that we found during our exploration, was from a conversation we had with two senior food security analysts. When faced with the question whether it would be preferable to have false negatives or false positives in the monitoring tool, they advised to err on the side of false positives, as a no-regrets policy..

This last example also illustrates that asking for such values at a very high level did not, in most cases, help to determine a more grounded and practical policy for design decisions, as most people can only really begin to think about them in relation to concrete cases. This led us to the decision not to ask the facilitators in the interviews what the most important values were for them, as we had originally planned, and instead to use their stories and insights as input to derive a general set of design requirements for our tool. These design requirements can be found in the section 6.3.

5.1.3 The interview guide

The final interview guide in Appendix A was structured around three objectives: Establishing the interviewee's experience, obtaining information on assumption building and the analysis process, and discovering current practices and ideas about a possible monitoring tool. As we had limited time with each participant, the interview questions were selected rigorously to focus on only the questions that were absolutely necessary. The first part was therefore quite to the point. We only asked the participants about their current function and organisation, and how many IPC analyses they had been part of and in which capacities.

The second part, where we delved deeper into the consensus process, was considerably longer. Here, it was our goal to understand how the assumption-building process for IPC analyses and projections is structured and how assumptions for current analyses and projections are developed throughout this process. We addressed this by investigating how key decisions were made and important factors were determined during the analysis. Specifically, we wanted to know about each of those factors: what do those factors look like, on which (sub)national level are they determined, and who determines them at what time and based on what information and reasoning? These questions were especially important for us to know in light of any possible monitoring solution: we need to know who decides what about the assumptions to know what future users and stakeholders to take into account, we need to know the shape and data used for the factors contributing to the assumptions to know how we might formalize and monitor them, and we need to know when and how everything happens so we can design a tool/process that is aligned to the current practices of the IPC. As different analyses might differ in setup and details, we asked the participants to relate these questions to the

last business-as-usual analysis they had been a part of and checked then afterwards which parts were general practice or specific to that analysis.

For the last part, we inquired about current monitoring practices. We planned to ask if and how they normally monitored the validity of their projections, how and who decided whether an analysis update was needed, and what information they used to support that decision. However, while we did ask those questions, we realized during the interviews that it was also very insightful to discuss with them the possibilities of a monitoring system. All the interviewees had many years of expertise and were very good at pointing out what parts of a monitoring system would or wouldn't work and where the challenges and issues would be in developing such a system. Naturally, this became a standard extra question in the interviews.

5.2 Procedures

We originally planned to interview 7 IPC facilitators. These people were recommended to us by our IPC contact person due to their extensive experience at the IPC and sharp analytical skills. However, as most IPC facilitators have very full schedules due to their profession, we only managed to interview 3 facilitators. Nonetheless, this number proved to be enough already to obtain an extensive description of the IPC analysis process, given the facilitators' many years of experience. Each interview session lasted 1-1.5 hours and took place online.

To ensure the participants' anonymity and shield them from potential professional harm, we synthesized each of their inputs into a separate general overview of how the IPC analysis cycle works and what the challenges are for monitoring. They then reviewed this overview to ensure we understood them correctly and that no identifiable information was contained within the document. To present a full picture of the IPC analysis cycle, we subsequently combined these syntheses into one shared overview, which is presented in the next section. When necessary, the individual overviews are available upon request.

5.3 Results: The IPC analysis cycle in practice

In this chapter, we will give an overview of the IPC analysis cycle based on our interviews in relation to our case study: How can we improve IPC analysis monitoring? The description below will, therefore, not repeat the IPC manual but rather give qualitative insight into how it is executed in practice. We will sketch the background for the challenges we have uncovered surrounding the monitoring of IPC analyses and show the opportunities for interventions that could be involved.

We will walk through the important stages in the IPC analysis cycle in the following sections to answer the following questions: **How, when** and by **whom** is/are:

- An IPC analysis initiated?
- The analysis parameters of the IPC analysis determined?
- The contributing factors, key drivers, projection assumptions and risk factors determined?
- The IPC analysis monitored?
- Decided to do an IPC analysis update?

After we explain who the important actors are in this process, we will continue to move through each stage in the analysis cycle as illustrated in Figure 5.1

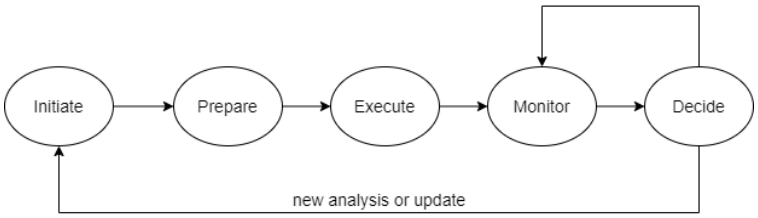


Figure 5.1: The different stages in the IPC analysis cycle

5.3.1 The actors

The Technical Working Group (TWG)

The TWG is a country-level group responsible for initiating, planning, and coordinating the analysis for its country and ensuring that the financial requirements are met. During the monitoring period, they also meet regularly and finally decide if an analysis update needs to be done. This group is composed of relevant national stakeholders and usually includes representatives of the government, United Nations agencies and NGOs. As the core purpose of this group is to create political and organisational feasibility, it is not a guarantee that each of its members has a high expertise in food insecurity.

The Regional Coordinator

The regional coordinator provides quality control and support for the IPC for a larger region in the world, such as Central Africa. They also monitor food insecurity in different countries and can poll the TWG on whether to conduct an analysis update. The regional coordinator has extensive expertise on food insecurity and is usually accredited as a level 3 IPC facilitator. Level 3 IPC facilitators are allowed to "act as IPC global experts who can lead IPC training and analysis facilitation across countries and regions." [16]

The Facilitators

Facilitators lead the analysis process. They are responsible for quality control and ensuring that the classifications concur with the evidence for each respective region. They also guide the analysts in reaching a technical consensus. They have strong expertise in food insecurity and are usually accredited as level 2 or level 3 facilitators. Level 2 IPC facilitators are allowed to act "as resource persons to support IPC Level 1 Trainings and Country Analysis Workshops, by facilitating small-group works." [16]

The Core Team

The core team is an unofficial group within the IPC analysis that is operationally responsible for the primary data collection and data analysis. They are the main leads in the analysis process. This group consists of a combination of:

- Members of the TWG
- Support staff: people who are responsible for the logistical arrangements.
- Facilitators. These can be facilitators from the country itself but also be cross-country learning experts.

Given its variety of members, this team hosts a mix of expertise in relation to food insecurity.

The analysts

The analysts are experts within the country responsible for doing the actual unit analyses. This group is a mix of people with expertise on different food insecurity topics, such as livelihoods, nutrition, food prices, conflict, health, gender, and more. They have to have had at least level 1 training. However, this training is not always refreshed for everyone, which means that the level of food insecurity expertise can vary within the group. Typically, there are around 50 to 60 analysts participating in an IPC analysis.

5.3.2 Initiating the IPC analysis

It is the responsibility of the TWG to initiate IPC analyses. Normally, these analyses are performed once or twice per year in each country based on that country's seasonal calendar.

Frequency:

The reason for the frequency of the analysis is the outcome of two counteracting factors:

- The first factor is that people try to extend the periods between analyses as long as possible, as performing an analysis is both costly in terms of time and money.
- The second factor is that there is a limit to this extension: IPC analyses and the data used for those are maximally valid for a year, so after this, there needs to be an IPC analysis performed in either case to maintain a valid IPC classification.

Timing:

The reason for basing the timing on the seasonal calendar is that it is best to do this during the post-harvest season. At this moment, it is possible to assess how the season's harvest has performed and how well this harvest can support the people in the months to come.
The actual analysis dates, however, can vary from year to year because they are also impacted by other factors such as:

- **Data availability:** an essential source for the IPC analysis is the primary survey. If this is delayed for some reason, the analysis will also be delayed.
- **Large shocks:** When there are major developments at play that influence food insecurity, this might also cause the analysis to be pushed forward or backwards. For example, the analysis might take place earlier than usual if circumstances are getting so dire that there needs to be a new analysis performed to account for the new situation. It might also be pushed back when there are developments happening with a delayed impact that decision-makers want to take into account as well. When, for example, developments are taking place in other countries, that might start to affect them as well.

5.3.3 Preparing the IPC analysis

There are a few important activities that the core group needs to undertake before the analysis to set the scope of the analysis. These are:

- Determining the areas of analysis
- Determining the validity periods
- Gathering evidence

Determining areas of analysis

The IPC always gives out separate classifications for different areas in a country. They call these different areas "units" or "areas of analysis". It can differ per country and per analysis how the country is divided into these areas of analysis and which are used.

- **Who** determines this: the core team
- **When** does this happen: around three months in advance
- **How** are they determined:
  The core team sits together to determine which areas will be considered for the analysis. Usually, these areas are fixed and will be repeated from one analysis to another. Sometimes, however, there are not enough resources to analyse all the areas in a country. Then, the core team picks a subset of areas that are more likely to have serious food insecurity issues, given what the contributing factors point towards. The number of areas of analysis can vary between countries, ranging between 10 to 300 areas.

Determining the validity periods

Validity periods (VPs) are periods within which a certain IPC classification or projected classification will be assumed to hold. Each validity period corresponds to a current or projection analysis and stretches out

to the next projection period or until the end of the validity of the whole analysis (e.g. a year after the analysis was performed). Within validity periods, it is assumed that the food security conditions will remain largely the same. The number of validity periods within a year can range between one and three periods, depending on the number of projection analyses that have been performed.

- **Who** determines the VPs: usually the core TWG with the facilitators
- **How** are they determined: VPs are usually determined based on three factors:
  - **Time and resources:**
    As explained before, people usually try to extend the time between analyses as long as possible. This means that the validity periods often sum up to a year. To account for the food insecurity fluctuations within this year, the IPC often recommends that countries do two projections in a year. Embedded in this recommendation, however, is the expectation that people do a projection update anyway between the first and second projection to provide more up-to-date information for decision-making. It will be, namely, always hard to perfectly predict what will happen in 7 to 9 months in the future. However, sometimes an analysis may be so extensive that it simply leaves no time for the analysts to do projections. Then, there will only be the current analysis, whose validity period will cover the whole year.
  - **Seasonality:**
    In the case where there are projection analyses, their periods are usually timed based on seasonality. Most of the countries where the IPC works have namely an important seasonal component to explain food insecurity: In the countryside, the household economy is based on agriculture and in the cities, the poorer people see seasonality in the food prices. Therefore, the VP of the current analysis usually covers the post-harvest period, the VP of the first projection covers the lean season, and another projection covers, if needed, something in between.
  - **Decision-makers' needs:**
    In some countries, there is a strong request by their governments to have the results at a certain moment, or a decision-making process like an HNO (Humanitarian Needs Overview) or HRP (Humanitarian Response Plan) may require to have the results at a certain date. In these cases, the validity periods may be changed or tweaked to fit the decision-makers' needs.

Gathering evidence

- **Who** determines what information is needed: the core team
- **Who** is responsible for uploading the evidence into the repository: facilitators and helpers

Information needs

In order to perform the analysis, evidence needs to be gathered on each aspect of the IPC analytical framework. This means that data needs to be gathered covering all the main aspects, e.g.

- **Outcome data** (See Section 4.1.3)
- **Contributing factors:** (See Section 4.1.2)
  - Hazards and shocks: what was their severity and magnitude? How did they impact households' food insecurity?
  - Vulnerability: what are their sources of food and income? What are their usual coping strategies?
  - Government policies and issues
- **Food Insecurity dimensions:** (See Section 4.1.1)
  - Food availability: Understanding what is going on at the markets,
  - Food accessibility: Knowing whether people able to purchase or borrow food
  - Food utilization: Checking what foods people prefer and how their access to water is

Information sources:

There are different data sources that can be used as evidence in the analysis.

The most important source for the analysis is the **seasonal survey.** This survey is performed on a representative household sample in each of the areas to get an insight into what the current level of food insecurity is for different households and what the driving factors are for their food insecurity. What questions are put in this survey, is guided by the analytical framework. It is the role of the core team to refine the questions from the analytical framework, see which ones are applicable and see which ones they may or may not need to delve into. So important data that is gathered in the survey are things such as:

- Outcome indicators on consumption, livelihood change and nutrition
- Information on vulnerability
- Information on FI dimensions

Apart from the survey, there are also other sources of information that are often used to inform the analysis. These can come from partners such as FEWSNET, the World Bank or WFP, or from the local governments or ministries. Examples of such data sources are:

- Food prices
- Data on crop production
- Seasonal calendar of the country
- Mortality rates
- Climatological data

Preparing the evidence repository
When the surveys are performed, and the data is gathered, the facilitators, with a team of analysts, analyse all the information sources and upload them to the platform so there is information on each area that the analysts can use for their analysis.

5.3.4 Executing the analysis

The setting

The actual analysis is performed in a workshop lasting one or two weeks. In these two weeks, analysts analyse the food insecurity situation in each area of analysis. To this end, they are divided into subgroups of 2-4 people, each subgroup being responsible for a number of areas. Each facilitator is assigned a number of subgroups. These subgroups can, for example, be assigned by province or other larger subnational region.

During the analysis, sometimes plenary sessions will be held where important information is shared or analysis outcomes are discussed. As IPC analyses typically happen in one large room where all the analysis teams are sitting at tables in their own groups, transitioning from the subgroups to these plenary sessions is quite easy: someone can just walk to the microphone and grab everyone's attention.

In general, the main steps in the analysis are:

1. Informing the analysts
2. Doing the current analysis:
   (a) Determining contributing factors
   (b) Determining phase classification
   (c) Determining key drivers
3. Doing the projection analysis:
   (a) Discussing national projection assumptions
   (b) Doing unit-level projection assumptions
   (c) Determining risk factors to monitor

Some of these steps might happen concurrently or interchangeably. See Figure 5.2 for an overview of the timing.

Informing the analysts

Typically, the analysis will start with a plenary session where someone will explain to the analysts the setup of the workshop and the evidence that is available, e.g. what evidence is available, how this evidence is organized and perhaps advice on how to approach the evidence, so analysts are on the same page.

Then, somewhere in the beginning or in the middle, different agencies can also come to give special sessions on different topics. When these sessions happen exactly depends on who's available at what point. Example sessions could be:

- WFP giving a presentation on expected or ongoing humanitarian food security assistance.
- FEWSNET giving a presentation on markets and market prices in different areas, as well as food price projections.
- Conflict experts giving a presentation on conflict and what the expected evolution of the situation is.

Doing the current analysis

Contributing factors and phase classification

The starting point for the analysis is identifying the contributing factors. Analysts usually start by reviewing all the evidence they have for contributing factors and analysing it one by one. The main source of evidence they have for this is the seasonal survey.

Example questions they might investigate:

- What are the livelihoods in this area? How are people making their ends meet?
- How did the harvest go, and how is their livestock doing?
- What are the different shocks that are affecting this area? How severe have they been?
- What are the prices in the markets?
- What have been the kind of impacts, and how do all these translate into your food availability, your access to food, your food utilization, etc.?

In doing this investigation, they identify contributing factors based on the evidence on each aspect of the analytical framework. After the contributing factors have been identified, the analysts look at the outcomes, e.g. livelihood change, food consumption, nutrition and mortality. Converging on all this evidence, they decide on the phase classification for the area and fill in the population table.

Key drivers (KD)

After the contributing factors have been identified and the area is classified into one of the five phases, the analysts' next step is to determine the key drivers. Key drivers are a subset of the contributing factors: those that have the biggest impact on the food insecurity of that area.

Example reasons for picking certain drivers:

- "In this situation, we see that the food prices are very high, so that should be a KD"
- "The production is very low compared to the last five years, so that should be a KD"

On which regional level are they determined:
Key drivers are always decided by analysts for a specific area. So in principle, it's possible to have different key drivers for every single analysis area. In practice, however, this never happens. So, typically, there are certain key drivers that tend to feature in at least most of the analysis areas. Quite often, these key drivers have a lot to do with shocks and with aspects of food access.

Doing the projection analysis

National projection assumptions

Before the projection analysis is done on a unit level, there is normally a plenary session to discuss projection assumptions on a national level. These are projection assumptions that will apply to most of the areas, or at least a large extent.

What does this session look like?
Usually, the core team prepares an initial draft with projection assumptions. Those are then projected on the screen and discussed in plenary with all the analysts. However, other setups are possible as well. In some countries, there is, for example, a lot of FEWSNET expertise. In these cases, they often prepare a presentation. In some other countries, there is no preparation at all, just a plenary conversation. Here, they ask the analysts in the room if someone can name an assumption that should be considered, and based on the ideas that people throw out in the room, they develop the projection assumptions together with everyone. But as this can take a very long time, usually at least some form of presentation or list is prepared.

Unit-level projection assumptions

After the plenary discussion, the analysts tweak the national assumptions to fit them into the context of their specific areas of analysis. This is necessary as there might be circumstances at play that will be unique for their specific area and therefore cause different food security dynamics. For example, when a large extent of the livelihoods in an area are dependent on mining, these livelihoods might be affected by very different factors than typical agricultural livelihoods. Hence, the analysts will add a specific set of projection assumptions on mining developments for that area.

There are normally two ways in which analysts tweak national projection assumptions:

- They might completely remove or add certain projection assumptions.
- Or for more basic assumptions such as climate related assumptions, analysts may change what impact that specific assumption might have on their areas. For example: much rain might mean different things for different areas: for one area it might mean good harvests, for another area it might mean flooding.

About projection assumptions

Structure:
Each projection assumption is normally about one factor that influences the food insecurity. One can therefore view each projection assumption as consisting of two components:

- The first component: the assumption about how this factor will develop. For example, the rainfall will continue to be lower than normal
- The second component: the assumption about how this development will impact the food insecurity (dimensions). For example, the continued low rainfall will cause lower food production and limit food access.

Sources for assumptions:
Key drivers, assumptions and risk factors are generally linked. As key drivers are the main causes for food insecurity in an area, it is logical to investigate how they will behave in your projected scenario. Hence, each key driver should get a corresponding projection assumption. However, there are other types of assumptions next to key driver-informed assumptions that should be developed, that might not be key drivers for the current situation but that could have an influence on the food security situation in the future. So often, there might be between 3 to 5 key drivers in an analysis, but around 7 or 8 projection assumptions.

How they are developed:
They are normally developed by applying critical thinking and inference on the data of the current situation, using contextual knowledge and historical evidence. So projection assumptions are normally based on trends in the past, or on forecasts of agencies that have expertise on those. They can give for example:

- Rainfall forecasts
- Price projections
- Conflict projections

Types of assumptions:
Most of the time, projection assumption are about:

- Climate related things
- Economic related things:
  - Price forecasts
  - Market prices
  - Inflation
- Movement (displacement, migrations)
- Trade, petty trade, cross-border trade

Sometimes they are about more rare but high-impact events, such as a volcanic eruption or a completely new phenomenon. In the last case, analysts then also have to figure out how that will affect food insecurity. For example, during COVID19, they needed to find a way to map that to food insecurity. So they linked it, for example, to the closure of businesses: As people get their incomes from those businesses, when they close, this will affect their household incomes. In this way, they could then model how that will impact their ability to obtain food.

Problems with projection assumptions:
Sometimes, projection assumptions are formulated in a way that is very general, not precise, and sometimes even unhelpful for the analysts. For example: "The trend of the quantity of rainfall will be lower than normal for (....), which will cause (....)". The problem with such a projection assumption is that it is hard to know how to interpret it: lower than normal can be very, very low or just a bit. The same goes, for example, with the statement: "The general inflation will remain high and will negatively impact the purchasing power of the households.". This statement begs again the question: To what extent will the inflation rise, and to what degree will this impact be?

Risk factors

As one of the last steps in the unit analysis, risk factors are defined by the analysts. Risk factors are factors that need to be monitored because they may raise the need for an analysis update: if you are basing your prediction on assumptions, then you will want to know whether these assumptions are still correct, or else your prediction might be incorrect as well. Other risk factors that need to be monitored are events that are unlikely to happen (and therefore not assumed on), but that do have a high impact when they happen and almost certainly lead to an analysis update. An example of this is earthquakes.

About setting thresholds on risk factors:
Normally, there are not any thresholds determined for the risk factors because the projection assumptions are not quantified. As most assumptions are so general and so loosely defined, there is no way one can establish a threshold or make a rule for when the update of the prediction needs to be revised. For the very clear cases, such when it was assumed that rainfall would be below normal and the rainfall turned out to be above normal, then it is clear that there is a need to update. But for example, with food prices, it is never assumed to what extent the price would be above or below the average. So it's difficult to set up a threshold for that.

5.3.5 Monitoring the IPC analysis

When the analysis is published, it will be monitored throughout the year by different actors, which can differ per country. For the IPC, it is most important to know when their analysis needs to be updated, as the analysis informs numerous other decision-makers.

Monitoring practices

There is always some kind of food security monitoring happening in every country. Normally, this is done by the country, the data analysis team or the TWG. However, this monitoring varies in implementation and in who is monitoring what.

Generally, the factors that are normally monitored are:

- Food prices: the World Food Program has a weekly monitoring of the market prices
- Indicators on water and health:
  - The Ministry of Health and the government typically have sentinel sites and surveillance for the number of children that are being admitted on a weekly basis from facilities and the access to water
  - Disease epidemics
- Climatological factors:
  - Basic monitoring: The government and the Ministry of Agriculture collect usually rainfall data on a weekly basis to show how it is progressing, especially during the rainy season.
  - Advanced monitoring: There are more scientific forecasts done by CHIRPS, which are not very regular, as they depend on when people have had a meeting about them.

Deciding on projection updates

Who decides:

- The TWG are typically the ones to decide to do a projection update. They meet periodically to discuss the country's food security changes, among other things.
- The regional coordinator also sometimes polls the key people in a country on whether there should be a projection update happening. This could also be a starting point for a conversation on this.

What information is used to determine the need for an analysis update?
Often, there is no access to new outcome indicators such as the FCS, so normally, updates on contributing factors are used to determine the need for an analysis update. So this will be data on:

- different kinds of shocks
- humanitarian food security assistance
- rainfall
- markets. production and market prices.

Rules for a projection update:
There are no hard and fast rules for when to do projection updates. Generally, people decide to do a projection update when they see that the situation is going in a very different direction from what they had projected. So, an example might be: if your rainfall deviates considerably from what was expected, then that's when you might want to do a projection update.

The general threshold for deciding on an analysis update is: if you think that your phase classification has changed by at least one phase difference, then you should do one. If projection assumptions have changed but won't lead to a change in phase classification, then it might not be necessary to do an update, as the message to decision-makers will stay the same.

When do projection updates normally happen?

1. If the decision makers call for it:
   They have their own sources of information that might trigger them. The households might, for example, complain that oil prices and commodities have gone up. Another case is when there is a new government that subsequently wants to have new information. In this case they can request an analysis update, which is almost always obliged due to their authority.

2. When something very clear and big is happening.
   For example, in a certain country, they started to have some signals that something was going on in the field, like people migrating and reports of clear cases of starvation. This led them to do an analysis update. So, in this case, they weren't monitoring any risk factors but just saw that their analysis was no longer reflecting what was happening on the ground, so they did an analysis update. Other examples of such analysis updates happen when there are hazards, high displacement, high inflation, disrupted supply chains, or unexpected price spikes for staple foods.

This last example also illustrates the current problem with projection updates: because there are no hard rules for when an analysis update has to be done, these analysis updates often happen too late. This shows, therefore, also a clear area for improving assumption monitoring because currently, updates only happen when something very obvious is telling us that the analysis is not holding anymore. Instead, an analysis update should also be done when a combination of data sources indicates the need for an update, such as data on prices, rainfall, currency or availability of commodities on the market.

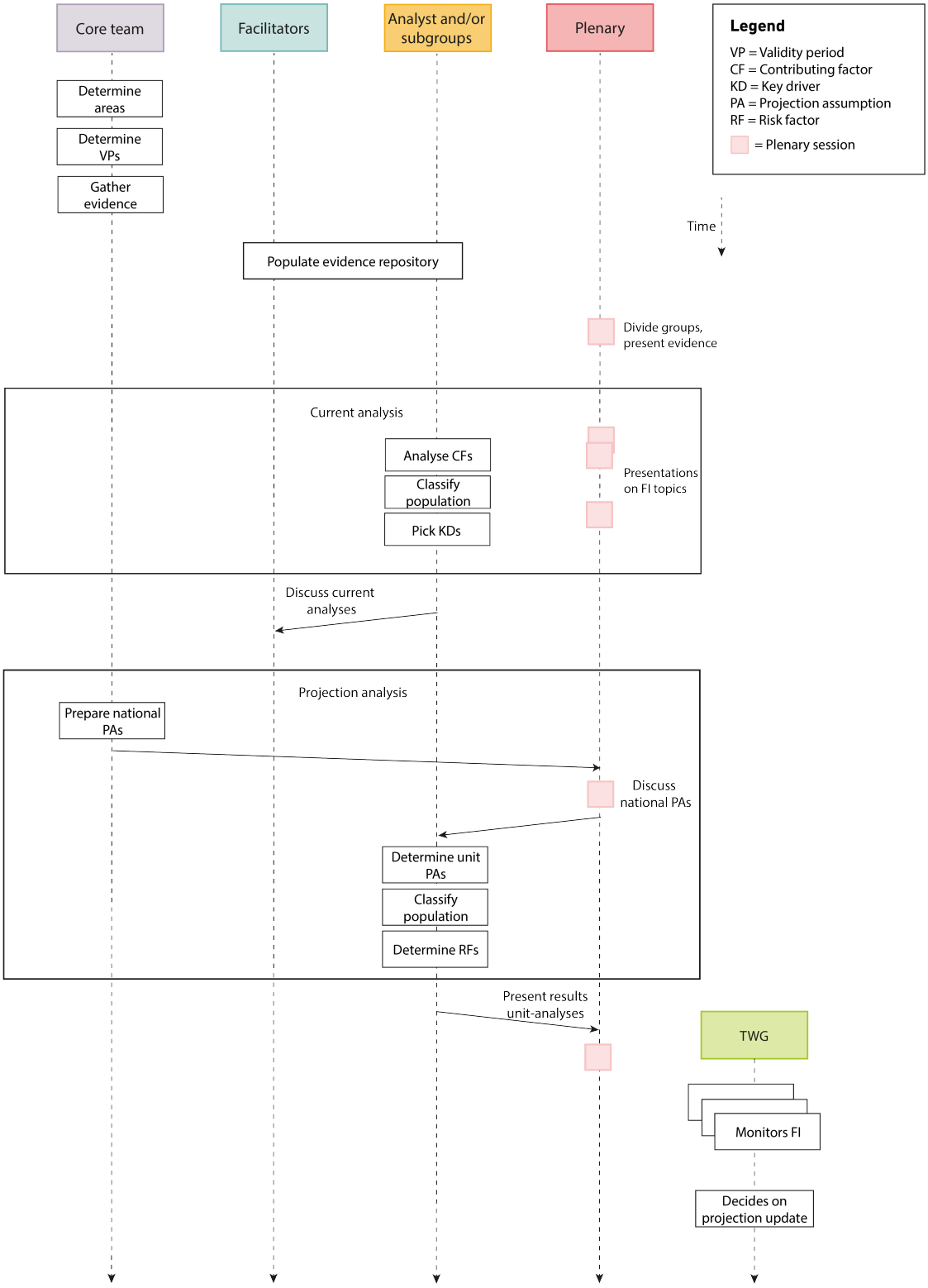


Figure 5.2: Schematic overview of the different steps taken by actors during the preparation, execution and monitoring stages of the IPC analysis cycle

Chapter 6

Analysis: The IPC monitoring problem and how to fix it

Given the understanding we had gained about the IPC analysis cycle in Section 5.3, we have synthesized this below into a succinct explanation of what we see as the current challenge with IPC monitoring and decision-making surrounding analysis updates.

6.1 The problem

IPC analysis updates happen too late

In many countries, the IPC has a critical role in providing decision-makers with actionable information on when, where and to what extent people are suffering from food insecurity. As such, they play an important role as an early warning system, as their yearly analyses and projections can activate many people to start taking preventative measures. However, each year after the analysis of a country has been performed, the IPC's function as an early warning system diminishes when they react too late to developing circumstances and only update their analysis when it is already obvious to all involved that the situation is deteriorating, as we found in part 5.3.5 of Section 5.3.

Important note: While we will go deeper into the role of the IPC in this problem in the following sections, it is essential to understand that there is a limit of control they have over this situation, as it depends on donors and other funders whether there is actually money made available to do an analysis update. So the IPC's ability to react will always be conditional on whether other stakeholders are willing to pay.

The cause: uncertainty about validity

The cause for this late reaction is that, at certain moments, it is not clear whether an IPC analysis is still valid and, therefore, it is not clear whether an analysis needs to be updated. This is due to the fact that the IPC does not always have a good overview on whether the assumptions on which the analysis is based are still correct, either because those assumptions:

- were formulated in vague statements, which makes it hard to determine whether they were still correct.
- were not perceived as assumptions in the first place and, therefore, not monitored for their validity.

To be clear, we are talking here about a broader range of assumptions than just the projection assumptions, as we will see when we start analyzing the current structure of IPC analyses.

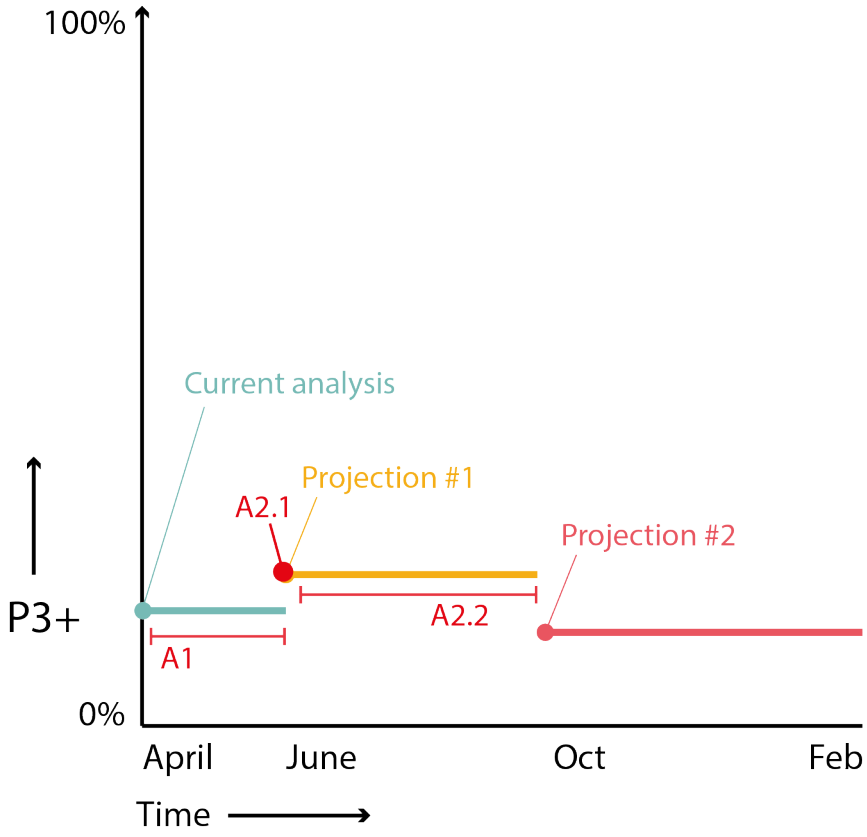


Figure 6.1: Assumptions inherent to IPC AFI analyses

So what assumptions are we talking about?

To understand this, we need to take a look at Figure 6.1. This depicts a typical IPC food insecurity prediction for a given year and country. The y-axis denotes the percentage of people within an area who are in phase 3 or higher (P3+). The starting point for this prediction is the current analysis. The food insecurity classification for this moment in time is based on primary data on FI outcome indicators and contributing factors and can be seen as a ground truth. This classification is then expected to hold for the duration of the first validity period, which lasts up to the next projection.

Then, the next checkpoint is the first projection classification: Projection #1. As there is no outcome data for this moment in time, this projection is determined by forecasting how the contributing factors will evolve, and then predicting how these developments will affect the food insecurity situation in each area. These predictions are captured in projection assumptions. A good projection assumption normally consists of two parts: a statement on how a specific contributing factor will evolve and a statement on how this will impact food insecurity for a specific area. Again, this projected classification is expected to hold for the duration of its validity period, which either lasts up to the next projection or up to the expiration date of the primary data, which lasts up to a year after its collection. The same rules apply for each subsequent projection made in that year.

When analyzing these classifications, one can identify a few different types of assumptions that are being

made:

- The first assumption (A1) is about the validity period of the current analysis. While the current analysis itself can be seen as a snapshot of the ground truth, the expectation that the food insecurity and contributing factors will remain the same for the following X months is an assumption that should be monitored for its validity.
- The next ones are about the projections:
    1. The first ones (A2.1) are the assumptions that contributing factors will have developed to certain values at the time of each projection and that those will result in the corresponding food insecurity classifications. These are the projection assumptions.
    2. The second ones (A2.2) are the assumptions that those contributing factors and FI classifications will remain the same for each projection period up to the next checkpoint.

Even though we often don't have access again for the rest of the year to food insecurity indicators, we do have a lot of data throughout the year on different contributing factors, so it is very possible to use those to at least monitor whether these assumptions turn out to be correct. However, at the moment, this data on contributing factors is not optimally being used for deciding analysis updates due to two reasons:

- First of all, assumption A1 is currently not even identified in IPC protocols as a factor to monitor, so contributing factors for the current analysis are not monitored at all.
- Secondly, monitoring the validity of assumptions A2.1 and A2.2 is currently quite hard, as the projection assumptions for the projection periods are formulated so vaguely that it is hard to determine in many cases whether they are still valid.

6.2 The goal

These challenges have led to my renewed and better-specified design mission:

Improving the monitoring process for IPC analyses, so decisions on analysis (updates) can be made:

1. based on systematic data-driven argumentations.
2. in time.

To do this, we propose to develop two systems:

1. A system that helps to define clearer risk factors for contributing factors as well as projection assumptions. This means:
    (a) Making contributing factors **measurable** by developing mathematical representations for them.
    (b) Making it clear **which values** those contributing factors are expected to take on, both for the current situation and the projection scenarios.
    (c) Setting **thresholds** on when the contributing factors do not adhere to these predictions anymore.
2. A system for monitoring these risk factors that can:
    (a) indicate when they are exceeded,
    (b) and help users to make an estimation (possibly by means of HCAI) on whether the classification for that period still holds so they can decide whether or not they need to do an analysis update when contributing factors have led to uncertainty about the validity of projection assumptions.

6.3 Design requirements

As we wanted to develop our solution in a human-centered way, an important question for us was what requirements our tool should adhere to, both in terms of values/norms and in terms of identified user needs. Based on our interviews with the IPC facilitators and our contact person, as well as on the literature on design requirements for information systems in humanitarian decision-making [48], we developed a set of design requirements in the form of a set of ground rules to adhere to. See Section 6.3.1.
While in Chapter 2, we looked at norms and values from a quite abstract and high-up point of view, we discovered in this phase that it was much easier to arrive at our design requirements from the bottom up. It was namely easier to derive from more concrete examples and rules, some larger overarching norms, than to work from abstract norms to concrete guidance. We did map our found design requirements afterwards to broader norms and values by clustering them on their themes and purposes.

This allowed us as well to compare our bottom-up design requirements with the value hierarchy that we developed in Chapter 2 and see where there was overlap in our norms and where connections were missing. This comparison can be found Section 6.3.2 below.

6.3.1 The requirements

Below, one can find the list of the design requirements and their overarching themes. It is important to note that these requirements do not all have hard boundaries and might sometimes oppose each other. Many design decisions will, therefore, be about striking a balance between those different requirements. Good examples of requirements that require such weighing are Efficiency and No regrets: how much do we want to err on the side of caution before it infringes too much on the principles of economy of resources? Furthermore, this list of requirements is also incomplete and might even be inconsistent. They are derived from literature and discussions with the IPC by picking and adapting requirements when we thought that they would give practical and relevant guidance. As such, these requirements below are rather a snapshot in time than something definitive. However, we can deal with these imperfections as the requirements will rather function as a hatstand to make important considerations behind future designs explicit instead of some absolute decree.

However, one strong quality requirement for us, when it came to these design requirements, was that the IPC agreed with what was in our (temporarily) final list. We have therefore presented this list, as well as the rest of our problem analysis and proposed solution, to three IPC experts, one of whom was fully involved in the project and the other two were not. As such, they were able to agree with the requirements we have put here below:

Usefulness for organization

- **Consistency in approach:** The solution should be aligned with current IPC protocols, procedures and tooling and aim to avoid any unnecessary changes.
- **Relevance:** Data that is shown or gathered should aid in the decision-making process for analysis updates.
- **Usability:** The tool should be adapted to the technical knowledge of its users.

Accountability and soundness

- **Traceability/verifiability:** Users must be able to evaluate the reliability and credibility of data and information provided by the system and should be able to trace its reasoning and origins.
- **Meaningful human control:**
  - Users should set the boundaries within which AI is allowed to function

  - Users should not be asked to do tasks that they might not be capable of. In those cases, they should either get proper training beforehand or be supported during the task by the tool and/or other people helping them.

- **Agency:** The tool should enable users to get a good insight on the (likely) food insecurity situation so that they are able themselves to make well-informed decisions

- **Timeliness of information:** The tool should have up-to-date information where possible, and it should indicate clearly when information is old or might be out of date.

Economy of resources

- **Efficiency:** The tool should minimize unnecessary use of resources in terms of time and money.

- **Building on expertise/No reinventing the wheel:** Where possible and relevant, the tool should build on the expert knowledge already present in the IPC.

Preserving lives

- **No regrets:** When necessary, the tool should err on the side of caution: rather overestimate than underestimate the food insecurity situation. Unnecessary action is always better than inaction in times of need.

- **Timeliness of decision-making:** The tool should enable decision-makers to decide on projection updates on time when this might be reasonably inferred from the current data on contributing factors.

6.3.2 Requirements compared to value hierarchy

When we compared our design requirements with the norms and values identified in Chapter 2, we made two observations. The first was that not all of the norms from the hierarchy could be covered by our requirements. This is partly due to the fact that the design requirements we derived in the current chapter are about what requirements *the system* should adhere to, rather than the *design process*. This means that norms such as inclusion and preventing biases are not covered in our requirements, since inclusion, for example, is a requirement of our human-centered design process rather than a system requirement. However, as this is only part of the reason, it also highlights a gap in our design requirements. For example, we have not done an active inventory of the possible biases we need to watch out for in development, nor have we made any requirements about how we should mitigate those biases. So this should be an important addition to our requirements in the future.

The second observation was that not all of our design requirements could be mapped to our current norms for HCAI. We found that, while our design requirements grouped under "Preserving lives" could be mapped to the value "Safety", their active purpose to help other people could not be mapped to a specific HCAI norm, as it rather stems from the core principles of the IPC as a humanitarian organization. This shows that not all HCAI norms can be predetermined, as each context in which HCAI is developed may have its own specific norms in addition to the more general HCAI norms. As such an important takeaway for HCAI is: **If you want to develop responsible HCAI, you need to tailor it to the specific values of its institutional context.**

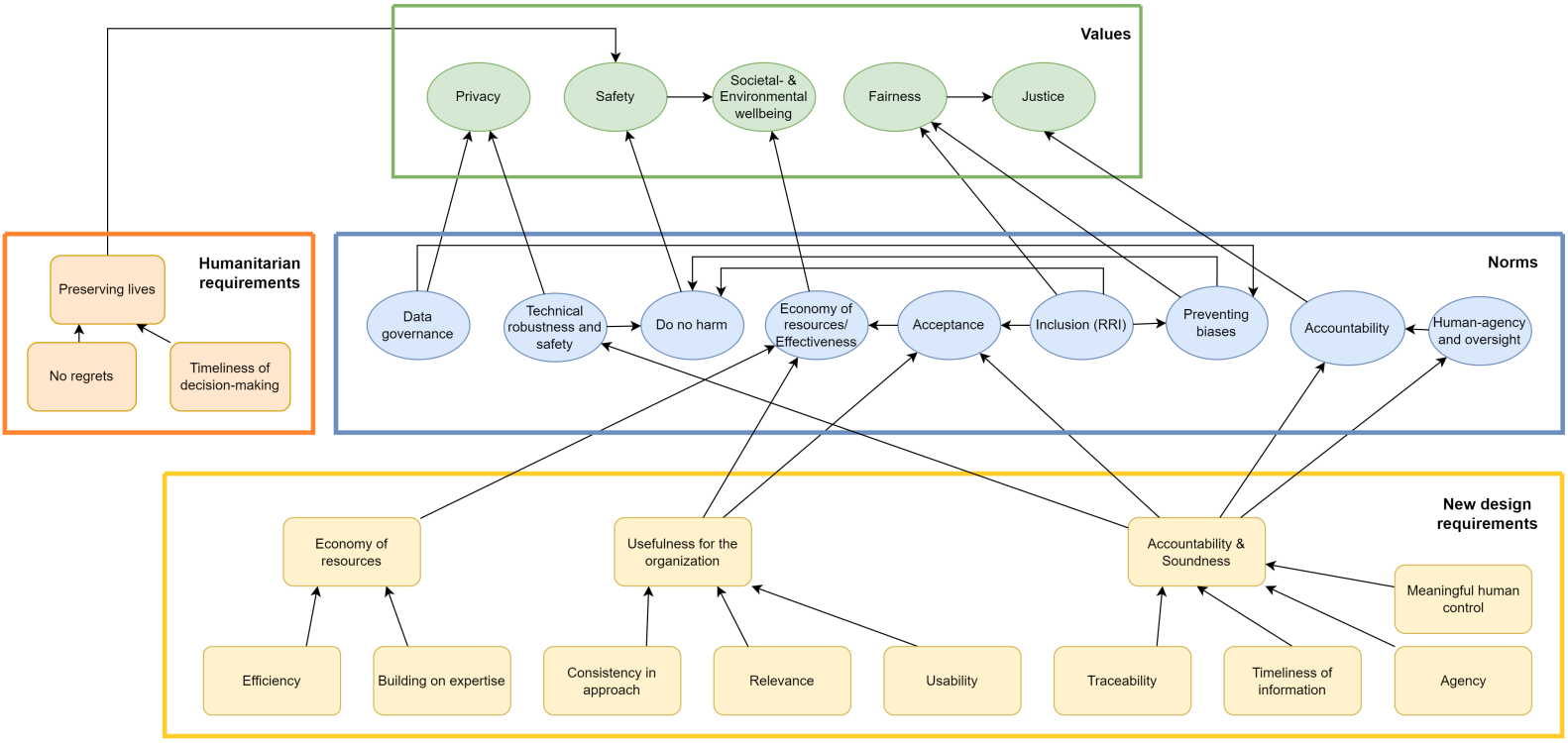


Figure 6.2: Design requirements compared to value hierarchy

6.4 The next steps

While we tried to make our proposal of Section 6.2 more detailed, we realized during our attempts that it depended on so many unknown factors that such a waterfall approach [56] would not be of much help. Each separate step in our proposal begs the question: who will do this and how, and there is not a definitive answer to be found beforehand. Many elements of this plan need to be step by step developed and tested before we can find an approach that we know will work. That is why, in moving forward, we decided to get more concrete and zoom in on one of the components of our monitoring process design: quantifying projection assumptions.

At this point, one might ask, "Why not start first with implementing AI instead of focusing on projection assumptions?" The reason we did not start with AI was that any future AI needs to be able to take these formalized projection assumptions as input to its model. Therefore, it is necessary to (a) figure out how this input should look and (b) test whether IPC analysts have the technical capacity to make such formalized assumptions.

Furthermore, there are several reasons why it is necessary that an AI model takes these assumptions as form of expert input into account:

1. **Enhancement, not replacement:** The first, most important reason why we need these assumptions, is that we want our FI model to say something about the validity of the analysis, not to have a second opinion. If you try to predict FI without taking any input from the IPC analysis, you are rather imitating or replacing the IPC process instead of enhancing it. If a model takes the initial analysis outcomes as leading, it can use the expert input on how certain assumptions will impact Food Insecurity and make estimations about the validity of the analysis that respects the original reasoning behind it.

2. **Unprecedented events:** We also need to use expert input to enable the FI model to work with a

full picture of the FI situation. As also indicated in [50], there are many different unprecedented/rare events that can happen that an expert can see happen and can take into account, that a purely data-driven model cannot, such as a looming El Niño, or a pandemic such as COVID19. In order to get model-estimations that are aligned with the real-world situation, it is important that such a model can be steered by the experts.

3. **Data availability:** Given our criterion from Section 4.2.1 that we want to have a model based on primary outcomes, this means that there is a quite limited number of data points on which an AI model can be trained. This makes it hard to estimate certain model parameters and create stable models, which is why we saw that many FI modelers opted for the non-primary FEWSNET data. Expert input could alleviate this problem by "warmstarting" the model and stabilizing it, as is amongst others explained by Wang et al. in [50].

A last reason why we first need to focus on quantifying projection assumptions, is based on our requirements for Agency and Meaningful Human Control. If we want our users to be informed by a model as to whether or not their analysis is still valid, they must first understand what their analysis and its assumptions explicitly mean before they can properly evaluate a model's output for its accuracy and reliability.

A sharp reader will note that we are discussing above explicitly projection assumptions instead of assumptions about the evolution of contributing factors. The reason for this narrowed scope is that we estimated that it would be challenging enough to let IPC people quantify assumptions that they had already an explicit idea about rather than sow confusion by talking about a larger set of assumptions that they had not considered up to that point.

So, in the next chapter, we will move on to the next phase of our project plan, "Develop," and discuss our experiment on letting analysts quantify projection assumptions by setting thresholds.

Chapter 7

Experiment: Can IPC analysts set thresholds for projection assumptions?

7.1 Introduction

In our problem analysis, we outlined our proposal on how the IPC AFI monitoring process can be improved so decisions on analysis updates can be made (a) based on systematic data-driven argumentations and (b) on time. An important component of this plan was to build a system that helps to make better-quantified risk factors by letting analysts set thresholds on mathematical representations of contributing factors. These outputs are not only essential for the development of any future AI for analysis monitoring, but also allow users to set boundaries within which a possible AI implementation would be allowed to function by indicating where they are and aren't sure that the projection assumption is still correct. In this way, these results can also help work towards our design requirement "Users should set the boundaries within which AI is allowed to function" of Meaningful Human Control.

The idea of setting thresholds was a recurring theme throughout the clarification and ideation phase of our design process. However, since the first time this idea was pitched to members of the IPC, the question of whether analysts would be able to set these thresholds has always been raised. In other words, did they have the necessary skills, expertise, and/or information to understand the question these thresholds pose, and are they able to give a meaningful answer? Following our other Human Control design requirement ("Users should not be asked to do tasks that they might not be capable of, or they should be supported in the process by the tool or other people helping them."), we concluded that an important first step towards implementing our monitoring process was to test whether this was the case or not.

Therefore, in this next phase we ran an experiment with analysts from Malawi on whether they could retroactively set thresholds on projection assumptions from the last AFI analysis they had performed for Malawi. We chose to run the experiment with Malawi's IPC team for the great expertise of their TWG and the richness of insights that Malawi can offer due to its seasonal variance, regional differences, and long-term influences of climatic factors.

This experiment allowed us not only to test the expertise of the analysts but also to analyze whether such thresholds would actually be useful during decision-making by comparing the thresholds afterward with the current data on the contributing factors. Furthermore, it also let us answer other questions such as: On which level(s) should the thresholds be set and what input should be prepared by the analysis leads?

In the rest of this chapter, we will outline our experiment setup and threshold design and discuss the results of the user study.

7.2 Experimental questions

For this experiment, many different factors could have been considered and investigated. After some refining, we have decided on the following seven questions to structure and guide our experiment setup. Below, we will highlight each of them and explain their relevance to our design and research process.

Q1 How difficult is it for analysts/facilitators to set thresholds on projection assumptions?
As doubt was raised by IPC members on whether analysts would be able to set these thresholds, the first important point is to investigate how hard they find it and what information or training they else need to be able to set them.

Q2 How difficult do analysts/facilitators find it to reach consensus on these thresholds?
As the IPC process is based on consensus, the setting of thresholds would also ideally be done in consensus. Therefore we want to also let them discuss their results and reach a technical consensus and see how difficult or easy this is for them. Moreover, as these thresholds are a method to make the interpretations of the projection assumptions explicit, this means that if they can't reach consensus, that the analysis might also be based on different assumptions about the same thing.

Q3 How much do thresholds for the same unit and contributing factor differ between analysts and between the consensus outcomes?
For the same reason as mentioned above, it is interesting to see how much interpretations (e.g. thresholds) differ for the same projection assumption. If they are very different, what does that say about the robustness of the AFI analysis outcomes? It is also interesting to see how the consensus process influences the final plenary outcomes. e.g. Do they converge to some kind of average? Do they completely gear towards the person with the strongest opinion? Do they get narrower or wider as an effect of the discussion?

Q4 How much do thresholds for the same contributing factor differ between units?
One of the questions that arose was on which level of aggregation these units should be set, e.g. on a national, regional or analysis-unit level. If it turns out that thresholds for units are very similar within regions or even within the country, this might mean that it could be set on a higher level of aggregation, which would save a lot of time and adhere to our Efficiency design principle.

Q5 How much are the thresholds of analysts influenced by input from facilitators?
Normally in the IPC analysis process, the facilitators provide the analysts with national or regional projection assumptions which they are then expected to adapt to projection assumptions for their own analysis units. The same structure would be possible for setting thresholds; the facilitators could give regional or national thresholds to the analysts to help them set the thresholds for their own units. However, it is a question how steering these regional inputs will be for the analysts. Will they simply take over the regional thresholds in this case? Or will the context of the unit have a stronger influence? If the former scenario happens to be the case, this can either imply that it is sufficient to just let facilitators set the thresholds, or that facilitators should not give their input in advance, or that some other measure should be taken depending on how (un)desirable this effect deemed.

Q6 How useful are the results for decision-making based on risk factor monitoring?
As our formulated goal of section 6.2 was to change the monitoring process to improve decision-making on analysis updates, it is of course necessary to see whether these thresholds actually are useful for informing analysis updates.

Q7 How much added value can the exercise of setting thresholds have during an actual AFI analysis?
However, aside from being useful for decision-making on analysis updates, we also envision that the exercise of setting thresholds might improve the (shared) understanding that analysts have of their analysis by making them think about their implicit assumptions. If this hypothesis is true, it would be a nice bonus and an extra reason to add this step to the analysis process.

7.3 Threshold design

A nontrivial matter in the design of our prototype and the experiment was to decide which thresholds we would ask the participants to set and what these thresholds would represent. We have ended up with two types of thresholds: validity and invalidity thresholds. See Figure 7.1.

The validity thresholds represent the first and most intuitive question to ask about a projection assumption: "Within which bounds are you certain your **projection assumption** is still **correct**?". The invalidity thresholds are more about exploring the edges: "Outside which bounds has the projection assumption become so incorrect, that you're certain that the **analysis** has become **incorrect** as well?"

The choice for the invalidity thresholds was inspired by the discussions we had in our interviews in Chapter 5. Some of the interviewees raised the problem that if you know that your assumption has become invalid (e.g. when a validity line is crossed), this doesn't necessarily mean that your whole analysis has become invalid as well. For this, you need to analyse whether this change in contributing factor(s) has created such a different situation that the FI classification has shifted as well. This implied that using only validity thresholds would not be sufficient for decision-making. However, in those same interviews, there were also situations mentioned where a large enough shift in a contributing factor would certainly cause the need for an analysis update. For example, if the food prices became much higher than what was expected. That is what led us to the idea of invalidity thresholds.

With the invalidity thresholds, there is also the added bonus of adding human control for any possible use of AI in decision-making on analysis updates. By indicating when we are certain that the projection assumptions are correct and when they aren't, we demarcate a zone of uncertainty between these two scenarios. This zone represents the situation where we are not sure what effect the change of contributing factors will have on the validity of the analysis without investigating further. In this space, we create relevance for an AI to give added insights as we indicate the limits of our own knowledge on knowing what will happen.

A question that had to be addressed with designing the validity thresholds, was when we would consider an analysis to have become invalid. Based on our interviews, the answer for this was quite simple: when the phase classification has shifted at least one phase for one or more regions. However, determining for one contributing factor whether it would shift whole phase classification would possibly be a too large cognitive step to make. Therefore, based on the advice of the IPC members who facilitated the experiment, we decided to use the validity of the corresponding food insecurity alignment[1] as a proxy for the validity of the analysis. This meant that we asked instead of when the phase classification would become invalid, when the corresponding alignment for that contributing factor would become invalid.

7.4 Method

7.4.1 Outline

To answer our research questions, we have divided 18 analysts into groups of 3, so we could compare results for different regions, units of analysis, contributing factors, and scenarios with or without facilitator input. This led to the setup illustrated in Figure 7.2. In Scenario 1, each analyst group is asked to set thresholds on two contributing factors (maize prices and rainfall anomaly) for two analysis units in the same region. In Scenario 2, the groups are asked to do the same, except that before they set the thresholds, they are shown a set of thresholds for their region that was set by a group of three facilitators. After these thresholds were set, we discussed them with two analysis leads (who are also facilitators) to assess their usability for decision-making.

[1] In the IPC process, they give each Food Security dimension (see Section 4.1.1) also a separate phase classification. They call these classifications food security alignments

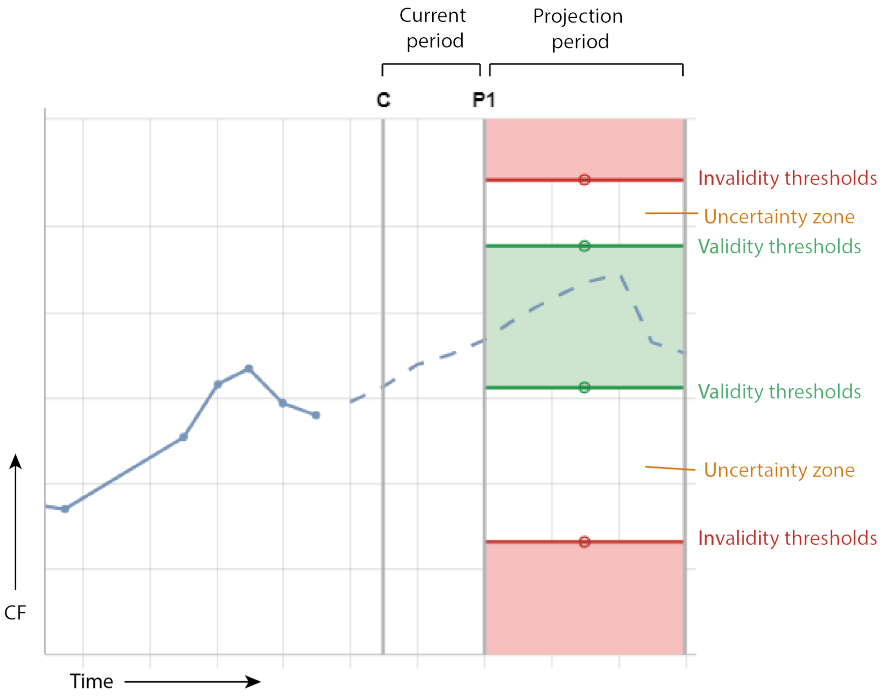


Figure 7.1: Validity and invalidity thresholds for a projection assumption

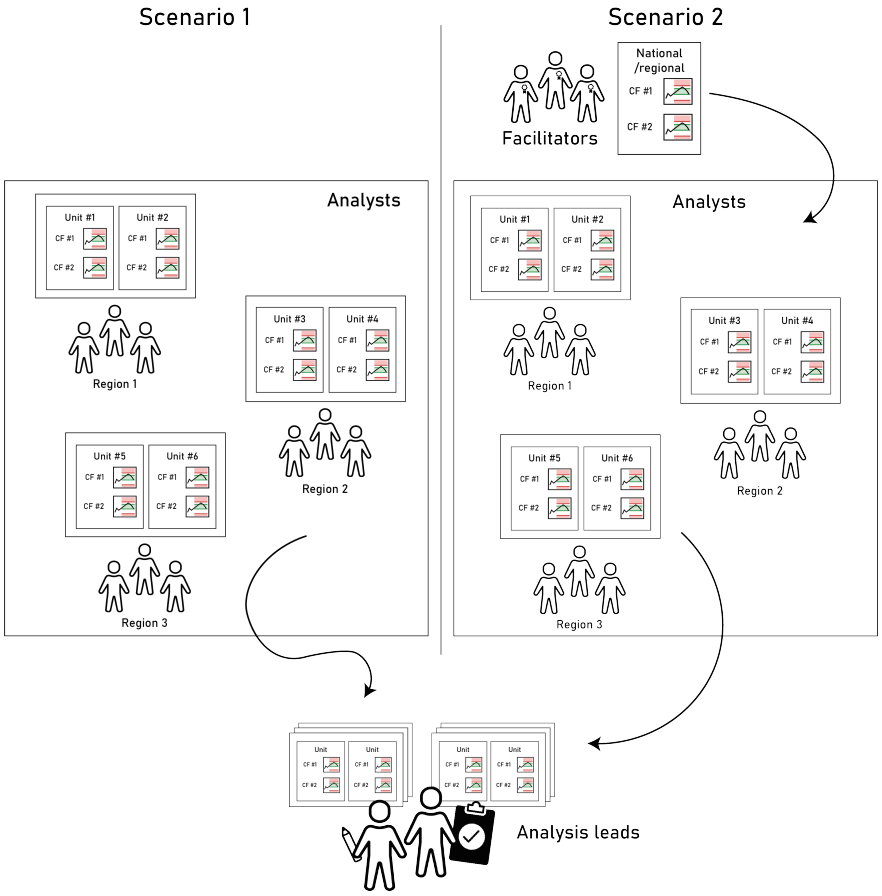


Figure 7.2: Experiment overview

For each group, we organized a separate virtual session of two hours to do the threshold exercises. Each session started with an explanation about the (in)validity thresholds and a demo of the prototype in which they could set them. After that, we discussed the national/regional projection assumptions for maize prices and rainfall, before we discussed the projection assumptions for their specific units. This was, in the case of rainfall, especially relevant as not all units had specified unit-level projection assumptions for that, even though it was indicated for most of the units to be a risk factor. As we wanted to compare the analysts' individual threshold estimations with those set in consensus, we started the threshold exercises by letting the analysts all set thresholds for each contributing factor and unit individually. After that exercise, we let them fill in a first survey about their confidence levels and how difficult they found it. Then the analysts were asked to discuss all the thresholds in plenary and reach a technical consensus on a shared set of bounds. To compare the effect of this consensus process on their perceived confidence and expertise, we let them fill in again a survey about it after they had reached a consensus. We closed the session with a last survey about their opinion of the feasibility and added value of setting such thresholds. See Figure 7.3 for an overview of the tasks that the analysts were asked to do.

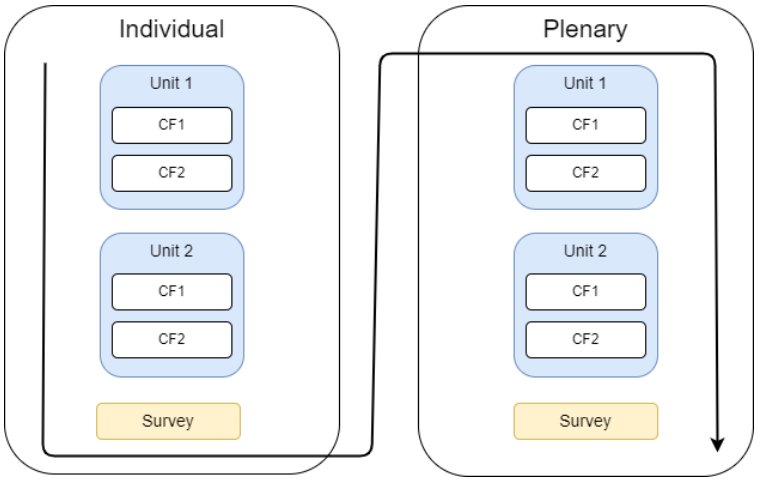


Figure 7.3: Setup for each group session

Due to some technical difficulties, time constraints and scheduling issues, it varied between sessions how many participants were available, whether they had enough laptops to do the individual threshold exercises, and whether we had enough time to set thresholds for both units. We were, therefore, not able to do all the exercises with all the participants. See an overview in Table 7.1 on which threshold exercises we managed to do for which units.

Scenario	Participants	Region	Set thresholds for	Individual		Plenary	
				1st unit	2nd unit	1st unit	2nd unit
1	3 analysts	Region 1	Unit 1, 2	yes	yes	yes	yes
2	1 analyst	Region 1	Unit 1, 2	yes	yes	no	no
1	3 analysts	Region 2	Unit 3, 4	no	no	yes	yes
2	1 analyst	Region 2	Unit 3, 4	yes	yes	no	no
1	1 analyst	Region 3	Unit 5, 6	yes	no	no	no
2	3 analysts	Region 3	Unit 5, 6	yes	no	yes	no

Table 7.1: Overview of participants and thresholds that were set in the experiment

7.4.2 Contributing factor selection

As mentioned in the outline, we had chosen to test the thresholds on two contributing factor indicators: maize prices and 1-month rainfall anomaly. The 1-month rainfall anomaly is the difference in rainfall that has fallen in a given month compared to the long-term average for that month. In graphs, this is represented

in percentages; a value of 100% means that a month was completely the same as average, and 50% means that there was twice as less rainfall than normal. The choice for these indicators was made in consultation with the analysis leads, who confirmed that these are two of the most important food insecurity drivers for that IPC country. Moreover, both contributing factors had publicly accessible data that was regularly updated, making them also viable candidates for any future risk factor monitoring system.

These two contributing factors also allowed us to investigate whether there is a difference in difficulty and doability between setting thresholds for those two contributing factors. Maize prices have namely a much shorter and clearer cause-and-effect chain towards any changes in food insecurity than rainfall anomalies might have. If the maize prices rise, this means poorer people can afford less maize, which leads to lower food consumption. In contrast, low rainfall timed wrongly might cause a bad harvest which might influence current and future household stocks and food prices. As one can see, in the rainfall cause-and-effect, there are many more ifs and buts and other variables that could influence the impact of low rainfall on food insecurity. The same goes for too high rainfall and flooding. That is why we hypothesize that analysts will have a harder time setting thresholds on rainfall than on maize prices.

7.4.3 Prototype

To facilitate the threshold exercise, we developed a web interface where analysts could view the historical and projected data of a contributing factor for a specific analysis unit and set their (in)validity thresholds for the specified projection period. This tool enabled us to automatically collect all the thresholds that were set by the analysts, as well as facilitate an interface where they could set the thresholds for the group. See Figure 7.4 for a screenshot of the UI. For both rainfall and maize prices, we used the Humanitarian Data Exchange API to automatically visualize the historical data for a given unit. For the projected maize price data, we used the same projected data that was used to inform the analysts during the original AFI analysis. For the rainfall data we did not have any projected data, as no projected data was shown as well during the actual AFI analysis.

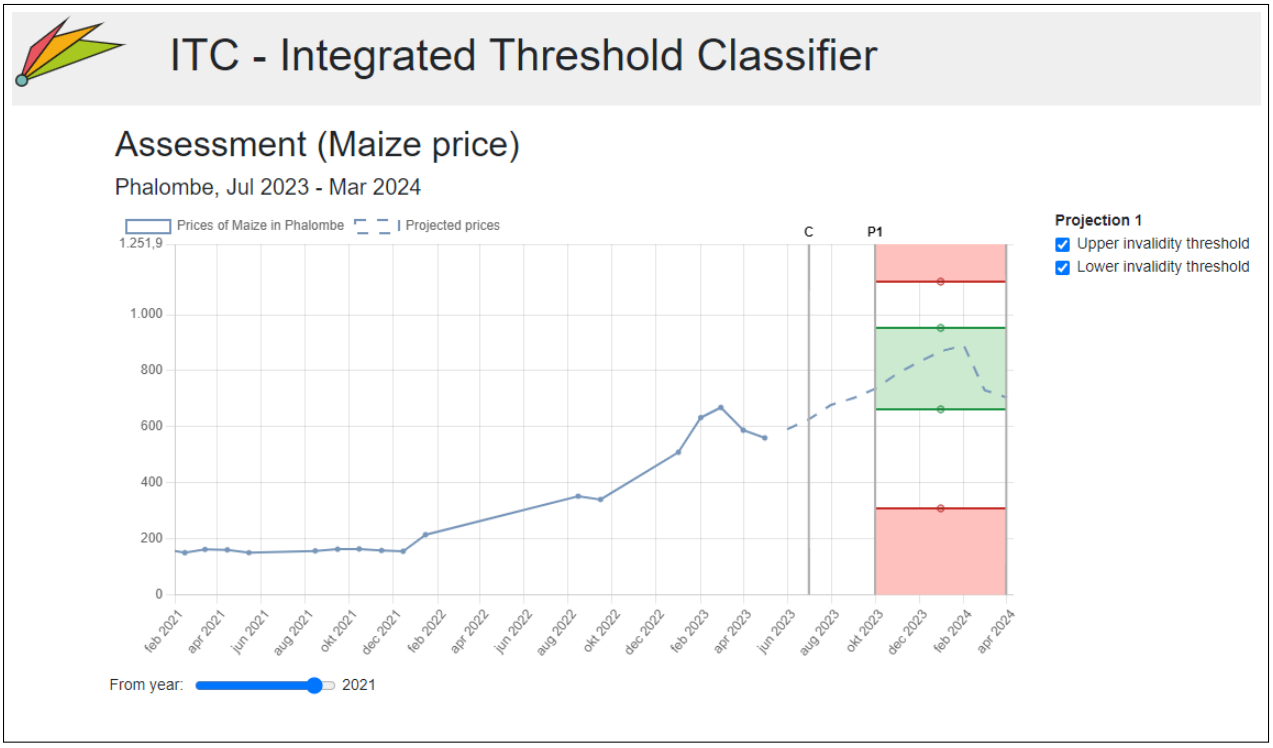


Figure 7.4: Screenshot of the ITC: Integrated Threshold Classifier, a prototype developed as proof of concept for quantifying projection assumptions

7.4.4 Survey and analysis

To answer some of our experimental questions of Section 7.2, we developed a set of survey questions to test them, as well as how we further wanted to analyse them. See Table 7.2 an overview of our survey questions. Where possible, we used semantic differential scales for our closed questions to prevent the acquiescence bias related to Likert scales [34].

After individual exercises		
Question	**Scale type**	**Scale**
How confident or unsure are you about the accuracy of the validity thresholds that you have set for the maize prices for both units?	Semantic differential	1= Very unsure 7= Very confident
How confident or unsure are you about the accuracy of the invalidity thresholds that you have set for the maize prices for both units?	Semantic differential	1= Very unsure 7= Very confident
How confident or unsure are you about the accuracy of the validity thresholds that you have set for the rainfall for both units?	Semantic differential	1= Very unsure 7= Very confident
How confident or unsure are you about the accuracy of the invalidity thresholds that you have set for the rainfall for both units?	Semantic differential	1= Very unsure 7= Very confident
How difficult or easy did you find it to set these thresholds?	Semantic differential	1= Very difficult 7= Very easy
What aspects of the threshold exercise made it easy or difficult?	Open question	
I had enough information to be able to set these thresholds	Likert	1= Strongly disagree 5= Strongly agree
I had enough expertise to be able to set these thresholds	Likert	1= Strongly disagree 5= Strongly agree
Is there any extra information or training we could give that would have made the exercise more doable?	Open question	
After the plenary session		
Question	**Scale type**	**Scale**
How confident or unsure are you about the accuracy of the validity thresholds you have reached as a group for the maize prices?	Semantic differential	1= Very unsure 7= Very confident
How confident or unsure are you about the accuracy of the invalidity thresholds you have reached as a group for the maize prices?	Semantic differential	1= Very unsure 7= Very confident
How confident or unsure are you about the accuracy of the validity thresholds you have reached as a group for the rainfall?	Semantic differential	1= Very unsure 7= Very confident
How confident or unsure are you about the accuracy of the invalidity thresholds you have reached as a group for the rainfall?	Semantic differential	1= Very unsure 7= Very confident
How difficult or easy did you find it to reach consensus with the other participants on setting the thresholds?	Semantic differential	1= Very difficult 7= Very easy
I agree with the thresholds that were set by the group	Likert	1= Strongly disagree 5= Strongly agree
I feel that as a group, we have enough information to be able to set these thresholds	Likert	1= Strongly disagree 5= Strongly agree
I feel that as a group, we have enough expertise to be able to set these thresholds	Likert	1= Strongly disagree 5= Strongly agree
After the whole session		
Question	**Scale type**	**Scale**
How doable was it to do these quantifications several months after the AFI analysis has taken place?	Semantic differential	1= Not doable at all 7= Very doable
How doable would it be to do these quantifications during an actual AFI analysis?	Semantic differential	1= Not doable at all 7= Very doable
To what extent did the exercise of setting thresholds create new insights for you about the AFI analysis and its implications?	Unipolar	1= To no extent 7= To a large extent
If you gained any insights, what were those?	Open question	
How much added value would the exercise of setting projection assumptions have in a real AFI analysis?	Unipolar	1= None at all 7= A lot
If you see any added value, what would this be?	Open question	
Do you have any suggestions for further improving the tool and/or process for quantifying projection assumptions?	Open question	

Table 7.2: Survey questions after individual exercises

7.5 Results

In the following section, we will present our outcomes structured according to the experimental questions we have posed at the start of this chapter.

Q1: How difficult is it for analysts/facilitators to set thresholds on projection assumptions?

To measure the self-perceived ability and confidence of analysts in setting thresholds individually, we have let them fill in the first part of the survey in Table 7.2 after they had performed the individual threshold exercises. To get a better insight into what makes the exercise more difficult and how we could mitigate that, we also included two open questions in this survey. As not everyone was able to do the individual exercises (due to circumstances explained above), we have gotten a total of 9 responses from analysts and 3 responses from facilitators on this survey. In Figure 7.5, we have visualised the analysts' responses to the question on how confident the analysts were in the accuracy of their thresholds. In Figure 7.6, one can find the analysts' responses on how difficult they found it and whether they had enough information and expertise.

Findings

A first important observation is the results skew more towards neutral to positive statements about their ability and confidence to set thresholds. This shows a promising view on whether analysts can better quantify their projection assumptions. However, it is important not to disregard the heterogeneity in most of these answers. This reflects namely also the heterogeneity of the analysts that the IPC works with. Due to their different backgrounds and domains of expertise, it is understandable that this will also cause differences in how easy or difficult they find it to quantify their projection assumptions. In any implementation of projection assumption quantification, it is therefore important to take this heterogeneity into account when preparing the analysts for such an exercise.

To test our hypothesis that rainfall thresholds will be harder to set than thresholds on maize prices, we also compared the differences of analysts' and facilitators' confidence levels between these two contributing factors per type of validity line. While we did find a close-to-significant average difference of 0.92 (SD=1.62) between confidence in validity lines for Maize prices and Rainfall ($t(11) = -1.96$, $p=.076$), no significant difference was found between the invalidity line confidences. This might be understood since invalidity lines were generally perhaps harder to set.

As IPC leads feared that the invalidity thresholds might be too hard to set or understand for the analysts, we also tested whether there was a significant difference between confidence levels in validity and invalidity thresholds by performing a paired t-test on these differences per contributing factor. For the validity and invalidity thresholds for Maize prices, we found an average difference of 1.33 (SD=2.0), which was close to significant ($t(8) = -1.99$, $p = .080$). For the difference between validity and invalidity lines for Rainfall, the results were inconclusive.

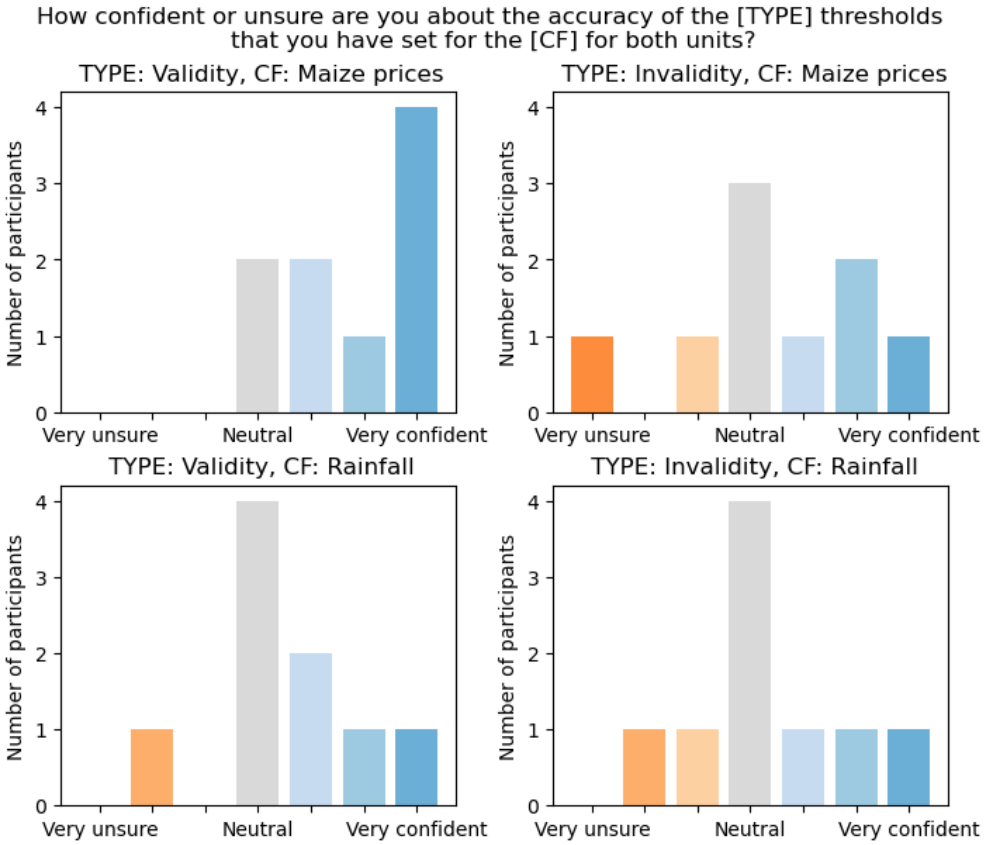


Figure 7.5: Analysts' confidence level for the different thresholds they had set individually (n=9)

Q2: How difficult do analysts/facilitators find it to reach consensus on these thresholds?

To measure how the analysts experienced the plenary exercise and how that influenced the confidence in the thresholds, we let them fill in the survey from the second part of Table 7.2. We asked them the same questions about confidence and ability but then in relation to the group's thresholds and competence. This allowed us to compare the results of the first survey and explore what effect the plenary session had on them. In Figure 7.7, the results are shown for their confidence levels for the group thresholds. In Figure 7.8, one can find the results for the other closed questions about the ability of the group and their support for the results.

Findings

Generally, the group exercise was received positively. As can be seen in the top left plot in Figure 7.8, analysts found it generally easy to reach a consensus on a shared set of thresholds. The one notable variation on this sentiment by the participant who found it "very difficult" represented the only instances where there was a longer discussion on where to set the (in)validity bounds, which was resolved in the end by meeting each other somewhere halfway. That can have contributed to the striking unanimity with which everyone agreed or strongly agreed to the thresholds they had reached as a group, as can be seen in the top right plot.

Furthermore, the results also suggested that the analysts viewed the ability of the group to set thresholds more favourably than their individual competence. Testing for these effects, we indeed found that both their assessments of whether they had enough information and whether they had enough expertise were higher for the group than for themselves: Average difference of information sufficiency was 0.44 (SD=0.53) ($t(8) =$

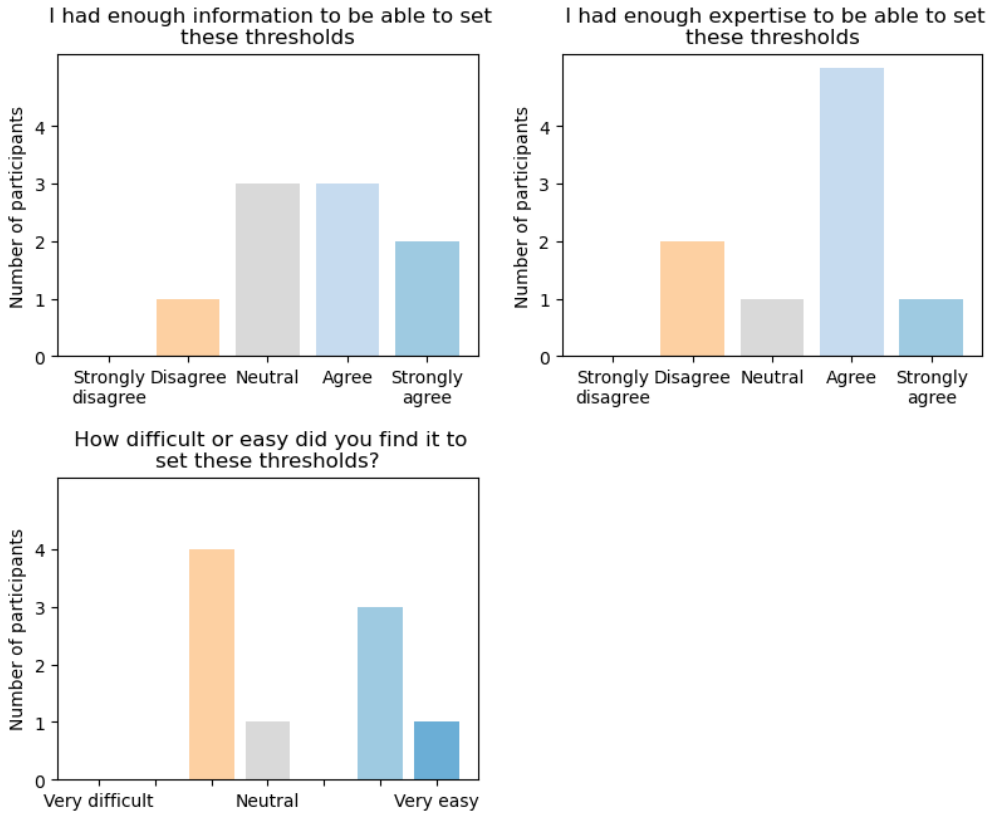


Figure 7.6: Analysts' self-perceived capability for setting thresholds individually (n=9)

-2.53, p=.035), average difference of expertise sufficiency was 0.89 (SD=0.78) (t(8) = -3.41, p=.009).

These differences also resulted in (close to) significantly higher confidence levels for the invalidity lines they had set as a group. The average difference in confidence for Maize invalidity thresholds was 0.78 (SD=0.83) (t(8)=-2.8, p=.025), the average difference in confidence for Rainfall invalidity thresholds was 0.33 (SD=0.5) (t(8)=-2.0, p =.08). For the validity thresholds no significant differences were found. That this difference is only clear for invalidity thresholds might be explained by the fact that some analysts still had some confusion about what the invalidity thresholds precisely entailed. Often in these cases, they obtained a better shared understanding of their meaning through the explanations of one of the other participants. This improved understanding could then have caused also the improved confidence in the set thresholds.

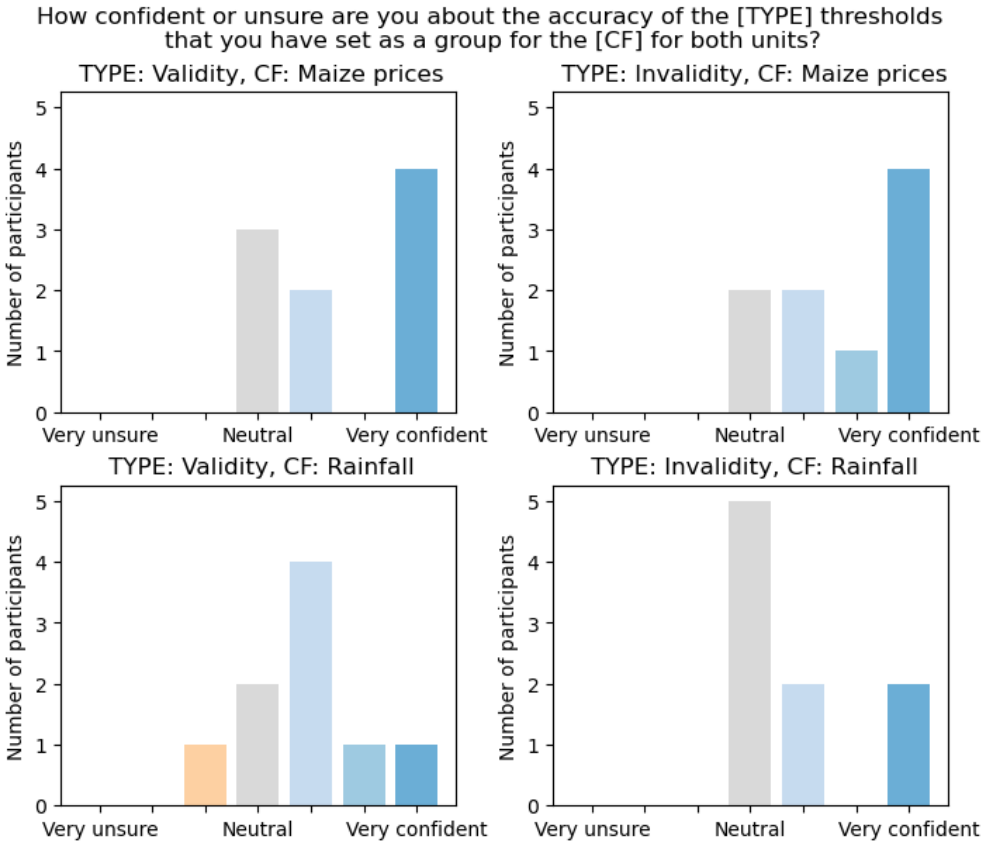


Figure 7.7: Analysts' confidence level for the different thresholds they had set as a group (n=9)

Q3: How much do thresholds for the same unit and contributing factor differ between analysts and between the consensus outcomes?

To answer this question, we collected the thresholds that each of the participants had set for the individual exercises from the prototype database, as well as the thresholds that were set by them for the group. In some sessions, only one participant was able to join and do the exercises. In those cases, the plenary outcome became, by default, their individual results. In Figures 7.9a, 7.10a and 7.11a, one can view the results respectively for regions 1, 2 and 3. On the x-axis of each graph, there is a tick for each participant that participated in that session. If they have done the individual exercises, then their individual thresholds are visualised above. If there are no individual thresholds visualised for them, this means they only participated in setting the plenary thresholds.

Findings

As can be seen in the plots, the thresholds can vary quite much between different individuals and between the different consensus outcomes. While this might cause worry about the consistency of the original projection assumptions, this can't be that easily concluded from these results. There are namely other factors that could have led to these differences, apart from inherently different implicit interpretations of the projection assumptions. First of all, as stated above: not all analysts had a perfect understanding yet of what the invalidity thresholds entailed (and what their implications were) during the individual exercises. This might cause some of the larger differences in invalidity thresholds. Moreover, at some points there was during the individual exercises confusion as well about what the rainfall anomaly meant exactly and how projection assumptions translated to those.

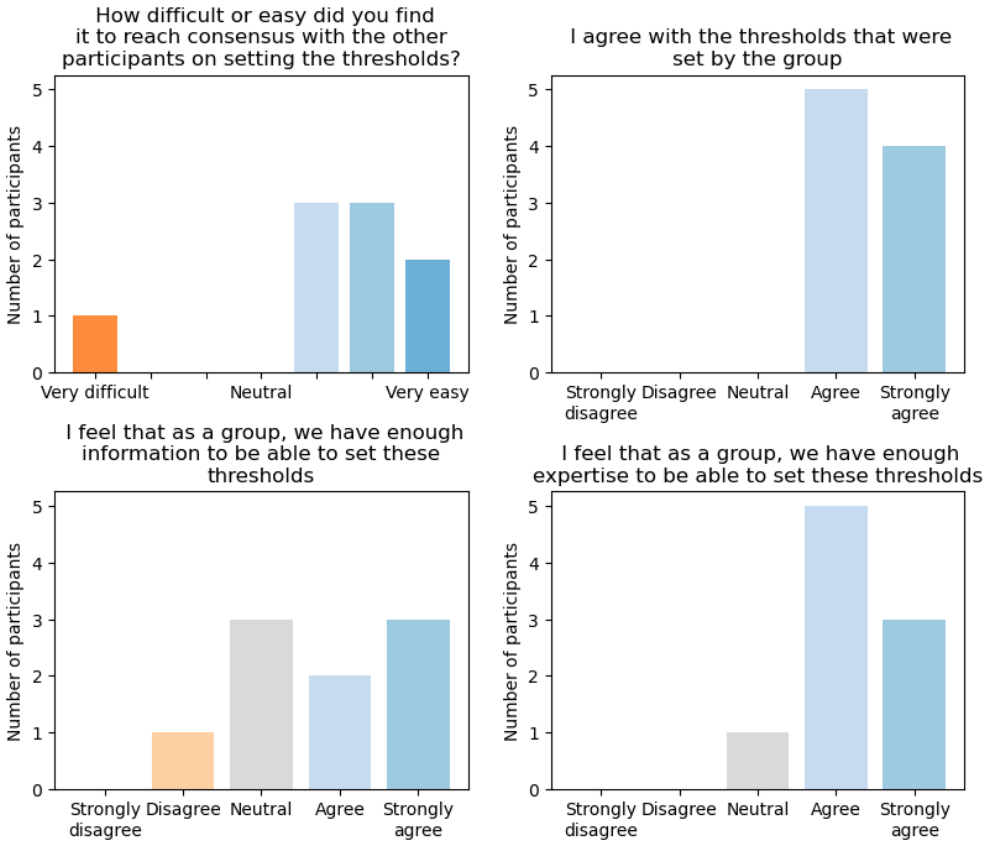
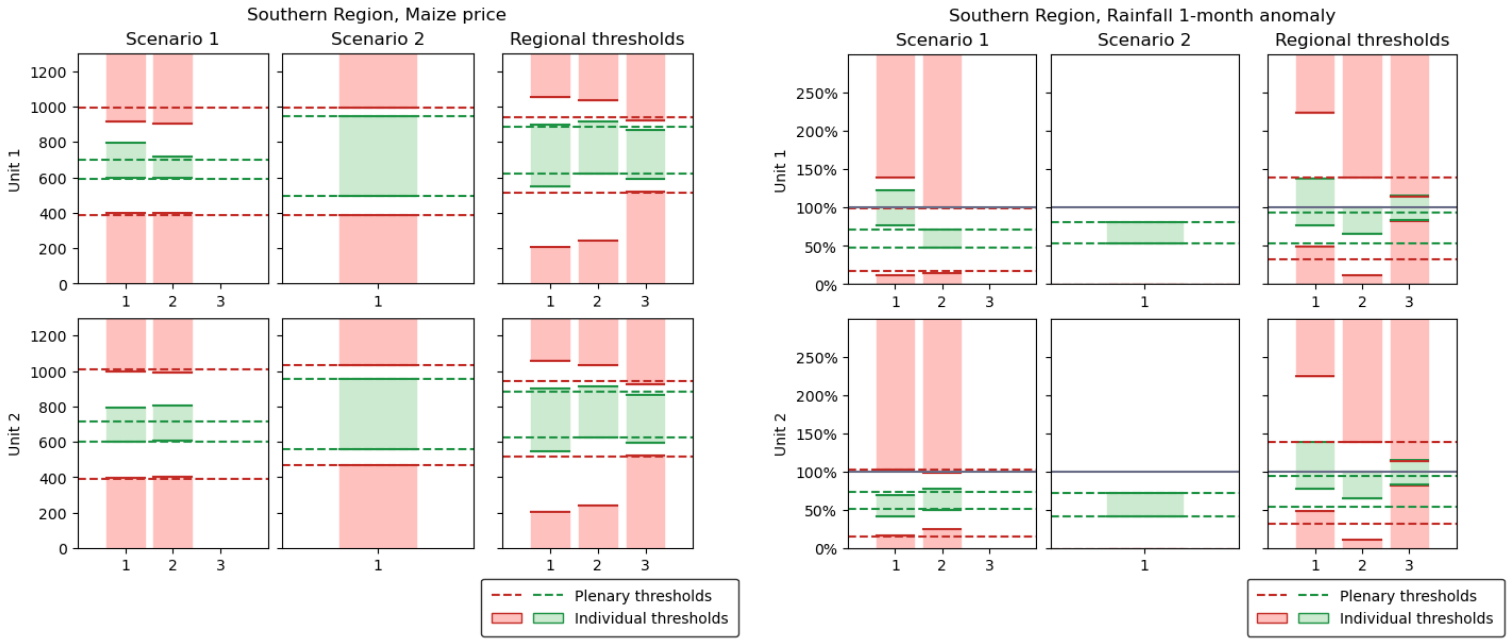


Figure 7.8: Analysts' perception of the capability of the group to set thresholds (n=9)

Given these differences, if such a projection quantification were to be implemented, it would be recommended to give analysts an extensive training together so they get a feeling for these thresholds and a share understanding of what they mean before they set them for an actual analysis. This recommendation also came forward from the open questions on how we could improve the setup.



(a) (b)

Figure 7.9: Individual and plenary thresholds for the Southern Region

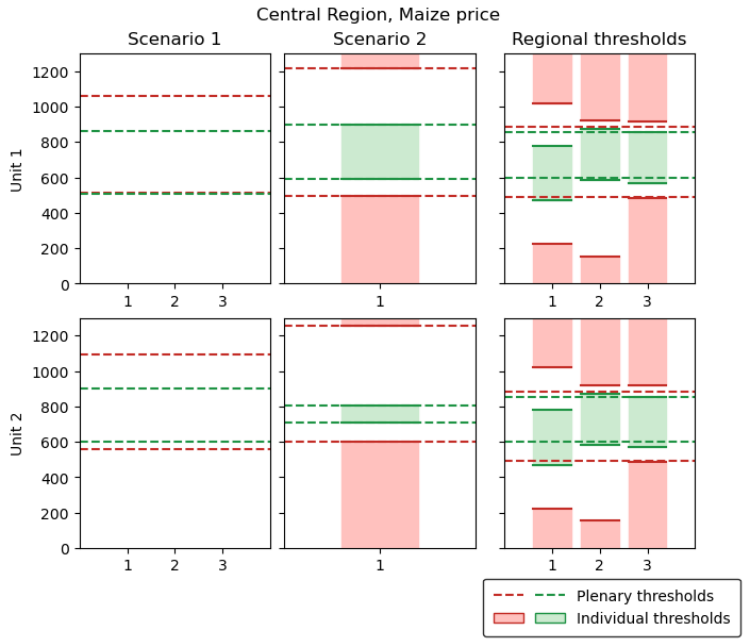
Q4: How much do thresholds for the same contributing factor differ between units?

In Figure 7.12, we have plotted the plenary thresholds for each unit in each region.

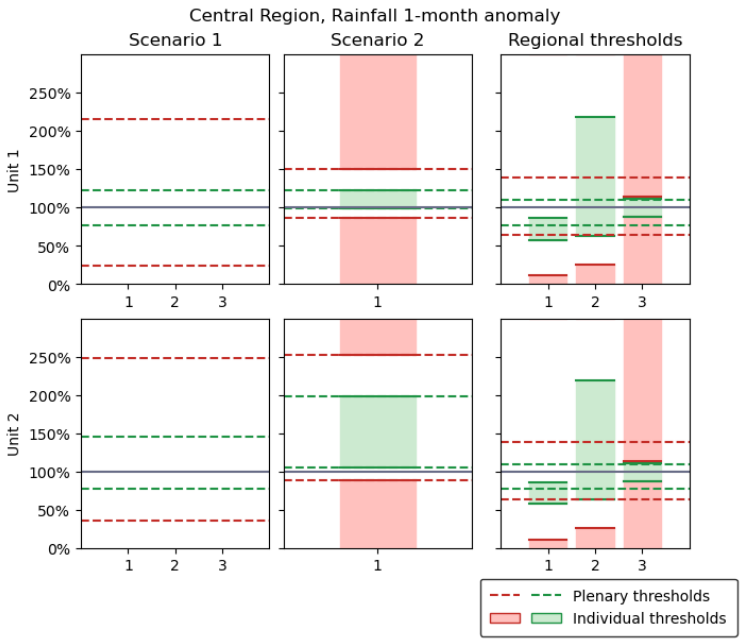
Findings

As one can see from the data, it is currently hard to properly compare thresholds between units, as the differences between the scenarios within units are generally much higher than the differences between the units.

However, when comparing the unit assessments that were done by the same groups (for example, the assessment of the scenario 1 group for the two units in the Southern Region), we do see, in many cases, a striking similarity. This similarity was often supported by the explicit assertions of the analysts that their two units could have (almost) the same thresholds, as they were in the circumstances anyway. However, this was not a given for all regions. As one can see in the Central Region, for example, within the same group, they had given the different units quite significantly different thresholds. This lets us conclude that it is not recommendable to set thresholds solely on a regional level, as the uniformity of assumptions within regions is not guaranteed throughout the country.

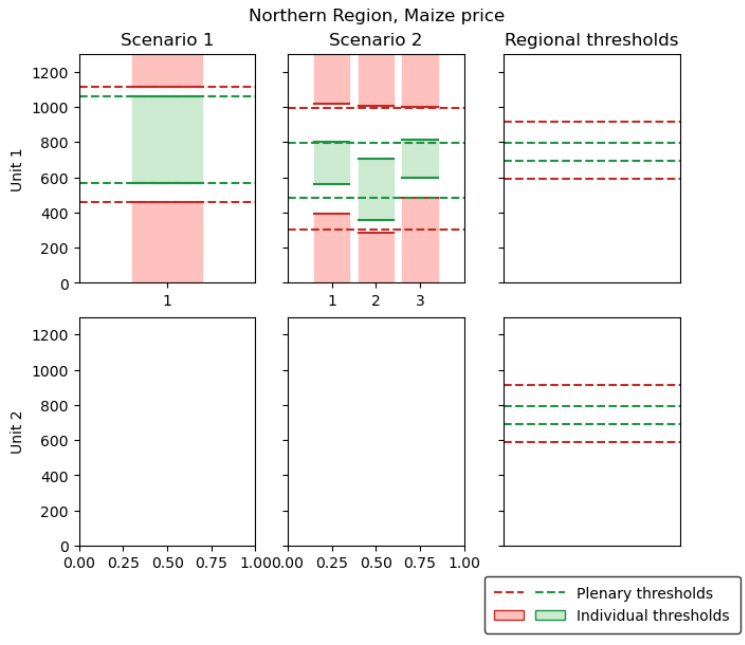


(a)

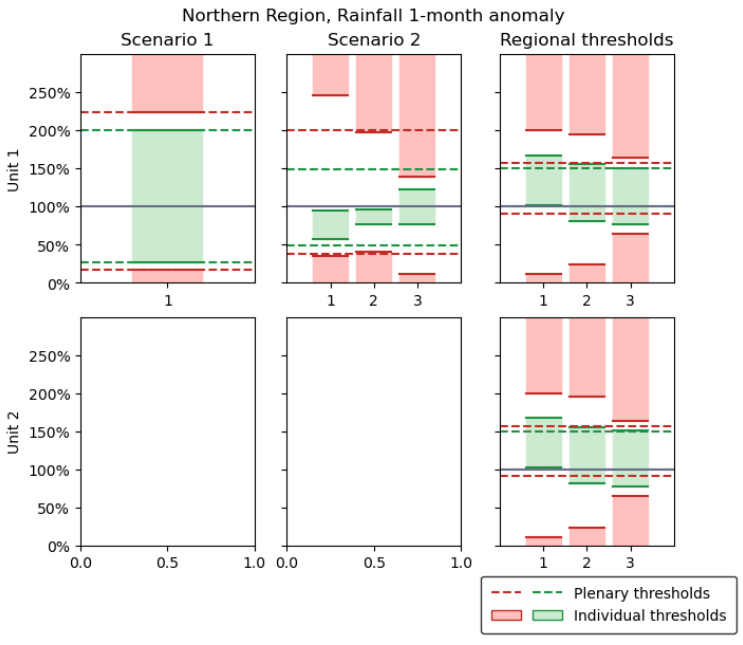


(b)

Figure 7.10: Individual and plenary thresholds for the Central Region

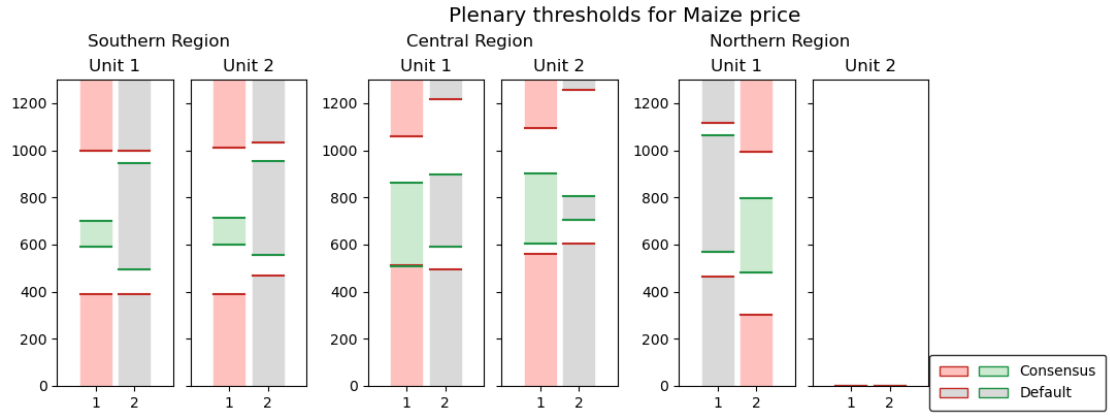


(a)

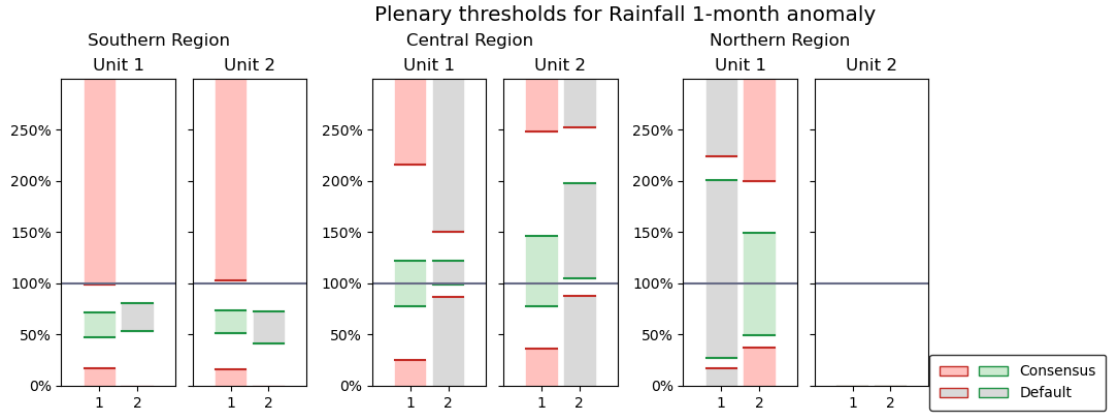


(b)

Figure 7.11: Individual and plenary thresholds for the Northern Region



(a)



(b)

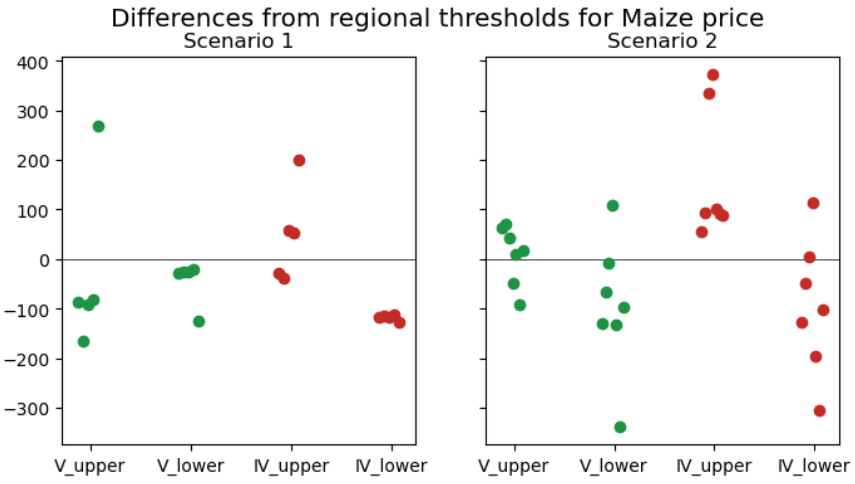
Figure 7.12: Plenary thresholds for each region. On the x-axis are the thresholds for scenarios 1 and 2. Grey-coloured thresholds are thresholds that became plenary by default due to there being only one person in those respective sessions.

Q5: How much are the thresholds of analysts influenced by input from facilitators?

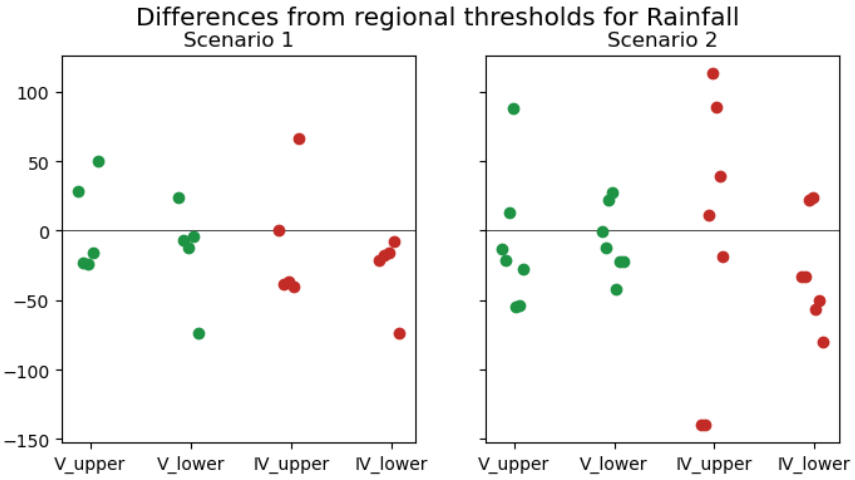
We investigated this question by subtracting from each individually set threshold its counterpart from the corresponding regional thresholds. See Figure 7.13.

Findings

As one can see, there is not a clear difference to be found between the region deviations of Scenario 1 and Scenario 2. To assert ourselves, we tested these differences with individual t-tests and found no significant differences. These lack of differences might be explained by the possibility that the facilitators had a less better understanding of the concept of the invalidity lines, as they were the first group to participate and our training was much improved upon afterwards, causing perhaps their validity lines to make less sense. Nonetheless, during the experiment sessions it did not seem like the analysts in Scenario 2 were actively considering the regional thresholds. This might imply that it would be better to not let the facilitators set regional thresholds at all, in order to remove unnecessary redundancy. However, one other analyst did remark that he was worried that not all analysts might be as adept at setting these thresholds and that some extra scrutiny or validation might be called for. In this case, properly considered and explained regional thresholds might be helpful for performing this validation.



(a)



(b)

Figure 7.13: Differences between individual and regional thresholds. Along the y-axis, each individual difference is plotted as a dot. The x-axis shows to which type of threshold they belong. Here, V_upper = upper validity line, V_lower = lower validity line, IV_upper = upper invalidity line and IV_lower = lower invalidity line.

Q6: How useful are the results for decision-making based on risk factor monitoring?

To estimate what impacts the thresholds might have had on decision-making, we compared our found thresholds with the actual values of the two contributing factors and their original projections from the analysis. See for an example the comparisons for Nsanje in Figure 7.14. We subsequently presented those results, as well as the (anonymized) results from the survey, to two analysis leads to discuss what implications they saw in these outcomes and how they generally looked at the usefulness of the tool.

Findings

We found that in the Southern Region, the maize prices had become much higher than what was projected during the analysis (See Figure 7.14). According to the analysis leads, this was caused by, amongst other factors, the 44% devaluation of the Malawi kwacha that happened in early November 2023 [11]. As they had

not anticipated this devaluation nor these extreme price peaks in the Southern Region (in the other regions the peaks were more in line with what was expected), the analysis leads concluded that an analysis update would have been necessary. However, the reason that one was not done at that time has had most likely to do with a lack of funding, as they surmised that the TWG would have been looking to do an analysis update with these prices.

The results for rainfall were much more erratic than one would have expected when looking at the thresholds of most participants. The projection period for the Southern Region started, for example, with a much higher peak than was expected, while for other regions, the rainfall dipped multiple times below the lower invalidity thresholds. However, while the rainfall crossed the regional invalidity lines significantly at different points, this did not lead the analysis leads to conclude that an analysis update was needed. That here the invalidity lines had less strong implications, was explained by them through its impact on two main phenomena: the impact on the harvest and the impact on flooding.

When considering the impact of rainfall on the harvest, it is important to look at the timing and the pattern of the rainfall before drawing conclusions. Within a period, the average rainfall could be normal for example, but if the rainfall within that period is very erratic, then it can still cause a crop to fail due to a prolonged dry spell. So, a high peak of rain at the start of the season does not necessarily mean a better harvest if it is followed by a prolonged lower-than-normal rainfall as what happened in the South. Moreover, any impact on the harvest will mostly be felt in the period *after* the projection period, so does not immediately necessitate an update for the projection period *itself*. When looking at the impact of rain on flooding, this also might not immediately merit an analysis update. The first thing that needs to be looked at here is measures to contain the flooding. Only if flooding has had a serious impact on food security will it become necessary to do an analysis update. So given these different phenomena, when using purely rainfall anomaly as an indicator, it is not possible to set very hard thresholds as the context and pattern of the rain as well as other factors, will really determine its impact on food security.

About the feasibility and usefulness of the tool:
In general, the two analysis leads were positive about the tool: they thought it would be good to start working with this kind of threshold application so that during an analysis, people don't just say "Yes, this is a risk factor that needs to be monitored", but also set thresholds so that they know something needs to happen if they are crossed. They also thought it was feasible because, according to them we are simply taking elements that are already there in the analysis and just reorganising them in a different way.

However, they pointed out that it would take quite some time and work to make something that could be incorporated into the analysis process. The first thing we need to do is to develop criteria for analysing these different indicators and some clear guidance on how to apply them to different situations. At the moment, analysts might namely misinterpret the rainfall graph, for example, and come to different conclusions due to this. So people need to be on the same page about the indicators, possible scenarios and their implications. Furthermore, we will also need to consider its impact on the planning of the analysis because it is usually difficult enough already to reach a consensus on the actual classifications without determining these extra thresholds. So there will also need to be careful consideration of how this will affect current practice and how many extra days will be needed to add such a step.

Q7: How much added value can the exercise of setting thresholds have during an actual AFI analysis?

To answer this question, we asked all participants to fill out a survey at the end of their session about whether they gained any new insights about the analysis and whether such a quantification exercise would provide added value during an actual AFI analysis. As an added check, we also asked how doable it actually was to set these thresholds, and how doable it would be to set such thresholds during a real-life AFI analysis. The results of this question can be viewed in Figure 7.15. In Figure 7.16, one can view respectively the results from the analysts answers and the facilitators' answers.

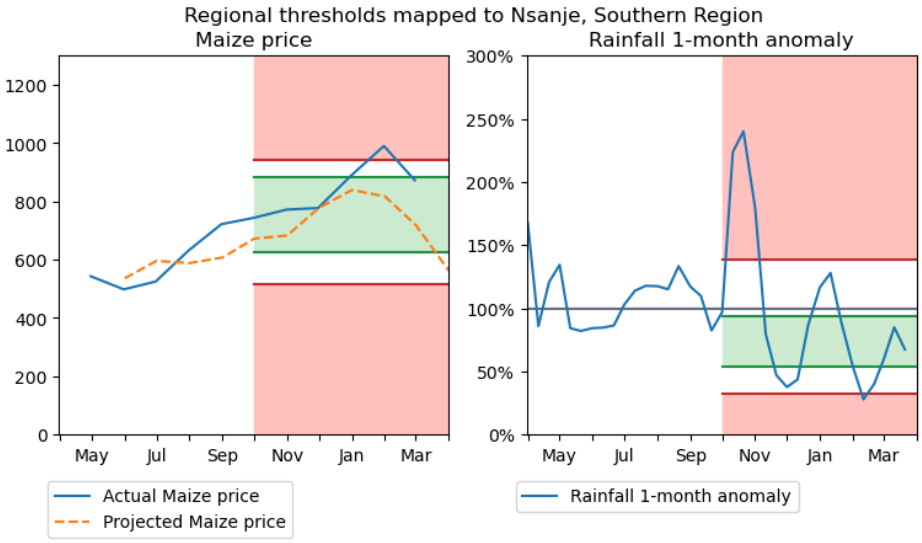


Figure 7.14: Actual price and rainfall developments of the past months for Nsanje, compared to the regional thresholds set by the analysis leads.

Findings

The results were remarkably positive. On the feasibility of setting thresholds in the current experiment setting, everyone responded either neutrally or favorably. When asked about the doability of setting thresholds during an actual AFI analysis, the responses were even more positive, leading to >70% of the participants saying that setting thresholds then would be very doable.
Apart from the exercise being feasible, they also saw it giving added value to the analysis. More than 60% of the analysts and all the facilitators stated that it gave them to a larger extent new insights and saw added value for doing the exercise in a real-life AFI analysis. The others were more neutral, and no one was actively opposed to the idea.

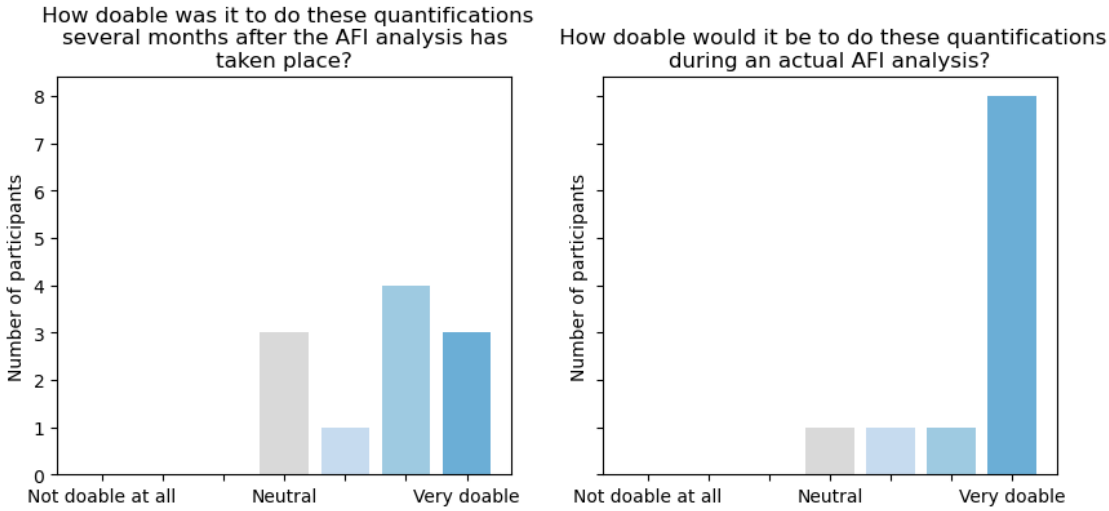
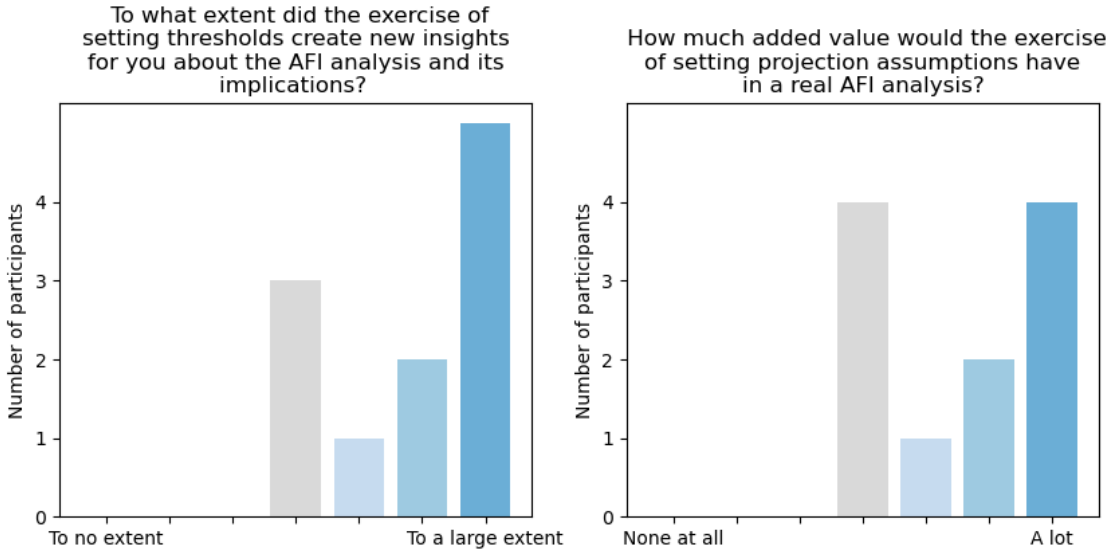
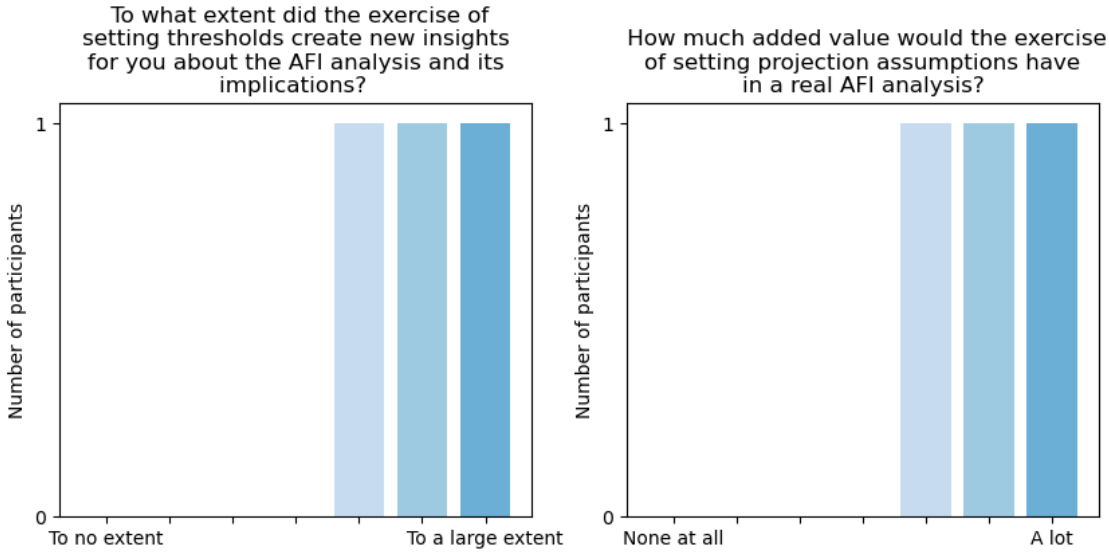


Figure 7.15: Feasibility of setting projection assumption thresholds as indicated by analysts



(a) Survey answers from the participating analysts (n=11)



(b) Survey answers from the participating facilitators (n=3)

Figure 7.16: Answers from participants about the added value of the threshold exercises and whether they gained new insights

7.6 Discussion & Conclusion

In this experiment, we examined whether analysts are capable of quantifying their projection assumptions by setting thresholds on the expected evolution of contributing factors. To this end, we introduced a prototype that, using validity and invalidity thresholds, lets analysts demarcate for which values they are certain about the validity and invalidity of their projection assumptions.

We found that while participants' self-assessed ability to set (in)validity thresholds and their confidence in them varied (see Q1), their general estimation of the doability of such an exercise was found to be considerably positive in Q7, especially if it was to be performed during an actual analysis.

Furthermore, in answering Q2, we discovered that the consensus process had a significant positive influence on the understanding and the confidence that analysts had in their thresholds as well as in the competence of the group in doing so. This also provides extra support for the consensus-based way of working of the IPC.

When comparing the different thresholds in Q3, we saw that there was a high variety in the thresholds between different groups for the same contributing factor and unit. However, this can be attributed to various different factors, so it's hard to make conclusive statements based on these results on the robustness of the original projection assumptions.

For Q4, we found that while some units in the same region might have very comparable thresholds, this was not a guarantee for all units. This leads us to conclude that any form of projection quantification should be done on a unit level in order to properly account for the different circumstances of each unit. This corroborates with what we found in our interviews in Chapter 5.

When investigating for Q5 whether the regional thresholds set by facilitators would have a large effect on the thresholds of the analysts who were shown them, we found that no such influence was actually perceivable. We see this as a positive outcome as it shows that analysts work with a higher level of autonomy than perhaps was presumed by the IPC.

When discussing Q6 with two analysis leads, we found that the thresholds would be useful in triggering action and letting people consider whether an analysis update was needed. For the Southern Region, it was even concluded that an analysis update would have been necessary, though its possibility was limited by other factors such as limited funding. However, to make such a system work, we would have to focus on creating good criteria and clear guidance on how to set these thresholds so teams are on the same page.

All in all, we found in Q7 that the responses about our threshold exercise and prototype were very positive. Participants gained new insights and saw considerately added value for doing such quantification during an actual AFI analysis.

7.6.1 Limitations & Future work

While the study's results are promising, several limitations must be considered. The first and most important factor is that the study was not performed during an actual analysis, so analysts might sometimes have reasoned about their thresholds using knowledge of what actually happened during the projection period. This setting also meant that while we were able to compare the thresholds with the actual data on the contributing factors, we couldn't properly investigate how such indicators would impact an actual monitoring process.

Furthermore, due to the short time we had for the training/introduction and due to not having the full analysis as context, the analysts had, in some cases, a lesser understanding of the thresholds and/or the indicators and were less able to fully consider their implications. This also resulted in less consistent thresholds than one would hope for.

Therefore, for future work, it would be of great value to test such a form of projection assumption quantification during an actual analysis and afterwards do a longitudinal study of how these thresholds impacted the monitoring process. Such a study could lay the groundwork for an actual first monitoring system and a real transformation of the IPC risk factor monitoring process.

Another factor that should be considered in any future work is the choice of indicators and how we set thresholds on them. As discussed in Section 7.4.2, using rainfall 1-month anomalies was already problematic for threshold setting, given it was a proxy for actually different types of factors that can influence food insecurity, such as droughts and flooding. These difficulties in setting thresholds on rainfall anomalies were also confirmed in our conversation with the analysis leads in Section 7.5. We should look into how we can best transform variables such as rainfall to represent more specific concepts, so any thresholds set on them can be more targeted on one phenomenon.

7.6.2 Conclusion

Our experiment has shown as a proof of concept that it is feasible for analysts to quantify their projection assumptions using thresholds. Implementing some form of threshold-based projection assumption quantification into the analysis process could not only provide extra insights during the analysis process but also help trigger earlier action during monitoring when food security crises arise. Moreover, such thresholds also allow for any future AI implementation for risk factor monitoring to be informed and contained by these thresholds, in this way setting a first step towards our design requirement for Meaningful Human Control. This experiment is a first small step towards realising a full-circle monitoring system for the IPC that considers the process of gathering input for monitoring during the analysis process as much as the actual monitoring and modelling process as well.

Chapter 8

Lessons for HCAI in humanitarian decision-making

In this chapter, we reflect on our case study of working towards a new monitoring process for the IPC to answer our third research question: "RQ3: Based on our case study, what lessons can we learn about the development of HCAI systems for humanitarian decision-making?". We will start with a short overview of the process we went through during our case study and afterwards elaborate on the lessons we have learned during this process. As this is the first case study that specifically tries to contribute to an HCAI methodology from the bottom-up (as far as we are aware), we do not have the full saturation of insights yet to provide here a whole framework. For that, more case studies should be performed to find out what elements seem to generally hold, and what is specific to different contexts. That is why we present our insights here clustered into 7 main lessons, in the hope that they will help future researchers when they shape a more coherent practice-informed methodology.

8.1 Our process

As described in Chapter 3, this research process started out as a plan to make solely an HCAI monitoring tool that would be developed following the traditional Clarify-Ideate-Develop-Implement cycle of Human-Centered Design. Instead, we found during interviews with IPC experts that the case was much more complicated, and the success of any solution would be highly dependent on its organizational and human aspects rather than solely the technical implementation. Therefore, based on these outcomes and the limitations we found in the current state-of-the-art of machine learning for Food Insecurity, we stepped away from the idea of an AI-first solution in Chapter 6, and instead developed a high-level plan for a new IPC monitoring process that leaned heavily on the assumption-building capabilities of IPC analysts. To verify whether this was possible, we tested in Chapter 7 whether they were able to quantify their projection assumptions using a threshold tool we had developed for this purpose.

In this research process, we developed different insights over time about what is needed in our view to make HCAI for humanitarian decision-making that is trustworthy, explainable, comprehensible, useful, and usable. These insights arose from different sources:

- The many discussions we have had with the IPC about their design requirements for HCAI and about their requirements for the development and use of FI Machine Learning models.
- The strong accountability and quality assurance systems the IPC uses to assess the quality of their data and their guidelines on how to interpret and use their different data sources
- Our concrete experiences during this research and development process.

8.2 7 Lessons for designing HCAI for humanitarian decision-making

1 Designing HCAI to adhere to an institution's values is a continuous balancing act

Currently, there is a lot of focus in the literature on defining some overarching set of principles that HCAI should adhere to. However, we found in Chapter 6 that these principles and their underlying values and norms are not sufficient if we want to develop HCAI that is truly centered around an institution and the people it serves. Through their purpose and context, they bring extra norms and values that need to be discovered and taken into account. An example of this is how the purpose of preserving lives in a humanitarian context brings added requirements for any HCAI tool that would be used there.

Moreover, when we mapped these HCAI principles to their underlying values and norms in Chapter 2, we found that these values and norms are neither absolute nor disjoint. Instead, they all exist on the same spectrum, where moving more towards one value means moving away from another. In other words: trade-offs need to be made between them. Furthermore, the trade-offs in those values can't also be determined beforehand as overlapping guiding principles such as: "40% safety and 60% efficiency".

Therefore, these trade-offs need to be reevaluated in every design decision in the process where those values come into expression. This requires continuous discourse with the stakeholders/domain experts throughout the design and implementation process to properly strike the balance each time such a decision pops up.

It also requires the designers and developers to look at these design choices with an active lens to identify those value trade-offs, so these trade-offs can be made consciously and responsibly. This requirement can be seen as part of the responsible innovation dimension of Anticipation, as defined by Stilgoe [45]. This process requirement teads to the next lesson:

2 Accountability comes from a valid process, not an external inspector and SHAP values

Explainable AI seems to be seen by many (such as in [42], [58], [32]) as the solution to fulfill the explainability requirement for accountability within HCAI. However, current attempts that are named in XAI literature are focused only on *AI* that can explain itself, not the *makers* that explain themselves. In our opinion, for true accountability and explainability, we also need to start asking the developers of AI to justify their decisions and ask questions such as:

- Why did you use this specific set of variables?
- Why did you choose these performance metrics?
- Why did you use these specific data sources?
- Why did you leave others out?
- What were the mechanisms that you were trying to capture in your model?

We see therefore a promising direction for research into how we can develop guidelines for translating different technical decisions to their ethical trade-offs and documenting them. Next to improving accountability procedures, this can also support users in gaining a better understanding of in which situations the algorithm can be trusted and when not.

The idea behind this stems from the procedures the IPC uses to assess the reliability of evidence in their analyses. For many outcome indicators, IPC analysts know for example how they are collected and on which questions they're based, which means they also know what implications those indicators have and in which situations they're relevant. If we want them to obtain the same kind of understanding of algorithms, information such as described above is then needed as well.

3 Develop for the user *and* the context

We found that in order to develop AI for a complex, high-stakes decision-making process, you cannot create a useful decision-support tool before you understand the current decision-making process that it will serve. Specifically, you need to know:

- Who makes the decisions?
- On what information and arguments do they base these decisions?
- Who collects and processes that information?

This information is needed for several reasons. First, it is necessary for truly understanding the problem that you are trying to solve: you need to know how the current decision-making process goes so you know which parts of the situation need to change and what bottlenecks need to be targeted to achieve that. Second of all, knowing on which arguments and information decision-makers currently base their decisions is imperative for understanding what elements should be explained about the AI in order for them to properly assess and use it. Lastly, knowing how the information is currently collected and prepared for decision-making can help with judging the suitability of different data sources for the AI, as well as finding possible sources for expert input if needed.

4 Don't build a spaceship if a bike might be enough: Start simple and only go complex afterwards if that is needed

In the literature we saw that many approaches to modeling Food Insecurity for monitoring purposes already started with quite complex models with many variables. Moreover, they were not developed with any specific decision-making process in mind. As an effect, these models were too complex to be fully explainable and did not answer the information needs of the IPC. While these authors spend a lot of time on making these complex models work, it would have been of much more benefit if someone had developed a monitoring dashboard that would just give decision-makers the relevant data they needed to assess the situation. Then, when they are enabled to best make the assessments themselves where possible, you can start implementing mathematical models to help in cases where there is uncertainty. This conclusion also follows the point made in "Stop explaining black-box machine learning models for high stakes decisions and use interpretable models instead" [37] to not make overly complicated (and uninterpretable) models when this isn't needed, especially when it comes to high-stakes decisions.

To quote an employee from the IPC: "You know what would already be nice? Just already a monitoring system that actually shows you the data. No extra analytics necessary first.".

5 But if you decide to go complex, use the knowledge that exists

As indicated before, once the simpler insights have been made available, there may come a point where it would be relevant and useful to add a more complex model to the process. However, for the sake of interpretability, efficiency, and inclusion, we would urge anyone in this case to use the knowledge that already exists in the institutional context. This knowledge comes in two forms:

Theoretical knowledge
There is a large body of literature on the mechanisms that drive food insecurity and how different factors contribute to this phenomenon. However, we found that much of this knowledge was not used in previous work on modeling food security. Instead, the general approach seemed to be to collect all the variables that might influence food security, feed them into a gradient-boosting machine, and analyze the results. In the same way that we incorporate laws of nature into physics models and economic theories into economic models, we think it would make much more sense to use our current knowledge of food insecurity to drive how we transform our variables and what mechanisms we model. This would as a natural effect also cause any

model to be more explainable by nature, as we would have determined its internal mechanics ourselves, instead of hoping that some mathematical optimization has drawn the right conclusion from the data we fed it.

Expert knowledge

Modeling food security is a specifically difficult problem, as it is a complex phenomenon with many input variables and differing contexts, but not much ground truth data. However, there is a whole range of experts available who have valuable knowledge and insights into those specific contexts and can help steer and tweak a potential model to be relevant and attuned to a specific context. More research is needed on how to best incorporate that input into models and use it to our advantage.

6 The HCD framework is a good starting point for implementing HCAI

But later phases need to be adapted to be able to work for AI development as well.

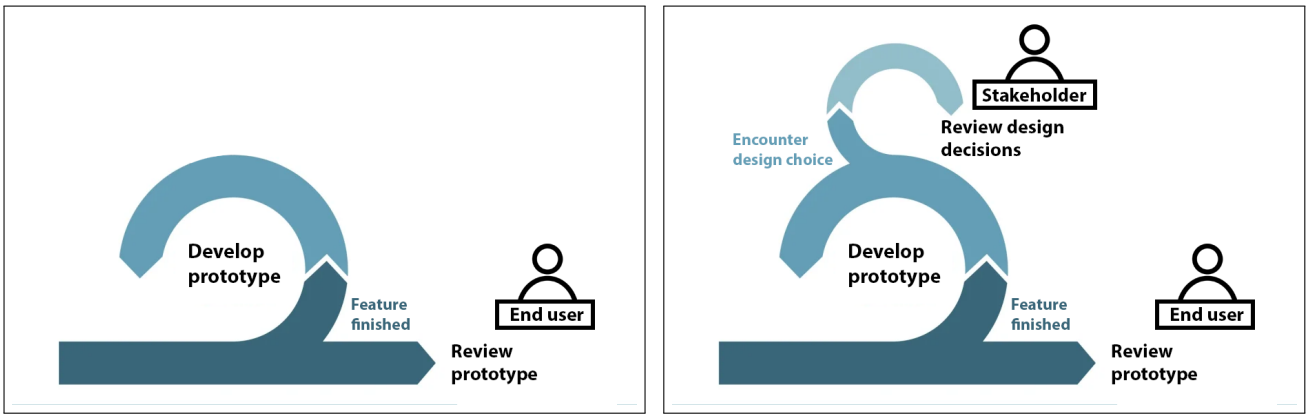
The first phase, the clarification phase, is a logical step in any design process and also turned out to be a logical one for developing HCAI as well. The interviews and literature exploration helped us to develop a good insight into what the problem was we needed to solve and what might be the possible steps to get there. In this phase, we also discovered that implementing AI was not the best solution at this point in time and that discovering how we could support expert input and determine the boundaries for human control was a first priority.

An important takeaway for doing HCAI using HCD techniques therefore is that after the clarification phase, there needs to be a reevaluation moment to check whether the original idea and problem definition still holds or whether another direction is actually needed. It is important to create space in the project to allow that the conclusion might be: "We can't implement AI here before we..", or even: "We can't implement AI here at all.". While in our research, this conclusion came more or less naturally, in larger projects, this reevaluation moment can be embedded more explicitly by using stage-gating in the same way as it was used in Stilgoe's 2013 paper [45] on Responsible Innovation.

In the next phase, the ideation phase, we did encounter that there is currently no clearly defined approach when it comes to defining specific user requirements for AI. Recently pre-published work by Xu et al. [59] might have helped to steer this process better. However, this gap did leave room for defining a set of user requirements that was tailored specifically to the needs and values of the IPC. We are therefore not sure yet how much can be standardized when it comes to finding requirements for AI or how much is dependent on the experience and analytical skills of the developer.

As we realized somewhere halfway through the ideation phase that we could not directly develop AI yet, we can't say from practice how well the next two phases of HCD work for implementing HCAI. However, given Lessons 1 and 5 of this chapter, where we concluded that continuous discourse with stakeholders and experts is needed throughout the development process to properly make ethical-technical design choices, we would like to propose one addition already to the HCD development phase as a way to meet this requirement.

In the traditional HCD development cycle, (rapid) prototyping and getting feedback on these prototypes follow each other in a repeated cycle, as illustrated in Figure 8.1a. However, as we pointed out before, it is not just the prototype that is important in HCAI, but also how it is developed. As the ethical and functional requirements are so closely interlinked with the technical design decisions made during the development process, we propose to add a much shorter iteration where, as technical choices come up that have ethical or functional impacts, developers discuss with a stakeholder in much smaller iterations about where they each time want to strike the balance, resulting in a process flow as illustrated in Figure 8.1b.



(a) Traditional design cycle

(b) Proposed design cycle with smaller iterations

Figure 8.1: New flow for efficient discussion of ethical-technical design decisions

If you are a software developer, you may recognize this shape from the famous Scrum method [41]. In the normal Scrum method, a product is iterated on every 2-4 weeks, with the improved prototype shown to the client at the end of each cycle. However, during these sprints, the prototype and development process is overseen by the product owner, who represents the interests and wishes of the client. In the case of HCAI, I propose to create a similar type of product owner who can quickly discuss with the developers the ethical trade-offs they want to make whenever a technical design choice arises, and who can then document these choices for the sake of accountability and to facilitate a wider discussion about them afterwards.

7 Doing Human-Centered AI properly takes a lot of time, *but* also generates a lot of valuable outputs

It is not to be underestimated how much more time it takes to do such an HCAI development process in a proper way. Scrutinizing and properly processing your data, running user tests, conferring with stakeholders; all those steps add up to the total development time of your system, given their time-consuming nature and their dependency on the schedules of other humans. Given the key norm within the humanitarian sector of economy of resources, it is important to weigh for each situation how extensive the HCAI approach should be.

However, when you do decide to do this process properly, it also means that you will have created something truly useful at the end of it. This can either be a successful artifact or new knowledge that some route or proposed solution will not work, which is a valuable insight in itself. Moreover, regardless of the outcomes, it can be exceedingly useful for the organization itself, as it will not only gain a great deal of insight into its own work processes and challenges, but also identify various opportunities and limitations with respect to new technologies and innovations.

Chapter 9

Conclusion

In this thesis, we present our interdisciplinary research on Human-Centered AI and Food Insecurity monitoring for the Integrated Food Security Phase Classification (IPC). Our aim was to develop Human-Centered AI (HCAI) to help improve the monitoring of Acute Food Insecurity (AFI), and in turn, learn more about what is needed to do HCAI in practice. We have endeavored to do this by answering three main questions and their respective sub-questions. In the next sections, we will present our answers and conclusions for each main question before ending with our recommendations for future research.

9.1 About developing a monitoring process

RQ1: How can we design an HCAI-supported monitoring process for Acute Food Insecurity (AFI) analyses that explicitly takes the validity of ongoing assumptions into account?

To answer RQ1, we have answered the following four subquestions:

1. ***How does the IPC analyze Food Insecurity and what computational models are currently available for predicting it?***
   By reviewing their extensive manual and learning materials, we gained an in-depth understanding of the IPC's analytical framework for assessing food insecurity, which is based on understanding the drivers of food security and correctly interpreting food security outcomes. We have analyzed the current state-of-the-art of Food Insecurity modelling and did not find any models that were suitable for our purpose, due to a lack of interpretability, data availability or use of IPC-approved training data.

2. ***What does the process in the IPC analysis cycle currently look like for developing and monitoring assumptions?***
   We obtained a detailed overview of the IPC analysis cycle by interviewing three experienced IPC facilitators, as well as by running many feedback sessions with our IPC contact person. We found that IPC projections are based on different kinds of assumptions, such as assumptions about how food insecurity drivers will evolve, how that will impact food security and within which periods projections are expected to stay the same. In addition, we found that some of these assumptions are formulated very qualitatively, and some are not even registered as assumptions by the IPC.

3. ***How can we change this process to make it produce assumptions that are more formally measurable and verifiable?***
   We designed a prototype which enables analysts to set thresholds on where they estimate food insecurity drivers to be during a given projection period, and for which values they are certain their projection assumptions will be invalid. We tested this prototype with 18 IPC experts from Malawi and found that they generally thought it was feasible to incorporate such a concept into the IPC analysis. More importantly, the majority of the participants saw considerable added value in setting these thresholds. Special care needs to be taken however in analyst training and indicator selection before integrating such a prototype into the IPC process.

4. ***How can we design a monitoring process that, using HCAI, assesses the validity of AFI analyses based on these assumptions?***
   Based on our interviews and problem analysis of the IPC, we determined two systems that have to be developed in order to make this monitoring process work: A system to define quantified analysis assumptions and a system to monitor them. For the first system, we need to determine which mathematical representations we want to use for different food insecurity drivers to make them measurable, as well as a mechanism for analysts to quantify their assumptions about these drivers. For the second system, there needs to be an alarm system to indicate when assumptions are becoming invalid, and there could be a place for an HCAI system to help decision-makers make assessments about Food Insecurity when it is not immediately clear whether the analysis should be updated or not.

   We did not reach the point in our research to build such an HCAI system, given the many steps that need to be taken before such a product becomes relevant in our case. Instead, we focused on the prototype as described in the question before. However, given the success of this prototype and the positive attitude of the IPC towards such a solution, we do see a potential success in our proposed solution for a new monitoring process.

9.2 About aligning the design process to HCAI

RQ2: How can we create a design process for answering RQ1 that fulfills the requirements for HCAI?
For RQ2, there were two subquestions to be answered

1. ***What are the requirements for HCAI in terms of the product and its development process according to literature?***
   For the ethical and technological requirements for HCAI, we inventoried requirements that were generally found in literature about trustworthy and responsible AI. We have related these requirements with each other through their intended purposes and values to obtain a more cohesive picture of the considerations and trade-offs that need to be made when developing HCAI. For the requirements regarding usefulness and usability, we found that the framework for Human-Centered Design (HCD) helps create useful and usable artifacts through the extensive interaction with the end-users in each stadium of the development.

2. ***How can we translate these requirements into our design process?***
   We integrated the process requirements for HCAI by adopting the HCD methodology as the main framework for our research. As the technological and ethical requirements we had found for HCAI were quite high-level and needed contextualization, we were not able to integrate those top-down into our design process or requirements. Instead, we developed a bottom-up set of design requirements based on our interviews and compared them afterward with the norms we had found in our literature review to test whether there were gaps.

9.3 About what we can learn from this

RQ3: Based on our case study, what lessons can we learn about the development of HCAI systems for humanitarian decision-making?

This case study provided many important takeaways for the future development of HCAI as a field and for practitioners aiming to work with it. These were takeaways about good practices for HCAI and important points to consider if we want to achieve certain norms that follow from it. Most of these takeaways can be summarized in these three recommendations:

1. Work closely with your stakeholders during every step of your process: understand their context, use their input and knowledge and develop what is actually needed,

2. Look for the ethical/functional trade-offs in your (technical) designs and document them,

3. Be prepared to make the time investment that the two recommendations above require.

9.4 Future research directions

There is much work to be done. Given the goal we found during our case study for Food Insecurity monitoring for the IPC ("Improving the IPC Acute Food Insecurity (AFI) monitoring process so decisions on analysis updates can be made (a) based on systematic data-driven argumentations and (b) in time."), it is safe to say that our research is not completed yet. Moreover, in our simultaneous process of figuring out how to do HCAI in practice, we identified several other areas for research as well. In the following two sections, we dive deeper into the future research directions for both research fields.

9.4.1 On modeling and monitoring Food Insecurity

As concluded above, to make our envisioned monitoring process work, we need two systems: 1) a system to quantify analysis assumptions and set thresholds on them and 2) a system that monitors them and assists in decision-making on analysis updates.

For the first system, there are still many open questions. Most of those are as much of a procedural nature as of a technical nature. We need to find, for example, better indicators to model the specific effects of rainfall, such as drought or flooding, but the challenge here is not as much how to develop them, as how to choose them. While there are already many models and indicators proposed for drought, such as SMART [21], SPI or RDI [3], there is not a procedure yet within the IPC on how to interpret these indicators and decide which ones are applicable for which situations, let alone on deciding what types of thresholds should be set for them. Moreover, there are subsequently also no methods yet for explaining these indicators to the analysts in a way that is intuitive to them and for creating a shared understanding on *how* to set thresholds on them.

For the second system, there is an even more open field to explore when it comes to assisting decision-makers with analysis updates: How can we best show the alerts generated by our thresholds? How can HCAI be used in cases where it is uncertain whether the Food Insecurity situation has changed? And how can make this HCAI such that it is interpretable and takes the outcomes and assumptions of the last analysis into account? How can we transform the IPC data pipeline to enable this transfer of information from the analysis into the algorithm? How do we deal with the limited amount of ground-truth data? And how do we present the AI outcomes in such a way that it fits the technical capacities of decision-makers and that it promotes their own reasoning capabilities instead of undermining them? And do all these measures actually improve decision-making? This is a non-exhaustive list of questions. We are open to suggestions for other issues to explore that we have not identified yet.

9.4.2 On Human-Centered AI

To structure our ideas for research directions into Human-Centered AI, we first want to take the reader along with our vision of what we would like the HCAI community to achieve in the coming years. Ideally, we would see that in some years, similar field guides for specific HCAI forms will be created like the one IDEO has made for Human-Centered Design. In this field guide, HCD is structured in three phases, and for each phase, numerous design activities and best practices are explained, which can be mixed and matched at the designers' discretion. Moreover, this field guide contains multiple concrete examples of how actual projects used HCD.

If we want to create such a field guide, different things are needed. First of all, more case studies need to be done. As we indicated in Chapter 8, we cannot give a complete methodology yet, as there are more real-life approaches needed in other contexts to find out what elements from each case study seem to be more universal for (specific forms of) HCAI and what elements is particular to each case. This will also allow us to create a framework that will fit for all the different cases within HCAI, as well as generate concrete

examples to explain how certain HCAI techniques can be applied.

Secondly, we hypothesize that due to the wide range of AI tasks (for example, decision-making vs text generation) and the wide range of application contexts, there need to be different field guides tailored to those different instances in order for them to be relevant. Therefore, more in-depth research is needed to create a taxonomy of those different types of AI and applications that can be used to differentiate which type of HCAI problems need a different approach.

Lastly, we need the methods to populate this field guide. Some of them already exist and just need to be adapted, such as techniques for interviewing stakeholders to understand the problem and find requirements. Other techniques still need to be developed (as far as we know), such as techniques for:

- How to explain different machine learning models to domain experts such that it is intuitive for them and developers can discuss with them on ethical-technical design choices
- How to give future AI users the right data literacy to properly use the models in different cases

And many more methods that we look forward to discovering in the coming years.

Appendix A

Interview guide

△ = question with high priority
□ = question of lower priority, useful for context

Introduce research objectives (5 min)

Thank you for participating in this interview. As this will be my first interview, apologies in advance for any fumbling on my part. I will first shortly explain our research purpose and the plan for the interview today, and then we can get right into it.
As communicated earlier via email, I would like to talk today about Food Insecurity monitoring by the IPC and the analysis process that precedes it. I would namely like to research if we can improve Food insecurity Monitoring by developing a tool that monitors assumptions underlying the IPC analyses. But in order to monitor analysis assumptions, I first need to understand how these are developed in practice, and in which forms they appear. In this way I can develop a monitoring tool that could actually fit into the current organization and that builds on the expertise that is present in the IPC process already.
My research questions for these interviews are therefore:

1. How is the process of assumption-building for IPC analyses and projections structured?
2. How are assumptions for current analyses and projections developed throughout this process?
3. How are relevant risk factors defined and monitored to potentially inform analysis updates?

So, how will the interview look like today?
We will start the interview with a short inquiry into your background with the IPC. Then I will walk with you through an analysis cycle. Next up, we will cover the current monitoring practices and lastly, we will close the interview. The whole interview should approximately take 1 hour.

If it is okay with you, I will record this interview, so I can transcribe it afterwards in peace and make sure that I have taken away the right messages. This transcription won't be shared with others. Any outcomes of this interview will also be fully anonymized. During the interview, you can also always make an "off the record" statement that I won't include in my research.

Do you have any questions about my research, the interview or the informed consent document?

Okay, so is it okay that I turn on the recording now?

Part 1: General info (5 min)

Ask the facilitator about their role within the analysis process.

1. △ What is your function and for which organization do you work?
2. □ How long have you been involved in IPC analyses?
3. △ How many IPC analyses have you been part of? and in which capacities?

Part 2: IPC analysis cycle (25 min)

So, in the coming part, I would like to walk with you through the last "business-as-usual" analysis cycle that you did. We will do this by first creating a schematic overview of the procedure, and then I will ask some specific questions about how certain analysis outcomes were developed.

First however, I would like to write down the context of the analysis. This context is mainly for me to place the analysis within the full picture and be anonymized in any publications so it can't be traced back.

A.0.1 Context

1. △ For which country was the analysis performed?
2. △ When was the analysis performed?
3. △ On what basis was it decided to conduct the analysis? Periodic or incidental reasons?
4. □ How big was the group of analysts and facilitators?

A.0.2 The process

Okay, let's talk about the analysis itself. As I have learned during the preparation of these interviews with Thomas, the procedure of how the analysis is performed can differ a bit per country. To get an overview of these differences and to get on the same page, I would first like to create a diagram with you to describe the procedure in this case.

I'm mostly interested here to find out **when** and **by whom** the current analyses and projections were determined. Specifically, I would like to find out how four important factors were developed throughout this process: contributing factors, the key drivers, the projection assumptions and the risk factors.

So, I will show you in a minute an example diagram of how such an analysis procedure might look like. The idea is that we afterwards create our new version based on your analysis in country X.

[SHARE SCREEN WITH ILLUSTRATOR]

1. Explain diagram & legend: This diagram shows which person or group does what over time. You can read the diagram from top to bottom.
2. Finish with: I would now like to know how such a diagram would look like for your process: When and by whom were the items such as the Projection assumptions discussed or developed?
3. So now, what should I copy or change from the example diagram? You can tell me and I will illustrate.

[STOP SHARING SCREEN]

Okay, so now we have an overview of the process, let's dive deeper into it.

A.0.3 Planning stage

Deciding on validity periods and information needs

1. △ How did you define the validity periods? (Which reasons factored into this decision?)
2. △ How did the analysis team start to determine what information they needed for the analysis process? When did this happen?
3. △ How did you decide which contributing factors were needed?
   (a) Which data sources/information factored into this?
   (b) Can you give some examples?
4. □ On what insights did you base your information needs?
5. □ How did these information needs guide the data collection and gathering?
6. □ How did you decide whether your information and data needs were sufficiently met? Did you set criteria beforehand for this?
7. □ Did you find out during the rest of the process that other data sources were needed? Why were they selected?

A.0.4 Preparation and Analysis stage:

To be asked about X:

1. What happened?
2. △ Which level? (example?)
3. △ Which format? How did you present or document it? (example?)
4. △ Which data sources? (example?)
5. △ Did these X differ much from the previous years?
6. Which reasoning?

*To be asked if was developed by **Analysts**:*

1. What instructions and information did they receive about X?
   (a) Which level?
   (b) Which format?
2. If given X: To what extent did the analysts adopt the X given to them?
   (a) In what situations did they make changes? Can you name some concrete examples?

Specific questions

*To be asked about **Plenary:***

1. □ What elements did the **facilitators/analysts communicate?**
2. □ What **analysis outcomes** were discussed? Were analysts required to make any **changes**/modifications to their outcomes?

*To be asked about **Current analysis:***

1. □ How did the contributing factors and outcome indicators inform the **population classification?**

*To be asked about **Projection assumptions:***

1. △ Do the PAs differ much from the CFs?

*To be asked about **Projection analysis:***

1. △ How do projection assumptions inform population figures for the projection period? Did you see cases where similar projection assumptions (across areas) led to different (changes) in population figures?

*To be asked about **Risk factors:***

1. △ How did the RFs relate to the CFs and PAs?

A.0.5 Communication stage

1. □ How did you decide on the key messages, in particular the key drivers and risk factors? On what level did you define those?
2. □ To whom did you communicate the key messages as well as risk factors to monitor? What did you exactly communicate on this and how?

A.0.6 Post-analysis

1. △ Did you monitor the FI situation after the analysis? àHow did you monitor it?
   (a) △ By monitoring the defined risk factors? Or any other factors?
   (b) △ Which data sources did you use to monitor those factors?
   (c) △ Which limitations were there during the monitoring?
   à What kind of support could have made this better?
2. △ Did the projections turn out to be true? (If you could determine that at all)
3. △ Did this help to inform a re-analysis or projection update?

Part 3: Current monitoring practices (15 minutes) 00:35

1. △ Do you monitor in-between analyses normally whether your projections are still correct?
2. △ How do you normally do that?
   (a) How do you determine what factors to monitor?
   (b) How often do you check them?
   (c) What tools and data sources do you use?
   (d) How do you define when a factor should trigger an alert?
3. △ How do you decide when to redo the analysis?
   (a) Which types of reasons can you use to decide on an analysis update (or new analysis)?
      i. Unforeseen shock(s) with a foreseeably high impact on FI?
      ii. Risk factor thresholds that are crossed?
      iii. Political factors, e.g. request by government or donors?
      iv. Other combinations of developments that make it highly unlikely that the current analysis assumptions are still valid?
   (b) What information or data is needed to support that decision?
   (c) How do you get others (relevant stakeholders) to agree on the necessity of conducting an analysis update?

Part 4: Closing statements (10 minutes)

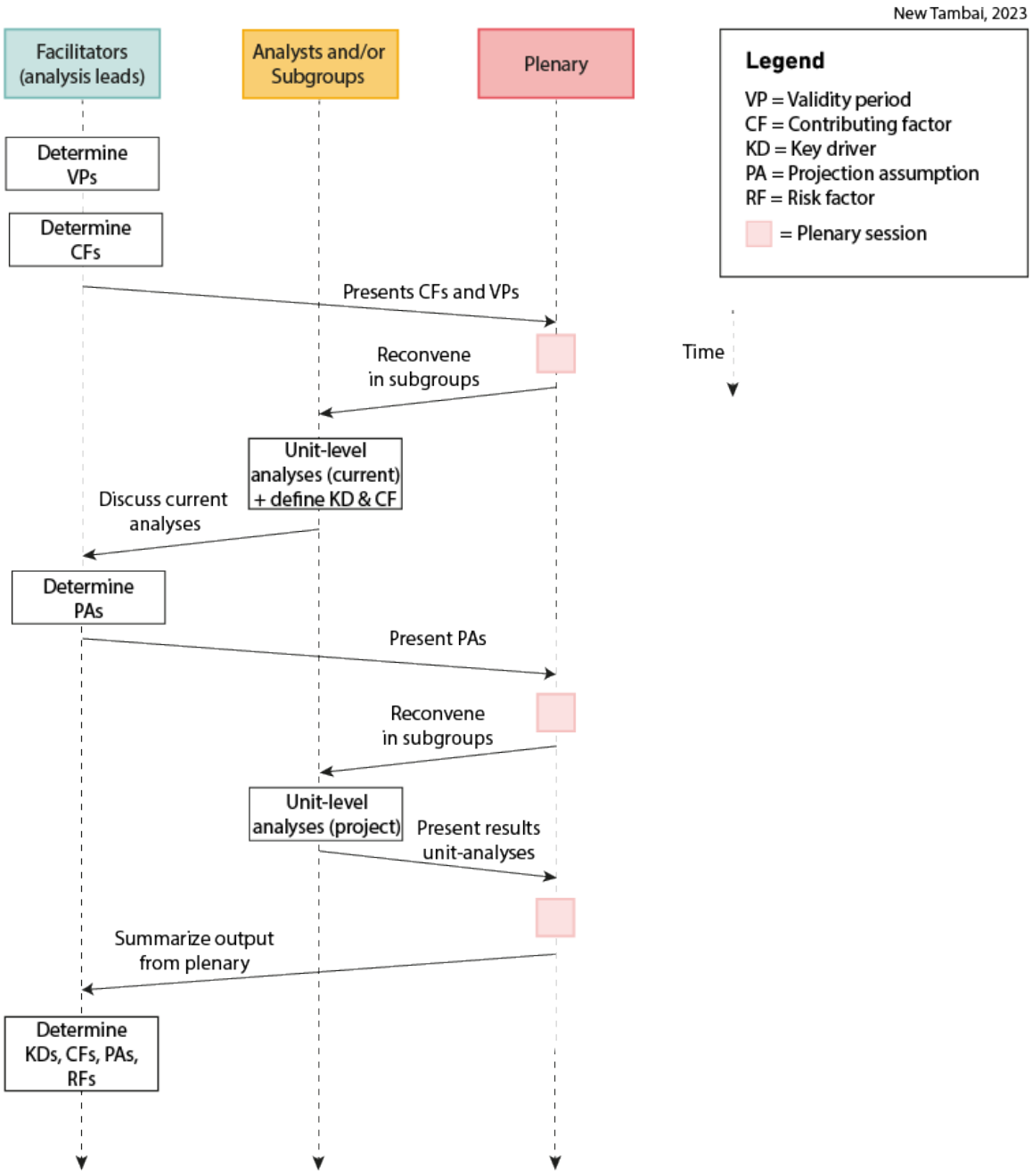


Figure A.1: Template process in the interview guide which needed to be adapted by the facilitators

Bibliography

[1] 2021. Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal. https://www.amnesty.org/en/documents/eur35/4686/2021/nl/

[2] Bo Pieter Johannes Andree, Andres Chamorro, Aart Kraay, Phoebe Spencer, and Dieter Wang. 2020. *Predicting Food Crises*. The World Bank. https://doi.org/10.1596/1813-9450-9412

[3] Mohammad Amin Asadi Zarch, Bellie Sivakumar, and Ashish Sharma. 2015. Droughts in a warming climate: A global assessment of Standardized precipitation index (SPI) and Reconnaissance drought index (RDI). *Journal of Hydrology* 526 (2015), 183–195. https://doi.org/10.1016/j.jhydrol.2014.09.071

[4] Jan Auernhammer. 2020. Human-centered AI: The role of Human-centered Design Research in the development of AI. In *Synergy - DRS International Conference 2020*, S. Boess, M. Cheung, and R Cain (Eds.). https://doi.org/10.21606/drs.2020.282

[5] Omar Cardona, Maarten Aalst, Joern Birkmann, Maureen Fordham, Glenn Mcgregor, R Perez, R Pulwarty, Lisa Schipper, and Sinh Bach. 2012. Determinants of risk: exposure and vulnerability.

[6] Hugo Deléglise, Roberto Interdonato, Agnès Bégué, Elodie Maître d'Hôtel, Maguelonne Teisseire, and Mathieu Roche. 2022. Food security prediction from heterogeneous data combining machine and deep learning methods. *Expert Systems with Applications* 190 (3 2022), 116189. https://doi.org/10.1016/j.eswa.2021.116189

[7] Virginia Dignum. 2019. *Responsible Artificial Intelligence*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-30371-6

[8] European Humanitarian Forum. 2023. Tackling the Global Humanitarian Funding Gap.

[9] European Union. 2010. *Charter of Fundamental Rights of the European Union*. Vol. 53. European Union, Brussels. 380 pages.

[10] FAO. 2022. Hunger and food insecurity. https://www.fao.org/hunger/en/

[11] FAO. 2023. Malawi national currency devalued. https://www.fao.org/giews/food-prices/food-policies/detail/en/c/1667165/

[12] Pietro Foini, Michele Tizzoni, Giulia Martini, Daniela Paolotti, and Elisa Omodei. 2023. On the forecastability of food insecurity. *Scientific Reports* 13, 1 (3 2023), 2793. https://doi.org/10.1038/s41598-023-29700-y

[13] GRFC 2023. 2023. *The Global Report on Food Crises 2023*. Technical Report. Rome.

[14] High-Level Expert Group on Artificial Intelligence set up by the European Commission. 2019. *Ethics guidelines for trustworthy AI*. Technical Report. European Commission, Brussels. https://ec.europa.eu/digital-

[15] IDEO.org. 2015. *Field Guide to Human-Centered Design*. https://www.designkit.org/resources/1.html

[16] IPC. [n. d.]. IPC Certification Programme - Process and Levels. https://www.ipcinfo.org/ipcinfo-website/e-learning/ipc-certification-programme/en

[17] IPC. 2023. IPC Overview and Classification System. https://www.ipcinfo.org/ipcinfo-website/ipc-overview-and-classification-system/en/

[18] IPC. 2023. IPC Self-learning courses: Acute Food Insecurity Classification. https://learning.ipcinfo.org/mod/scorm/view.php?id=184

[19] IPC Global Partners. 2021. *Integrated Food Security Phase Classification Technical Manual Version 3.1. Evidence and Standards for Better Food Security and Nutrition Decisions.* Technical Report. Rome. https://www.ipcinfo.org/fileadmin/user_upload/ipcinfo/manual/IPC_Technical_Manual_3_Final.pdf

[20] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2023. Trustworthy Artificial Intelligence: A Review. *Comput. Surveys* 55, 2 (2 2023), 1–38. https://doi.org/10.1145/3491209

[21] P. Krishna Krishnamurthy R, Joshua B. Fisher, Richard J. Choularton, and Peter M. Kareiva. 2022. Anticipating drought-related food security changes. *Nature Sustainability* 5, 11 (9 2022), 956–964. https://doi.org/10.1038/s41893-022-00962-0

[22] Jeff Larson, Julia Angwin, Lauren Kirchner, and Surya Mattu. 2016. How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[23] Stefan Larsson and Fredrik Heintz. 2020. Transparency in artificial intelligence. *Internet Policy Review* 9, 2 (2020).

[24] Lauren Landry. 2020. What is human-centered design? https://online.hbs.edu/blog/post/what-is-human-centered-design#:~:text=Human%2Dcentered%20design%20is%20a,tailored%20to%20your%20audience's%20needs.

[25] E. C. Lentz, H. Michelson, K. Baylis, and Y. Zhou. 2019. A data-driven approach improves food insecurity crisis prediction. *World Development* 122 (10 2019), 399–409. https://doi.org/10.1016/J.WORLDDEV.2019.06.008

[26] Fei-Fei Li and John Etchemendy. 2020. *2019–2020 Annual Report.* Technical Report. Stanford Institute for Human-Centered Artificial Intelligence.

[27] Giulia Martini, Alberto Bracci, Lorenzo Riches, Sejal Jaiswal, Matteo Corea, Jonathan Rivers, Arif Husain, and Elisa Omodei. 2022. Machine learning can guide food security efforts when primary data are not available. *Nature Food* 3, 9 (9 2022), 716–728. https://doi.org/10.1038/s43016-022-00587-8

[28] David Miller. 2021. Justice. In *The Stanford Encyclopedia of Philosophy* (fall 2021 ed.), Edward N Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[29] Marco Modica, Aura Reggiani, and Peter Nijkamp. 2018. Vulnerability, resilience and exposure: methodological aspects and an empirical applications to shocks. (2018). http://www.sustainability-seeds.org/.Enquiries:info@sustainability-seeds.org

[30] James H Moor. 1997. Towards a theory of privacy in the information age. *ACM Sigcas Computers and Society* 27, 3 (1997), 27–32.

[31] OIEWG. 2016. Report of the Open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction. https://www.preventionweb.net/files/50683_oiewgreportenglish.pdf

[32] Ozlem Ozmen Garibay, Brent Winslow, Salvatore Andolina, Margherita Antona, Anja Bodenschatz, Constantinos Coursaris, Gregory Falco, Stephen M. Fiore, Ivan Garibay, Keri Grieman, John C. Havens, Marina Jirotka, Hernisa Kacorri, Waldemar Karwowski, Joe Kider, Joseph Konstan, Sean Koon, Monica Lopez-Gonzalez, Iliana Maifeld-Carucci, Sean McGregor, Gavriel Salvendy, Ben Shneiderman, Constantine Stephanidis, Christina Strobel, Carolyn Ten Holter, and Wei Xu. 2023. Six Human-Centered Artificial Intelligence Grand Challenges. *International Journal of Human–Computer Interaction* 39, 3 (2 2023), 391–437. https://doi.org/10.1080/10447318.2022.2153320

[33] Steve Penson, Mathijs Lomme, Zacharey Carmichael, Alemu Manni, Sudeep Shrestha, and Bo Andree. 2024. A Data-Driven Approach for Early Detection of Food Insecurity in Yemen's Humanitarian Crisis. *World Bank Working Paper* 10768 (5 2024). https://doi.org/10.1596/1813-9450-10768

[34] Pew Research Center. [n. d.]. Writing Survey Questions. https://www.pewresearch.org/writing-survey-questions/

[35] Ibo Poel. 2013. Translating Values into Design Requirements. 253–266. https://doi.org/10.1007/978-94-007-7762-0{_}20

[36] Mark O. Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies* 1, 1 (1 2019), 33–36. https://doi.org/10.1002/hbe2.117

[37] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (5 2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[38] Fernando Rudy-Hiller. 2022. The Epistemic Condition for Moral Responsibility. In *The Stanford Encyclopedia of Philosophy* (winter 2022 ed.), Edward N Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[39] Stuart J Russell and Chai Faculty Director. 2020. *Center for Human-Compatible Artificial Intelligence Progress Report 2020*. Technical Report.

[40] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. 2020. Principles to Practices for Responsible AI: Closing the Gap. *CoRR* abs/2006.04707 (2020). https://arxiv.org/abs/2006.04707

[41] Scrum.org. [n. d.]. What is Scrum? https://www.scrum.org/resources/what-scrum-module

[42] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice. *ACM Transactions on Interactive Intelligent Systems* 10, 4 (12 2020), 1–31. https://doi.org/10.1145/3419764

[43] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice. *ACM Transactions on Interactive Intelligent Systems* 10, 4 (12 2020), 1–31. https://doi.org/10.1145/3419764

[44] Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human Computer Studies* 51, 5 (1999). https://doi.org/10.1006/ijhc.1999.0252

[45] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a framework for responsible innovation. *Research Policy* 42, 9 (11 2013), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

[46] UNDRR. 2021. Resilience - Understanding Disaster Risk. https://www.preventionweb.net/understanding-disaster-risk/key-concepts/resilience

[47] UNDRR. 2021. Vulnerability - Understanding Disaster Risk. https://www.preventionweb.net/understanding-disaster-risk/component-risk/vulnerability

[48] Bartel Van de Walle and Tina Comes. 2015. On the Nature of Information Management in Complex and Natural Disasters. *Procedia Engineering* 107 (2015), 403–411. https://doi.org/10.1016/j.proeng.2015.06.098

[49] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174014

[50] Dieter Wang, Bo Pieter Johannes Andrée, Andres Fernando Chamorro, and Phoebe Girouard Spencer. 2022. Transitions into and out of food insecurity: A probabilistic approach with panel data evidence from 15 countries. *World Development* 159 (11 2022), 106035. https://doi.org/10.1016/J.WORLDDEV.2022.106035

[51] WEF. 2023. *The Global Risks Report 2023*. Technical Report. World Economic Forum, Geneva.

[52] Joris J.L. Westerveld, Marc J.C. van den Homberg, Gabriela Guimarães Nobre, Dennis L.J. van den Berg, Aklilu D. Teklesadik, and Sjoerd M. Stuit. 2021. Forecasting transitions in the state of food security with machine learning using transferable features. *Science of The Total Environment* 786 (9 2021), 147366. https://doi.org/10.1016/j.scitotenv.2021.147366

[53] WFP USA. 2017. *Winning the peace: hunger and instability*. Technical Report. World Food Program USA, Washington, D.C,. https://www.wfpusa.org/wp-content/uploads/2019/03/2017-Winning-the-Peace-Hunger-and-Instability.pdf

[54] Wikipedia. 2024. Sequence diagram. https://en.wikipedia.org/wiki/Sequence_diagram

[55] Wikipedia. 2024. Values (Western philosophy). https://en.wikipedia.org/w/index.php?title=Values_(Western_philosophy)&action=history

[56] Wikipedia. 2024. Waterfall model . https://en.wikipedia.org/wiki/Waterfall_model

[57] Copyright Wisner and Duryog Nivaran. 2003. *At Risk: natural hazards, people's vulnerability and disasters Second edition 2003*. Technical Report.

[58] Wei Xu. 2019. Toward human-centered AI. *Interactions* 26, 4 (6 2019), 42–46. https://doi.org/10.1145/3328485

[59] Wei Xu, Zaifeng Gao, and Marvin Dainoff. 2023. An HCAI Methodological Framework: Putting It Into Action to Enable Human-Centered AI. https://arxiv.org/pdf/2311.16027

[60] Yujun Zhou, Erin Lentz, Hope Michelson, Chungmann Kim, and Kathy Baylis. 2022. Machine learning for food security: Principles for transparency and usability. *Applied Economic Perspectives and Policy* 44, 2 (2022), 893–910. https://doi.org/10.1002/aepp.13214