# Incentive-Tuning

Understanding and Designing Incentives for Empirical Human-AI Decision-Making Studies

Simran Kaur

# Incentive-Tuning

## Understanding and Designing Incentives for Empirical Human-AI Decision-Making Studies

by

**Simran Kaur**

for the purpose of obtaining the degree of
Master of Science in Computer Science
at the Delft University of Technology
to be defended publicly on
Monday July 8, 2024 at 14:00.



| | | |
|---|---|---|
| Student number: | 5765269 | |
| Project duration: | November 13, 2023 – July 8, 2024 | |
| Thesis committee: | Dr. Ujwal Gadiraju (Advisor & Chair) | TU Delft |
| | Dr. Myrthe L. Tielman | TU Delft |
| Daily co-supervisor: | Sara Salimzadeh | TU Delft |



An electronic copy of this thesis is available at
https://repository.tudelft.nl/.

TU Delft Delft University of Technology

*To moving halfway across the world and daring to do something different.*

# SUMMARY

With the rapid advance of artificial intelligence technologies, AI's potential to transform decision-making processes has garnered considerable interest. From criminal justice and healthcare to finance and management, AI systems are poised to revolutionize how humans make decisions across various fields. Their ability to analyze massive datasets and identify patterns offers significant advantages, including faster decision-making and improved accuracy. At the same time, human judgment and empathy are paramount for decision-making, especially in high-stakes scenarios. This has fueled explorations of *collaborative* decision-making between humans and AI systems, aiming to leverage the strengths of both human and machine intelligence.

Integrating AI effectively into decision-making processes requires a deep understanding of how humans interact with AI. To explore this dynamic, researchers conduct *empirical studies*. These studies delve into human-AI interaction, investigating how humans use AI assistance for decision-making and how this collaboration impacts results. These studies are thus crucial for shaping the future of human-AI decision-making. They not only illuminate the fundamental nature of this interaction but also guide the development of new AI techniques and responsible practices.

A critical and fascinating aspect of conducting these studies is the *role of participants*. The validity of such studies and the applicability of their findings hinges on the behaviours of the participants of the study, who act as the human decision-maker. Study participants might not necessarily embody the true motivations that drive the humans making decisions in the real-world. Effective *incentives* that motivate participants may lead to improved engagement and make participants more invested in the decision-making process. This can lead to richer data and more reliable results that accurately reflect real-world human-AI interaction.

Incentive schemes can thus be the bridge between the controlled environment of the study and the complexities of real-world decision-making. By carefully designing incentives that align with the study goals and participant motivations, researchers can unlock the true potential of empirical studies for investigating human-AI decision-making.

Thus, in this thesis, we highlight and address the critical role of incentive design for conducting empirical human-AI decision-making studies. We focus our exploration on *understanding*, *designing*, and *documenting* incentive schemes.

Through a thematic review of existing research, we lay bare the landscape of current practices, challenges, and opportunities associated with incentive design in human-AI decision-making, in order to to facilitate a more nuanced understanding. We identify recurring patterns, or *themes*, such as what comprises the components of an incentive scheme, how incentive schemes are manipulated by researchers, and the impact they can have on research outcomes. We further raise several questions to lead the way for future research endeavours.

Leveraging the acquired understanding, we present a practical tool to guide researchers in designing effective incentive schemes for their studies - the Incentive-Tuning Checklist. The checklist outlines how researchers should undertake the incentive design process step-by-step, and prompts them to critically reflect on the trade-offs and implications associated with the various design choices and decisions. To aid this effort, we supplement the checklist with detailed discussions and valuable suggestions.

Further, recognizing the importance of knowledge capture and dissemination, we provide tools to meticulously document the design of incentive schemes in the form of a reporting template and a collaborative repository.

By advocating for a standardized yet flexible approach to incentive design and contributing valuable insights along with practical tools, this thesis paves the way for more reliable and generalizable knowledge in the field of human-AI decision-making. Ultimately, we aim for our contributions to empower researchers in developing effective human-AI partnerships for decision-making.

# ACKNOWLEDGEMENTS

This thesis would not have been possible without the guidance and support of many individuals.

First and foremost, I would like to express my sincere gratitude to my thesis advisor, Dr. Ujwal Gadiraju. It is only your wisdom and guidance that steered this project in the right direction. Your invaluable expertise, patience, and encouragement throughout this journey were instrumental in shaping this thesis as well as me, as a person, over the past year. You believed in me more than I did myself, and for that I will forever be grateful.

My daily co-supervisor, Sara Salimzadeh, I am particularly grateful for our meetings. They were like lifelines – especially those times when I felt a little lost at sea. Your guidance on even the minute aspects of this thesis, coupled with your unwavering patience, helped me stay sane throughout this process. Thank you for critiquing, and at the same time, reassuring me - always with a smile.

I am also grateful to my friends, the *Dream Team*, and the one brain-cell we share that had to work overtime for all of us. You were the ultimate degree survival squad. Together we panicked, celebrated, and somehow managed to master computer science along the way.

To my partner, Vikram, thank you for being my one-man support army. You deserve a medal (and a plane ticket). If it wasn't for your constant love and reassurance, and simply sitting on calls together for hours as we worked, I never would have made it. You are my inspiration in everything I do. I love you.

Finally, to my parents, words can't express my love and gratitude. Your endless faith, love, and sacrifices have provided the foundation for all my achievements. You are my strength. I am unbelievably lucky and eternally grateful to be your daughter. Thank you for supporting me through everything I have ever wanted to do and always being there for me.

"*Last but not least, I wanna thank me*
*I wanna thank me for believing in me*
*I wanna thank me for doing all this hard work*
*I wanna thank me for having no days off*
*I wanna thank me for, for never quitting*"
*– Snoop Dogg*

# CONTENTS

# 1

# INTRODUCTION

"*AI Won't Replace Humans —*
*But Humans With AI Will Replace Humans Without AI.*"

Harvard Business Review

Artificial intelligence (AI) technologies have advanced highly in recent times, demonstrating remarkable predictive capabilities [1]. This has led to the integration of AI systems into several fields that involve *decision-making processes* such as criminal justice, healthcare, money lending, organizational management, and more [2–6]. An AI system's ability to quickly process large volumes of data and identify patterns can make decision-making faster and more reliable, offering numerous advantages, such as increased efficiency, scalability, and data-driven insights [7]. However, while AI can enhance decision-making processes, complete automation of such processes is not always desirable due to the importance of human judgment and empathy as well as ethical and legal considerations in certain contexts, specially high-stakes domains [8, 9].

Hence, there has been a growing interest in *augmenting* human decision-making with AI assistance, aiming to leverage the strengths of both humans and machines [7, 10, 11]. This concept is referred to as *human-AI decision-making*, though various other terms such as human-AI interaction, human-AI collaboration, and human-AI teaming are also used [11].

## 1.1. EMPIRICAL STUDIES IN HUMAN-AI DECISION-MAKING

To effectively integrate AI into human decision-making processes, it is important to develop a fundamental understanding of how humans interact with AI systems, how they incorporate AI advice into their decision-making, and how this collaboration impacts outcomes. To do so, researchers have been conducting *empirical studies* that investigate

the dynamics of human-AI interaction in the context of decision-making, exploring different factors such as algorithmic aversion, trust, reliance, fairness perceptions, explainability, cognitive biases, and more [12–22].

*Empirical studies play a crucial role in shaping the future of human-AI decision-making.* Empirical studies can shed light on the fundamental nature of human-AI interaction in this context. This understanding can serve a multitude of purposes. It can guide the development of new AI techniques that are not only more effective in assisting human decision-making, but also better align with human cognitive strengths and weaknesses. Further, these studies provide valuable insights for practitioners building AI assistance, allowing them to make informed technical and design choices that optimize the human-AI partnership. Finally, empirical data and insights can inform the creation of policies, infrastructure, and practices surrounding human-AI decision-making to establish regulatory frameworks that dictate the responsible and ethical use of AI in society [11].

*Crowdsourcing is a valuable tool for conducting such empirical studies.* Empirical human-AI decision-making studies often simulate real-world decision-making scenarios, with *participants* playing the role of human decision-makers and providing valuable insights into how humans interact with AI systems and the effectiveness of AI assistance [12, 23, 24]. Some studies also task participants with assessing the quality of AI-generated recommendations or explanations, or with providing feedback on AI systems and interfaces in order to identify areas for improvement in AI models [15, 25–27]. It is *crowdsourcing* that has enabled the research community to conduct large-scale experiments requiring human participation in various capacities [28]. It has hence proved to be a valuable tool for human-AI decision-making researchers, allowing them to conveniently recruit participants for their studies for engaging in decision-making tasks with AI systems.

## 1.2. INCENTIVES FOR CROWDSOURCING

The design of studies that recruit participants through crowdsourcing, or *crowdsourced studies*, involves several key aspects, such as crafting and allocating tasks to the crowd, providing incentives to participate, implementing data quality control measures, and aggregating individual contributions into usable data that can yield valuable insights [29–32].

*Among the several aspects of crowdsourcing, incentives emerge as a critical component.* Incentive schemes serve as the linchpin for motivating and retaining participants as well as ensuring active engagement and maintaining data quality [31, 33]. When participants feel fairly compensated and motivated by the incentives, they could be more likely to truly engage with the tasks, and emulate the real-world motivations of the human whose role they are playing [34, 35].

Hence, understanding and implementing effective incentive schemes is imperative for the success of crowdsourced studies.

The design of incentive schemes is thus in itself a field of interest, with researchers exploring how to appropriately reward crowdworkers while ensuring the reliability and validity of study outcomes within various domains [33, 36]. Common approaches include monetary rewards, such as payment per task or hourly rates, as well as non-monetary incentives like gamification elements, badges, or platform recognition for high-quality contributions [33, 37, 38]. Researchers have also focused on optimizing monetary incentive structures to strike a balance between motivating participation and controlling costs [39–41], while also considering ethical implications such as fair compensation and preventing exploitation [42, 43].

*Successful incentive scheme design hinges on various factors.* The effectiveness of incentive schemes lies in their ability to align with the objectives of the study while also appealing to the diverse motivations of participants. Moreover, the design of incentive schemes should consider the nature of the study and the tasks involved. Additionally, researchers must be mindful of ethical considerations, ensuring that incentives are fair and equitable for all participants. By carefully designing incentive schemes that address these factors and resonate with participants, researchers can maximize participation and data quality in crowdsourced studies.

## 1.3. RESEARCH SCOPE AND MOTIVATION

This research work lies at the intersection of the field of human-AI decision-making and the role of crowdsourcing in conducting empirical studies. Within this landscape, the scope of our exploration is focused on the *design of incentive schemes for crowdsourced human-AI decision-making empirical studies*. We capture this in Figure 1.1.

*Human-AI decision-making tasks go beyond simple crowdsourcing tasks.* While traditional crowdsourcing tasks often comprise of simple microtasks like data annotation or transcription [44, 45], decision-making tasks involve complex cognitive processes [46]. Participants of human-AI decision-making studies often contribute to high-stakes decision-making processes, analyzing information, making judgments, and assessing AI capabilities, ultimately impacting the very foundation of our understanding of human-AI interaction within decision-making [11].

Designing crowdsourcing experiments that reflect the real-world stakes of such tasks can be challenging. Translating these stakes to the context of crowdsourcing is not straightforward since participants are usually recruited from platforms where the primary incentive is monetary [47]. Thus, the monetary incentive - and not the perceived

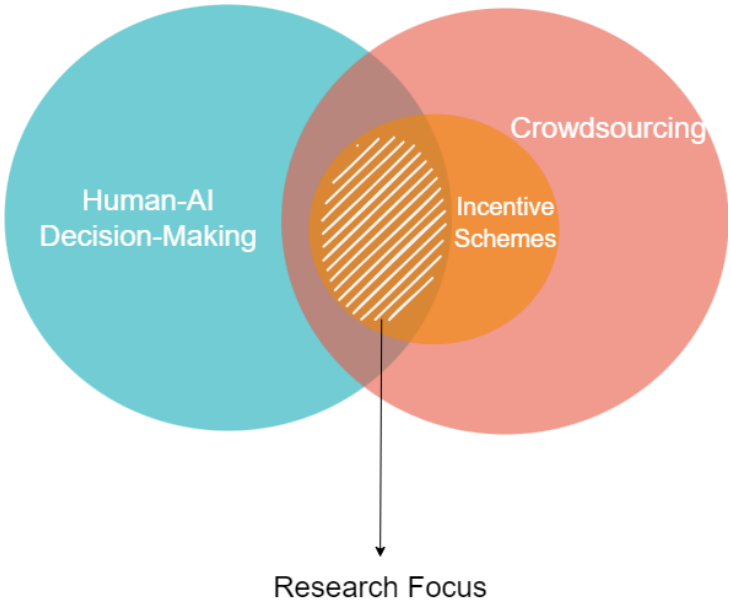


Figure 1.1.: Research focus of this thesis: incentive schemes for crowdsourced human-AI decision-making

stakes of the decision itself - can become the primary motivator for the decision-makers.

Despite these nuances, the current research landscape relies heavily on simplistic [19, 48–55] or ad-hoc [12, 18, 23, 56–60] incentive schemes. They seem to rarely be strategically tailored to align with the research goals or to address the unique challenges presented by the field of human-AI decision-making.

*Participant motivation can have implications on study findings.* How researchers motivate and incentivize participants can influence the inferences of the results of a study [61]. Researchers have in the past discussed the possibility of participant motivation affecting certain results, acknowledging it as the potential cause behind observations [23, 62]. While few studies enquire further and discuss the role or potential impact of incentives for their experiments [20, 62–67], there are also studies that do not describe incentive schemes at all [68–70].

These variations and inconsistencies in the design and descriptions of incentive schemes within human-AI decision-making literature highlight a critical gap: current practices *lack a standardized approach* to the design as well as the documentation of incentives, making it difficult to compare findings across different research projects. This hinders collaborative research efforts and the ability to generalize conclusions,

thus slowing down the development of a robust body of scientific knowledge in this field.

This lack of cohesion regarding the use of incentives in the literature has also caused the research community to suffer from a *fragmented understanding* of the role of incentives. This creates a significant knowledge gap, as we currently lack even an understanding of the nature of incentive schemes and the potential ramifications associated with them. This "unknown unknowns" situation makes it challenging to evaluate the implications of existing research findings and impedes the development of standardized practices for the future.

Thus, we identify that there is a pressing need for a more nuanced approach towards understanding and designing incentive schemes. Careful design of incentives, which addresses the complexities of the human-AI decision-making task while aligning the goals of the study with the motivations of the participants, is critical for conducting valid and robust human-AI decision-making studies. This makes our chosen research area worthy of in-depth scientific exploration.

## 1.4. RESEARCH QUESTIONS

Addressing the identified research gaps appears to be a three-pronged effort. Firstly, it is important to foster a deeper understanding of the current scenario of incentive design for human-AI decision-making studies. Secondly, we need to determine how to actually design incentive schemes, while addressing the unique challenges that the domain presents. Here, we also need to move towards standardization of the design process. Thirdly, we need to identify how to best document the incentive schemes, their design process and the associated outcomes, so that we can pave the way for future researchers to build upon existing knowledge.

Following from this reflection, we develop concrete research questions (RQs) that further guide our research effort. By articulating clear and concise research questions, we aim to effectively communicate the purpose of this work. The process of formulating the research questions involved delving into existing literature, bearing in mind the gaps identified in the previous section, and refining the scope of inquiry. Through this process, the following actionable research questions emerged, which lay the foundation for rigorous and impactful research outcomes:

***RQ1: How are incentive schemes currently designed for conducting empirical human-AI decision making studies?***

Human-AI decision-making empirical studies are focused on specific

domains or decision tasks, with various factors possibly influencing and shaping the structures of incentives. However, this hasn't been extensively investigated yet. In order to bridge this gap, we need to delve deeper into the existing literature and provide insights into the prevailing structures of incentive schemes as well as the discourse surrounding incentive design and its impact. Hence, RQ1 is formulated with the aim of investigating the current state of incentive design in human-AI decision-making literature.

> ***RQ2: How can incentive schemes be appropriately designed through a standardized process for empirical human-AI decision-making studies?***

RQ2 tackles the challenge of actually crafting incentive schemes for human-AI decision-making studies. There is an inherent complexity in designing incentives, which can be influenced by the study's goals, disciplinary practices, resource limitations, and several other potential factors. In exploring how to design, or *tune*, incentive schemes for human-AI decision-making studies, we seek a flexible yet standardized approach to foster comparability across studies.

> ***RQ3: How can the design of incentive schemes be documented through a standardized process to facilitate future research in human-AI decision-making?***

RQ3 acknowledges the importance of documenting the design process for incentive schemes for human-AI decision-making studies. While well-designed incentives are crucial for reliable research, the knowledge gained from creating them can be lost if not properly captured. As noted in the previous section, researchers describe incentive schemes in varying ways or they may even not describe them at all. This research question thus explores a standardized documentation process for incentive schemes.

These research questions serve as a roadmap, guiding the objectives, scope, and methodology for the rest of this thesis.

## 1.5. OVERVIEW OF METHODOLOGY

Firstly, to address RQ1, we conducted a *semi-structured review* and *qualitative analysis* of existing human-AI decision-making literature by means of a *thematic literature review*. A thematic literature review is a *thematic analysis* [71] of literature or parts of literature. We executed a meticulously defined search strategy guided by specific inclusion

criteria, to identify the literature to be reviewed. We further developed a dataset by extracting relevant fields from the shortlisted articles. This included *excerpts* describing incentive schemes or discussing participant motivation from within the articles. For qualitatively analysing the textual excerpt data, we specifically employed *reflexive* thematic analysis. We followed a rigorous methodology for conducting a *good* reflexive thematic analysis, as outlined by Braun and Clarke [72].

Secondly, in order to address RQ2, we developed a *checklist*, called the Incentive-Tuning Checklist, as a standardized solution to guide researchers in tuning incentive schemes for their human-AI decision-making studies. The construction of the checklist was informed by the *insights* obtained from the thematic analysis as well as *prior literature*. Specifically, our thematic analysis highlighted several core components that form an incentive scheme. These were refined and incorporated into the *items* of the checklist. Other items of the checklist were identified by reflecting on various other insights from the thematic analysis as well as lessons from prior literature. The discussion and suggestions accompanying the checklist were also informed by related literature.

Finally, to address RQ3, we leveraged the *Incentive-Tuning Checklist* to define a standardized reporting *template*. We further used GitHub [1] to set up a *repository* for hosting incentive scheme data that we gathered during the literature in order to facilitate open-access and collaboration. The source data format for documenting incentive schemes in the repository is directly based on the items of the checklist and the template.

## 1.6. SUMMARY OF CONTRIBUTIONS

This thesis makes the following novel contributions to the field of human-AI decision-making research:

1. *Insights into the current state of incentive design.* The objective of RQ1 was to shed light onto the practices currently prevailing in incentive scheme design for human-AI decision-making studies. The themes identified through a thematic literature review form the basis of our contribution towards addressing this objective and to the field of human-AI decision-making. Our critical discussion and reflection on the themes culminated in several insights and directions for future work. These are captured in Section 2.5.

2. *The Incentive-Tuning Checklist.* The checklist we provide for addressing RQ2 is a unique tool that can guide researchers in designing incentive schemes for their human-AI decision-making

[1] www.github.com

studies. Captured and discussed in-depth in Section 3.3, this checklist is a first-step contribution towards standardizing the approach to incentive design in this context.

3. *Documentation template and repository*. Lastly, in addressing RQ3, we contribute a standardized template that researchers can use to describe incentive schemes within their research articles. We also contribute a public-access, collaborative repository for conveniently storing and accessing incentive schemes and the rationales behind incentive design decisions for published research. These are described in Sections 4.2 and 4.3.

All the supplementary materials and resources associated with this work are captured in this public GitHub repository: https://github.com/simrankaur1509/IncentiveTuning.

## 1.7. REPORT OUTLINE

The rest of this thesis is structured as follows: Chapter 2 describes the methodology and outcomes of the thematic literature review addressing RQ1. Chapter 3 builds upon the insights presented in Chapter 2 to propose a checklist in order to address RQ2. It discusses the checklist and its implications in-depth and provides case studies to demonstrate its application. Chapter 4 provides a standardized reporting template and collaborative repository to address RQ3. Chapter 5 concludes this work by revisiting the research questions we set out to address in Chapter 1, through the lens of the findings. It highlights the implications and limitations of this thesis, finally paving way for future work.

# 2

# INVESTIGATING THE CURRENT STATE OF INCENTIVE DESIGN: A THEMATIC LITERATURE REVIEW

"*Research is formalized curiosity.*
*It is poking and prying with a purpose.*"

Zora Neale Hurston

*In this chapter, we use thematic analysis to unravel the current landscape of incentive schemes employed by researchers for conducting human-AI decision-making studies. By peering into the existing literature, we aim to answer our first research question: "How are incentive schemes currently designed for conducting empirical human-AI decision making studies?"*

*We begin by outlining the theoretical underpinnings of thematic literature review, subsequently detailing the search and screening process of finalizing the articles to be reviewed. Then we delve into the thematic analysis methodology employed for this study. Notably, our analysis embraces reflexivity.*

*As we unveil our identified themes, accompanied by detailed descriptions and illustrative examples, we shed light onto the discourse surrounding incentive scheme design for human-AI decision-making studies. Furthermore, we reflect on the observations we made and raise questions that warrant further exploration, laying down the path for future research endeavours.*

## 2.1. TOWARDS RQ1

In this section, we lay the groundwork for our investigation into incentive design for human-AI decision-making studies by exploring two key analytical tools: thematic analysis (TA) and thematic literature review. Then, we outline the motivation behind using these methods to address our primary research question:

> *RQ1: How are incentive schemes currently designed for conducting empirical human-AI decision making studies?*

### 2.1.1. THEMATIC ANALYSIS

*Thematic analysis* (TA) is a method specifically designed to analyze qualitative data [71]. It is used to identify and interpret meaningful patterns, or "*themes*", that emerge within a dataset. These themes can provide insights to address various kinds of research questions.

TA has long been popular in qualitative research, as it is considered a versatile tool that can be applied to many disciplines and fields of study [73]. Researchers have often used it to explore people's experiences, perceptions, and representations of concepts within various domains [74–78].

#### REFLEXIVE THEMATIC ANALYSIS

Reflexive thematic analysis is a specific approach to thematic analysis that differs from other approaches in terms of the underlying philosophy and procedures for theme development. Designed and refined over the years by Braun and Clarke [71, 72, 79, 80], it has been described as a theoretically flexible method that can be used within a range of frameworks to address different types of research questions. We will delve deeper into the implications of *reflexivity* in Section 2.3.

### 2.1.2. THEMATIC LITERATURE REVIEW

A *thematic literature review*, also previously referred to in the literature as a "thematic analysis of literature", "thematic review of literature", or "systematic review and thematic analysis", is essentially a thematic analysis applied to a collection of research articles. Just as a thematic analysis examines qualitative data to generate themes, a thematic literature review analyzes a body of literature surrounding a specific research topic for the same. It can thus be described as a thematic analysis, where the dataset is a corpus of research articles or smaller extracts from within research articles.

Thematic literature reviews aim to explore and identify patterns, themes, and concepts within a body of literature, and can be leveraged to gain an understanding of the main aspects and trends in a particular area of research [81]. They typically result in a narrative synthesis that discusses the identified themes and how they relate to the broader research question.Thematic reviews are also less rigid in their approach, as the TA process is more interpretative and involves qualitative analysis subject to the researcher's position [82].

Thematic analysis and thematic literature reviews are most commonly used in fields like healthcare and medicine [83–86] as well as economics and business [87–89]. While less frequent in computer science (CS) and human-computer interaction (HCI) research, there have been some recent publications which demonstrate a growing adoption of this approach [90, 91].

### 2.1.3. WHY TAKE A THEMATIC APPROACH?

We posit that a thematic approach is perfectly suited to addressing our primary research question, RQ1: "*How are incentive schemes currently designed for conducting empirical human-AI decision making studies*?"

Thematic analyses excel at identifying recurring themes and patterns within a body of text. Further, thematic literature reviews transcend the role of summarizing and analyzing existing research. They aim to foster a deeper understanding of the current state of knowledge within a specific field. This is achieved by meticulously examining research articles and identifying the underlying themes that connect seemingly disparate studies, through following the widely practiced methodologies of TA. By uncovering these thematic threads, it becomes possible to construct a richer tapestry of knowledge, revealing how researchers have approached a specific topic and how their ideas have evolved over time.

This aligns perfectly with our goal of understanding the current practices employed by researchers in designing incentive schemes for human-AI decision-making studies. Researchers often describe the incentive schemes they employed while conducting these studies within their research papers. While traditional thematic analysis often utilizes interview or survey data, it can be effectively applied to other forms of text as well. In this case, the descriptions of incentive schemes presented within research papers can function as the researchers' statement on how they designed and implemented incentives for their studies. By conducting a thematic analysis on these descriptions, we can uncover meaningful insights. Eventually, it can allow us to gain a deep understanding of the current landscape of incentive design in human-AI decision-making research, by revealing latent knowledge such as the different concepts involved in incentive design practices, the

approaches and trends favoured by researchers, and the connections and relationships between these various aspects.

Ultimately, we believe that a thematic literature review will allow us to interpret the narrative surrounding incentive design in human-AI decision-making studies. It will reveal the *current state of incentive design* in order to address RQ1 and further pave the way for identifying areas of importance and future research directions.

## 2.2. LITERATURE REVIEW METHODOLOGY

We conducted a *semi-structured* literature review by means of a thematic literature review for this study. This section outlines its scope and details the criteria used to select relevant papers. It further elaborates on how the search for relevant literature was carried out as well as the process of screening relevant papers for review.

### 2.2.1. SCOPE OF REVIEW & INCLUSION CRITERIA

We focused on *empirical*, *human-subject* studies that investigate human-AI collaboration in *decision-making tasks*. These are studies that aim to evaluate, understand, or improve human performance and experience within the decision-making context.

Further, since the research questions are centered around exploring incentive scheme design, we also aimed to only investigate studies that recruit crowdworkers to play the role of the human decision-makers. Since crowdworkers are mostly incentivized with monetary rewards, our focus for further work and analysis is also scoped to monetary incentives only.

After scoping the research focus as described above, the inclusion criteria for papers was defined as follows:

- The investigated tasks must be centered around decision-making activities in human-AI collaboration. Papers focusing on tasks with different goals, such as debugging, model improvement, co-creation, or gaming, are excluded.
- The study must involve human decision-makers specifically in the context of crowdwork. Papers that solely focus on in-house personnel to act as the human decision-makers, such as employees of an organization or students of a university, are excluded.
- We exclusively target studies with an *evaluative* focus. These studies assess the effectiveness, usability, or impact of human-AI collaboration within a decision-making task. This excludes purely formative studies that aim to explore user needs to inform the design of AI systems. Formative studies often rely on qualitative

methods like interviews. We explicitly exclude purely qualitative, interview or survey based studies. The crowdworkers must have a *task* to perform.

*Note: Following the specification of this criteria, from here on we use the terms "participant" and "crowdworker" interchangeably. We also refer to "empirical studies" directly as "studies".*



### 2.2.2. SEARCH STRATEGY

We identified that Lai *et al.* [11] have shared a list[1] of research papers that was compiled considering an inclusion criteria that is a superset of the inclusion criteria defined for this work. Thus, we included all 81 studies presented within it to be screened again against our specific criteria. This covered the potentially relevant literature until the year 2021.

To ensure comprehensiveness for more recent research, we conducted a search focused on publications between 2021 and November 2023. Our search targeted the proceedings of key venues in the field, including the ACM Conference on Intelligent User Interfaces (ACM IUI), the AAAI/ACM Conference on AI, Ethics, and Society (AIES), the ACM Conference on Human Factors in Computing Systems (CHI), the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), the ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW), and the AAAI Conference on Human Computation and Crowdsourcing (HCOMP). The search keywords were "human-AI collaboration" and "human-AI decision-making".

An initial screening using titles and abstracts identified 86 potentially relevant papers, resulting in a total of over 160 papers papers to be evaluated further. Following the initial screening, we meticulously re-evaluated each paper against the defined inclusion criteria. This rigorous assessment resulted in a final selection of 97 papers deemed in-scope for this study. Figure 2.1 illustrates the sequential steps of searching, screening, and including research papers for the thematic analysis.

### 2.2.3. DATASET DEVELOPMENT

We first compiled the bibliographic information of the selected papers into a spreadsheet. This captured details such as title, authors, publication year, venue, and an accessible link for each included paper.

Then, for each study, the following information was extracted for each paper: the objective and research goals of the study, the study (and task) domain, the perceived risk or stakes associated with the

[1] https://haidecisionmaking.github.io/

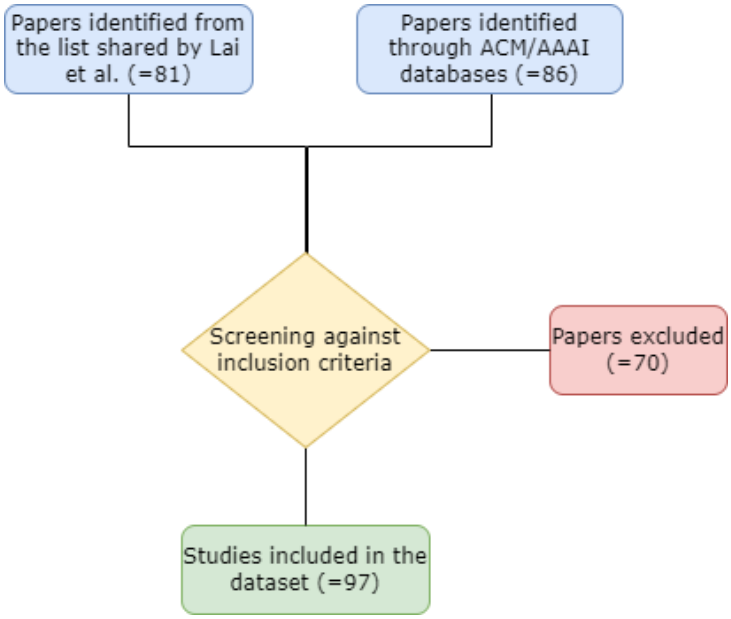


Figure 2.1.: Flowchart illustrating the paper selection process in creating the dataset for thematic analysis

task, study setup details such as the intended audience and participant information (number of participants, platform, filtering criteria, etc.), the role of participants and the task they had to perform, summary of the employed incentive scheme and task completion time, the excerpts describing the incentive scheme, and any excerpts containing discussions on participant motivation or incentive design.

The *excerpts* describing the incentive design and any discussions surrounding incentives served as the *dataset* for the thematic analysis. The rest of the information was deemed to be related to incentive design, and compiled in order to potentially facilitate future analysis.

This dataset is made available for public access here.

## 2.3. DATA ANALYSIS METHODOLOGY

For *qualitative analysis* of the dataset of excerpts developed in the previous section, we employ a *reflexive thematic analysis*.

*Practicing reflexivity is a significant part of our methodology.* Reflexivity is a conscious effort to acknowledge the researcher's role and its potential influence on how data is interpreted. It has also been framed as a way to embrace and value researchers' subjectivity [92]. Further, Braun and Clarke [72] emphasize that reflexivity goes

beyond self-reflection. It is a critical examination of both the knowledge produced from the research and how we produce it. This means justifying and reflecting on the chosen methodologies and being transparent about their potential implications and limitations.

Embracing reflexivity throughout the process of conducting this TA manifested in several ways:

*Acknowledging our positionality*: We recognized how our background and experiences might influence our interpretation of incentive design practices, captured in a positionality statement in Section 2.3.1.

*Justification of reflexive TA approaches*: We explicitly described the specific reflexive TA approaches we employed and explained our choices, outlined in Section 2.3.2.

*Iterative process and reflexive journaling*: We executed each of the six phases of reflexive thematic analysis iteratively, as described by Braun and Clarke [72]. We maintained a reflexive journal to capture evolving thoughts throughout the analysis. Additionally, the iterative nature of the thematic analysis, facilitated by tools like Atlas.ti, allowed for continuous refinement and ensuring that the final themes truly represent the data. We describe this process in Section 2.3.4.

*Critical reflection on developed themes*: Finally, we engaged in a critical reflection on the generated themes through discussing the potential implications of our observations and raising questions for future work in Section 2.5.

### 2.3.1. POSITIONALITY

Thematic analysis hinges on a researcher's ability to systematically extract meaning from qualitative data. However, researchers themselves are not blank slates. Their background, experiences, and biases - encapsulating their positionality - can influence how they interpret and analyze the data [72].

In order to successfully conduct *reflexive* thematic analysis, it is important for the researcher to acknowledge their position. This allows the researcher to be aware of their potential personal perspectives and yet aim to conduct a critical analysis. Thus, we acknowledge our position by means of a positionality statement given below.

#### POSITIONALITY STATEMENT

As a Master's student in Computer Science, my background lies primarily in the technical aspects of human-computer interaction (HCI) and artificial intelligence (AI). This academic foundation equips me with a strong understanding of the underlying concepts that shape human-AI collaboration. Through my involvement in human-AI decision-making projects, I have firsthand experience in studying and designing incentive schemes for the kind of studies that I have analyzed. This has directly

exposed me to the complexities of human behaviour within this domain and the challenges of aligning participant motivations within human-AI collaborative studies. Witnessing these dynamics firsthand has given me perspective on how well-designed or ill-designed incentives can influence these studies.

In undertaking this thematic analysis of incentive design in human-AI decision-making studies, I acknowledge the potential limitations of my position. While my experience is an asset, it may also lead to preconceived notions about how incentive design is, or should be, undertaken within such studies.

Ultimately, I believe that my position as a computer science student possessing a blend of theoretical knowledge and practical experience allows me to approach this research with a comprehensive lens. Further, I plan to rigorously adhere to reflexive thematic analysis methodologies and actively seek out diverse viewpoints within the literature. By acknowledging the limitations of personal perspective yet employing a rigorous approach, I strive to deliver an insightful analysis on this topic.

### 2.3.2. APPROACHES

Here, we justify the selection of the thematic analysis (TA) variations we employed, outlining how we approached *knowledge generation* within the analysis.

The different variations of reflexive TA, as described by Braun and Clarke [72], are briefly summarized in Appendix A.

We adopted a specific combination of these variations, as detailed below:

*Orientation to data*: We chose the *inductive* approach.
This core approach allows themes to emerge directly from the data itself, minimizing the imposition of pre-existing theoretical frameworks. We began with a broad research question and allowed the data to guide the identification of key themes related to incentive design.

*Focus of meaning*: We chose the *semantic* approach.
We paid close attention to the explicit content of the data, focusing on how incentive design is operationalized and discussed within the studies. This approach helped to ensure that the identified themes accurately reflect the researchers' practices and avoids misinterpretations.

*Qualitative framework*: We chose the *critical* approach.
This approach focuses on unpacking the broader meaning and implications around the topic. By critically analyzing the data, we aimed to uncover underlying assumptions and existing norms, providing a deeper understanding of incentive design.

*Theoretical frameworks*: We chose the *realist, essentialist* approach.
This approach aims to capture the objective truth and reality as expressed within the data. By adhering to this framework, we ensured

that our analysis remains grounded in the actual practices prevalent in the studies, providing an authentic representation of the current state of incentive design.

By employing this combination of reflexive approaches, we strive to achieve a rich and nuanced understanding of how incentive design is discussed and implemented within the recent literature on human-AI decision-making studies.



### 2.3.3. CHECKLIST

Braun and Clarke [71] also provide a "15-point Thematic Analysis Checklist" which outlines key criteria for conducting high-quality thematic analysis. It emphasizes the importance of systematic coding, clear documentation, and coherence between analysis and data. By meticulously following the checklist, we aimed to minimize bias and enhance the quality and reliability of our thematic analysis.

The checklist, applied to the TA conducted within this study, can be found here.

### 2.3.4. PHASES

This section details the iterative phases employed in our approach to the process of reflexive thematic analysis. The process is summarized in Figure 2.2.

#### 1. DATASET FAMILIARIZATION

This initial phase involved a thorough immersion in the data. We read and re-read the selected papers to gain an understanding of the research landscape and the terminology surrounding incentive design in human-AI decision-making. We also made notes within the spreadsheet to keep track of any interesting points that emerged.

#### 2. DATA CODING

This phase involved systematically assigning codes to segments of text in the data. Here, we practiced *inductive* coding to allow themes to emerge organically from the data itself. We conducted two complete iterations of coding all the excerpts. Atlas.ti[2] proved to be a valuable tool for this phase. It was used to assign and organize codes for relevant parts of the excerpts.

The final coded quotations and codebook generated through Atlas.ti are available here.

[2] https://atlasti.com/

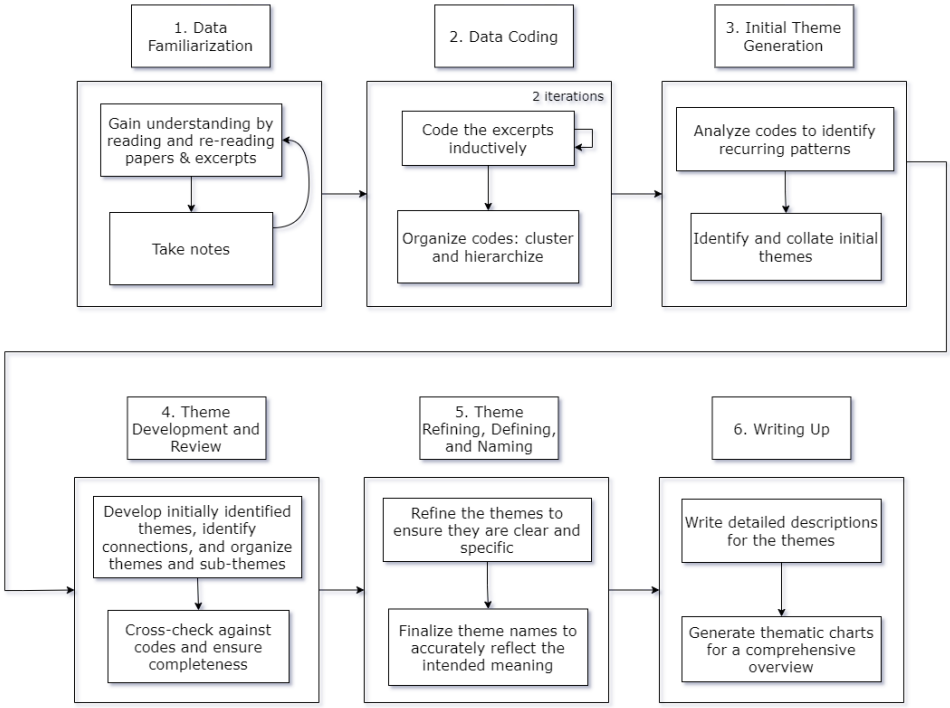


Figure 2.2.: Flowchart summarizing the phases of reflexive thematic analysis as performed for this study

### 3. Initial Theme Generation

The phase of initial theme generation involved analyzing the codes generated in the previous phase and identifying broader patterns that emerged. Here, we sought to identify recurring patterns across the assigned codes and transform them into meaningful themes. By systematically analyzing the codes and employing a *critical* approach, we were able to identify a set of initial themes.

### 4. Theme Development and Review

This phase involved developing and refining the initial themes. Multiple iterations were done to refine and group themes and sub-themes into meaningful thematic clusters. Atlas.ti allowed for visualizing code co-occurrence and creating code maps, aiding in the identification of potential thematic connections. We examined the coded data in relation to each theme, ensuring that the themes accurately represented the data and offered a coherent narrative.

#### 5. THEME REFINING, DEFINING, AND NAMING

In this phase, we further polished the themes and finalized their names, ensuring clarity and specificity, preserving the *semantic* meaning from within the data.



#### 6. WRITING UP

The final phase involves writing up the results of the thematic analysis. We present the identified themes organized into thematic charts, along with detailed descriptions and supporting evidence drawn from the coded data in the next section.

## 2.4. RESULTS

Here, we detail the themes that have been identified after performing a rigorous reflexive thematic analysis on the dataset of excerpts that describe incentive schemes in human-AI decision-making literature. We also accompany the descriptions of the themes with specific examples from the literature, as guided by the Braun and Clarke [72] reflexive TA checklist (mentioned in section 2.3.3).

### 2.4.1. THEME 1: COMPONENTS OF AN INCENTIVE SCHEME

Occurring in 88/97 papers, the components of an incentive scheme is the most common theme. We further identified two sub-themes within it, namely *base pay* and *bonus*. These are shown in Figure 2.3.

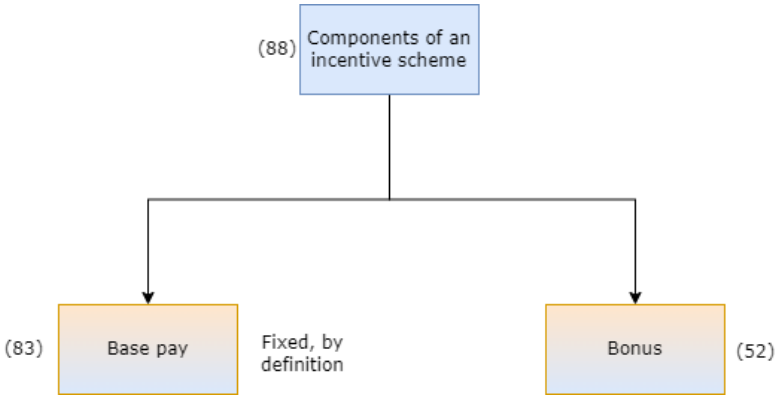


Figure 2.3.: Top-level thematic chart for Theme 1: Components of an incentive scheme

47/88 papers mentioned both a base pay and a bonus. For instance, Lim, Dey, and Avrahami [93] mentioned, "*Participants were each given*

*$3 for completing the study ($1 base and a $2 bonus to motivate performance)...*". Similarly, Dietvorst, Simmons, and Massey [94] wrote, "*Participants received $1 for completing the study and they could earn up to an additional $1 for accurate forecasting performance.*"

In 41/88 papers, researchers only mentioned a base pay and not a bonus. For instance, Poursabzi-Sangdeh *et al.* [95] wrote, "*Each participant received a flat payment of $2.50.*" Meanwhile 5/88 papers only mentioned a bonus, and not a base pay, like Biran and McKeown [96]: "*...we offered (relatively) large bonuses to the two annotators who made the most virtual money.*"

#### THEME 1.1: BASE PAY

83/97 papers mentioned base pay. Figure 2.4 captures this theme and its sub-themes.

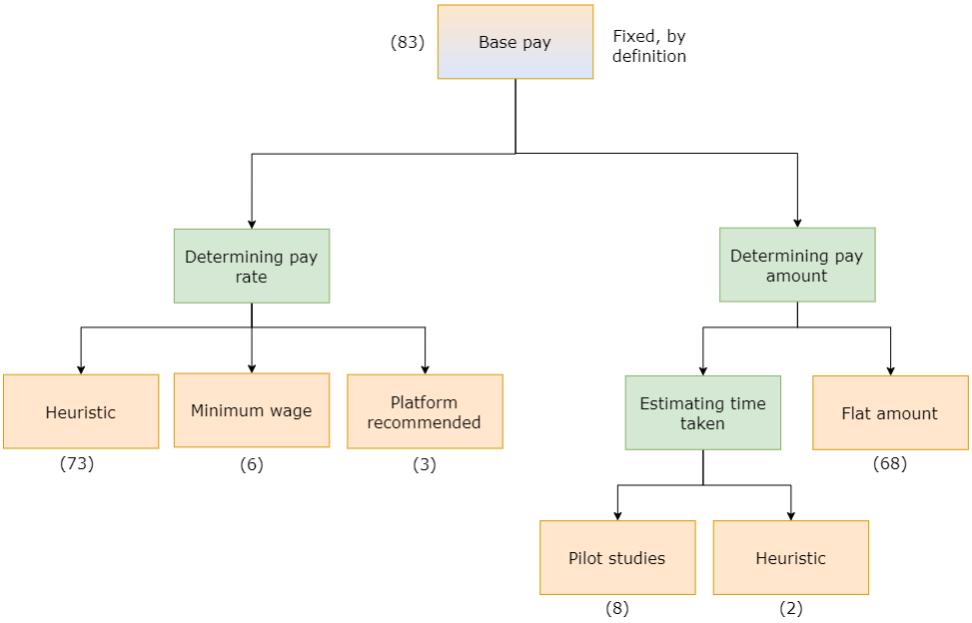


Figure 2.4.: Thematic chart for Theme 1.1: Base Pay and its sub-themes

Base pay is identified as the core or basic payment that participants receive for completing the tasks assigned to them. To describe base pay, researchers either directly identify a *pay amount* or they identify a *pay rate*, i.e., a certain amount per hour, or both.

For instance, Buçinca *et al.* [51] wrote, "*Each worker was paid 2 USD.*" Meanwhile Yurrita *et al.* [97] mentioned, "*Participants were rewarded based on a $12 hourly rate...*". Chromik *et al.* [52] described both:

"*£3.75 per completion (=£7.09/hour)*"

Pay rates and amounts are most frequently determined *heuristically*, with no explanations given for how the amount was decided. In a few cases, the pay rate is explicitly said to be informed by *minimum wage*, or *platform recommendations*. For example, Liu *et al.* [98] mention, "*To provide fair compensation to our participants, Mturkers were offered an equivalent of United States federal minimum wage...*" while He, Buijsman, and Gadiraju [99] mention, "*All participants were rewarded with [...] deemed to be "good" payment by the platform...*".

Researchers usually just present a *flat amount*. In a few cases, researchers mention the *estimated time taken* to complete a task along with the pay amount or rate. Estimation is mostly done through *pilot studies*, and in a couple of cases, *heuristically*. For instance, Jahanbakhsh *et al.* [100] mentioned: "*From our pilot studies with our research group, we determined that the average time for completing the task was approximately an hour. Therefore, we set a compensation of $17 for the task.*"

#### THEME 1.2: BONUS

52/97 papers mentioned bonuses. A maximum possible bonus *amount*, per task or for the whole study, is decided *heuristically*. For instance, Lai, Liu, and Tan [27] mentioned that, "*Each participant was compensated $2.50 and an additional $0.05 bonus for each correctly labeled test review.*" Meanwhile Hou, Lee, and Jung [66] mentioned "*All participants were paid $7.50, including the base value $5.50 plus a $2.00 bonus payment...*".

A *payout scheme* outlines how the final bonus amount which is to be paid to the participants is to be calculated. These schemes can be classified into different *types*, based on the criteria which is used to determine the payout. These are identified as *performance-based*, *completion-based*, and *luck-based*.

Figure 2.5 illustrates the high-level theme and sub-themes of bonuses.

*Performance-based* schemes are the most commonly used. Participants are paid based on their performance, which is evaluated using certain *performance metrics*. A policy for *mapping performance to pay* is usually also presented. For instance, Dressel and Farid [23] noted that, "*The participants were paid $1.00 for completing the task and a $5.00 bonus if their overall accuracy on the task was greater than 65%...*".

For *completion-based* bonuses, a fixed amount is paid when a participant completes a specific task, such as responding to optional surveys. For instance, Bansal *et al.* [18] mentioned, "*Participants received [...] a fixed bonus of $0.25 for completing the survey...*".

For *luck-based* bonuses, there is usually a *randomized* element to determining whether a participant receives a bonus. Randomization is

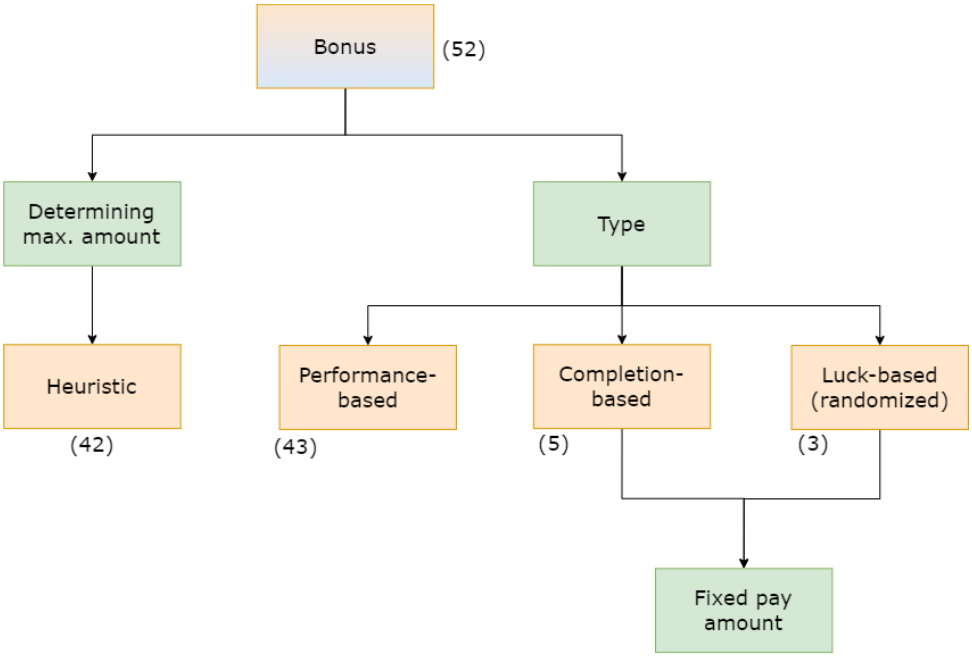


Figure 2.5.: Thematic chart for Theme 1.2: Bonus and top-level sub-themes

introduced within the design in different ways, such as introducing a specific chance for a participant to earn a bonus, or picking one task at random for each participant to evaluate based on some criteria. For example, Lu and Yin [101] mentioned, "*...we randomly selected one prediction task in the sequence to check whether the subject's final prediction on that task was correct. If so, the subject would receive a $1 bonus on top of the base payment.*"

Further, there are sub-themes that emerge specifically within performance-based bonuses. These are captured in Figure 2.6, and discussed below:

Firstly, *performance evaluation metrics* are used to evaluate participant performance, with the most common being *accuracy*. For instance, Alqaraawi *et al.* [102] mentioned, "*...participants received an additional performance-based bonus of £0.5 for each correct answer...*"

Other examples of metrics include in-game currency (gamified score for game-based tasks) [103], custom defined metrics such as 'precision + coverage' [59], and expert judgements (experts evaluating participant performance) [104].

Secondly, there are several ways to *map performance to pay*. These include policies such as fixed payout when performance is over a

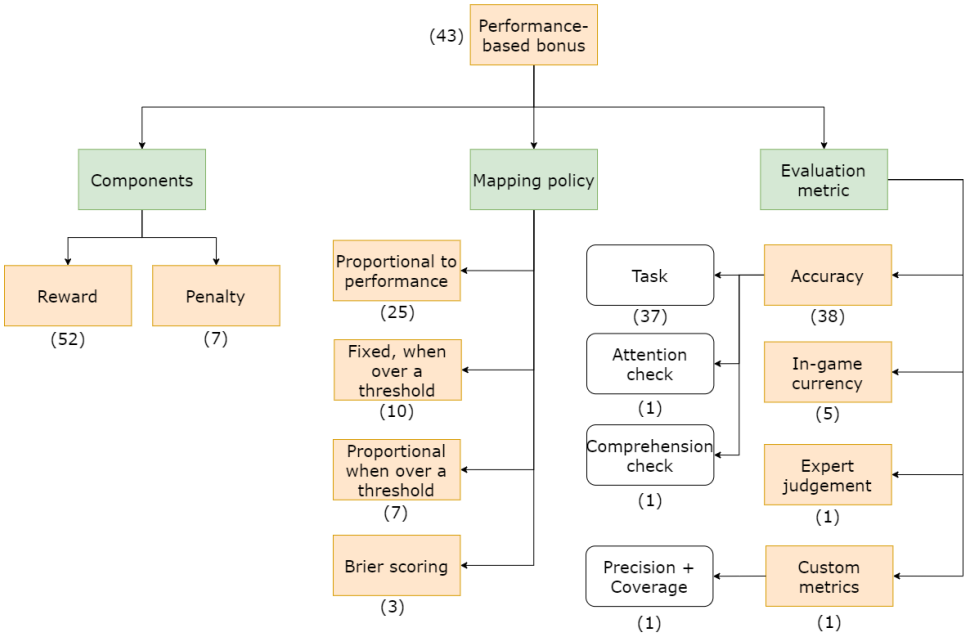


Figure 2.6.: Thematic chart for *performance-based bonus* and its sub-themes

threshold [23], payout proportional to the performance [27], payout proportional to the performance when performance is over a threshold [105], and payout calculated through the Brier scoring [106] function [107]. For instance, Dietvorst, Simmons, and Massey [12] describe their policy as, "*Participants were paid a $0.50 bonus if their official forecasts were within five percentiles of students' actual percentiles. This bonus decreased by $0.10 for each additional five percentiles of error [...]. As a result, participants whose forecasts were off by more than 25 percentiles received no bonus.*"

Further, a performance-based bonus can also have a *penalty* (negative incentive) alongside a *reward* (positive incentive). For instance, Zhang, Liao, and Bellamy [17] mentioned, "*...a reward of 5 cents if the final prediction was correct and a loss of 2 cents if otherwise...*".

The design specifications of such policies are often heuristic, with only a couple of cases being grounded in prior literature [17, 107].

### 2.4.2. THEME 2: MANIPULATION OF INCENTIVES

The next identified theme is that incentives are *manipulated* for various purposes. Figure 2.7 illustrates the chart for this theme and its sub-themes.

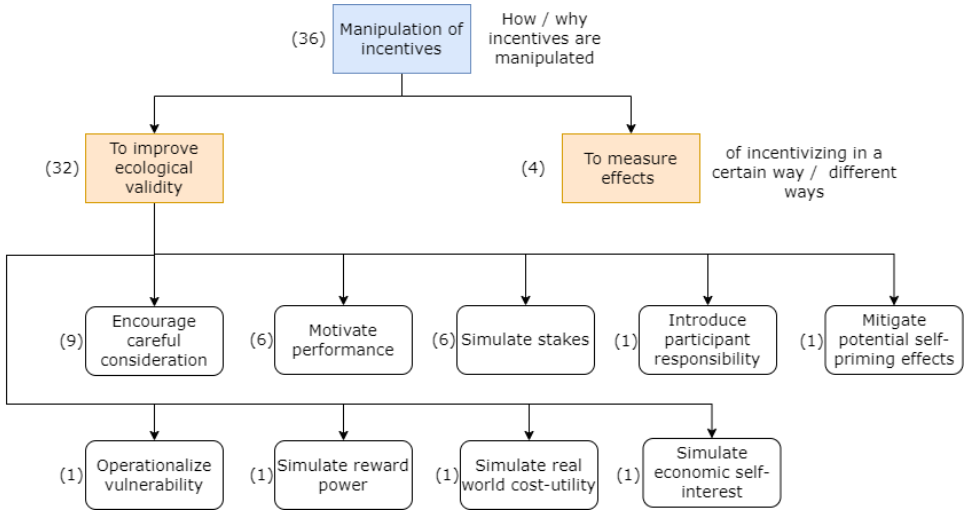


Figure 2.7.: Thematic chart for Theme 2: Manipulation of incentives

36/97 papers mention using, or manipulating, incentives for certain purposes. Largely, the purpose is identified as *improving the ecological validity* of the experiments. For instance, Zhang, Liao, and Bellamy [17] said, "*We took two measures to improve the ecological validity. First, the decision performance was linked to monetary bonus...*".

Researchers mention using incentives (mainly bonuses) to: "motivate performance" [21], "encourage participants to pay attention" [23], "simulate stakes" [108], "simulate reward power" [66], and more. We identify these as examples (and not sub-themes) of the broader theme of improving the ecological validity.

Finally, bonus schemes are also intentionally varied (such as paying high vs. low bonuses) to *measure the effects* of the different schemes on participant performance or research outcomes. For instance, Yin, Wortman Vaughan, and Wallach [13] mentioned, "*We also posited and pre-registered two additional hypotheses: [H3] The amount at stake has a significant effect on people's trust in a model before seeing the feedback screen. [H4] The amount at stake has a significant effect on people's trust in a model after seeing the feedback screen [...] to test whether the effect of stated accuracy on trust varies when people have more "skin in the game".*"

### 2.4.3. THEME 3: IMPACT OF INCENTIVES

The third theme is identified as the *impact* of incentive schemes on the results of a study. Figure 2.8 illustrates the chart for this theme and its sub-themes.

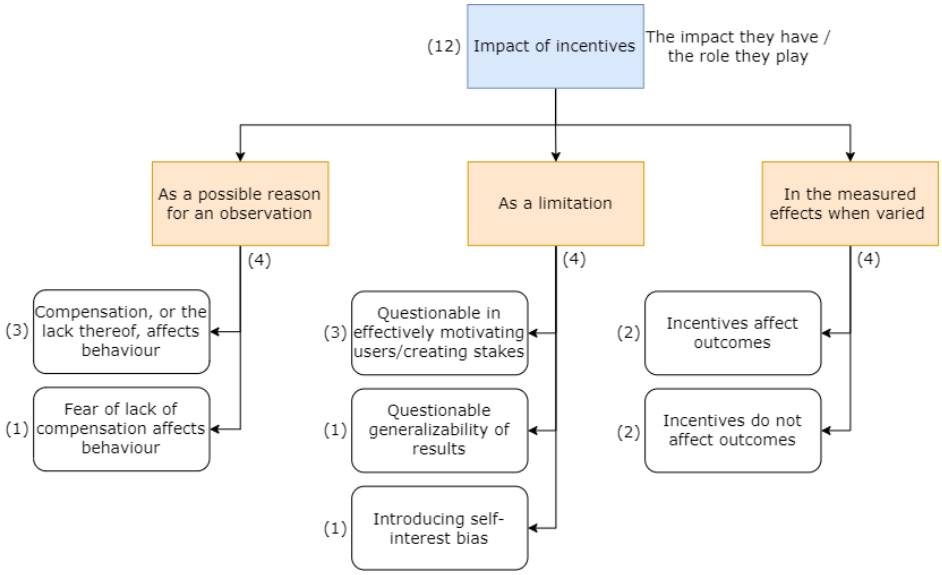


Figure 2.8.: Thematic chart for Theme 3: Impact of incentives

12/97 papers mention the impact of incentives, reflecting on the potential role that incentives played in their studies. Some researchers discuss the limitations of incentive schemes, questioning whether they can effectively replicate real-life scenarios or motivate crowdworkers, acknowledging uncertainty regarding the generalizability of their results. For instance, Dietvorst, Simmons, and Massey [12] noted that, "*Because real life end-users can have very different demographics characteristics and non-monetary incentives and operate in higher-stake environments, we cannot reliably generalize the finding to real workplaces...*".

In a few cases, incentives are also attributed to as the *potential reason behind an observation*, such as a particular trend being observed because participants may or may not have been motivated to perform due to compensation, or the lack thereof. For example, Gemalmaz and Yin [103] note that, "*There are two important caveats to the analysis. First, it relies on non-incentivized self-reported data near the end of the experiment. Thus, we cannot verify that subjects reflected carefully on their answers...*".

Lastly, the few papers that intentionally manipulate incentives to measure their effects also describe their results, leading to a discussion on the impact of incentives. For these studies, different researchers found that incentives may or may not affect performance or outcomes under different conditions. For instance, Vasconcelos *et al.* [62] described that, "*...we found an impact of rewards on overreliance...*". Meanwhile Yin, Wortman Vaughan, and Wallach [13] found that, "*...the*

*amount at stake does not have an effect on laypeople's trust in a model, at least for the limited range of stakes used in our experiment.*" We note that such studies were conducted for dissimilar tasks and domains, under differing conditions, measuring the effects of rewards on different variables.



### 2.4.4. THEME 4: COMMUNICATION OF INCENTIVES

This theme highlights the different trends in the *communication* of incentive schemes to participants. Figure 2.9 illustrates the chart for this theme and its sub-themes.

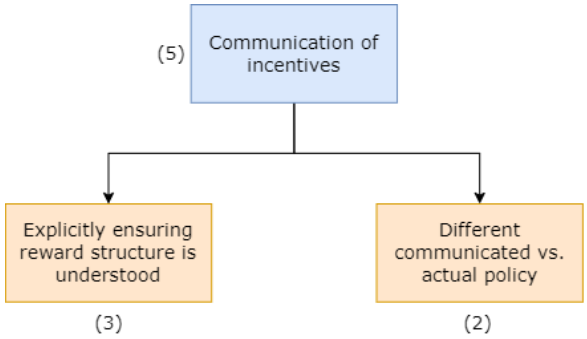


Figure 2.9.: Thematic chart for Theme 4: Communication of incentives to participants

5/97 papers discussed the communication of their incentive schemes. In some cases, researchers explicitly ensure that participants understand the pay structure before they begin the task. For example, Green and Chen [107] described that, "*...participants were incentivized to report their true estimates of risk. We articulated this to participants during the tutorial and included a question about the reward structure in the comprehension test to ensure that they understood.*"

There are also a couple of instances where the bonus scheme communicated to participants differs from the actual method used to calculate the final payment received by participants. For instance, Das and Chernova [109] mentioned that, "*...participants were told that at least the top 50% of participants would be given a $2 bonus based on the thoroughness of their evaluations; unbeknownst to them, all ultimately received the bonus.*"

### 2.4.5. THEME 5: NO MENTION OF INCENTIVES

8/97 papers did not mention anything regarding incentive schemes [68, 70].

## 2.5. DISCUSSION

Here, we discuss the key observations made from the results of the thematic analysis. We reflect on the implications of these observations, raising questions that can guide further investigation.



### 2.5.1. PAY AMOUNTS

Our investigation revealed that the current strategies for determining base pay and bonus amounts are predominantly heuristic in nature. Furthermore, a lack of consistency was observed within the calculation and reporting of base pay and bonuses.

For base pay, researchers employed a disparate approach, with some specifying a flat amount and others opting for a pay rate. Additionally, the explanations behind the base pay calculation varied. While some studies explicitly outlined parts of the process, such as estimating task completion time for amount calculation, others lacked such transparency. For bonuses, there was rarely any discussion regarding how the allocated amount was determined. These observations highlight the need for a more standardized approach towards determining base pay and bonus amounts as well as reporting them.

> ***Is there an optimal or appropriate pay amount? Can there be a standardized way of determining it?***

Could future work explore data-driven methods for establishing optimal pay levels? Can we develop a standardized methodology for calculating base pay and bonus amounts across different tasks and platforms?

### 2.5.2. ASPECTS OF BONUS SCHEMES

The examination of bonus schemes further revealed many different approaches within the different aspects of bonuses, such as a variety of types of bonuses (performance-based, completion-based, luck-based), bonus calculation policies (fixed, proportional, etc.), and performance evaluation metrics. Notably, not all studies comprehensively addressed each aspect that we identified. Future work should aim at making the process of bonus design more comprehensive. Further investigation is needed into which approaches are most suitable under different task conditions.

> ***Can we develop a standardized methodology that comprehensively recommends specific bonus calculation methods based on study conditions?***

Is one type of bonus scheme more appropriate or suitable than the other? Which metrics are most suitable for evaluating performance? What is the best way to map performance to pay? Can these policies be optimized? Under what conditions?

### 2.5.3. USE OF REWARDS AND PENALTIES

The analysis highlighted the use of bonuses alongside base pay, with some studies even employing penalties to simulate high stakes. Further research is needed to determine the effectiveness of these approaches.

***Is there value in the use of rewards and penalties? Does it create "higher" stakes?***

Does the inclusion of bonuses and penalties truly create "higher stakes", or does it introduce ethical concerns? Is there a balance to be struck between positive and negative incentives that optimizes participant motivation without compromising ethical considerations?

### 2.5.4. IMPROVING ECOLOGICAL VALIDITY

The analysis highlighted concerns about the ability of incentives to create "stakes" and enhance the ecological validity of studies. Future research should explore methods for evaluating the impact of incentive structures on participant behaviour and their ability to enhance ecological validity in actuality.

***Can it be evaluated and/or ensured that incentives have the intended effect on participant performance and behaviour? How?***

Can some practices be devised to objectively assess whether incentives truly improve the ecological validity of an experiment?

### 2.5.5. EFFECTS OF INCENTIVES

The analysis highlights a complex relationship between incentive schemes and research outcomes in crowdsourced studies. We saw that some studies found that incentive schemes affect outcomes, while others did not. This suggests that there's no single "formula" for incentive design. The effectiveness of an incentive scheme likely depends on several factors, such as the experimental conditions and design. This begets a deeper exploration within this focus.

***Which kind of incentive schemes affect outcomes? Which do not? Under what conditions?***



### 2.5.6. LIMITATIONS OF INCENTIVES

The analysis highlighted that researchers show concerns regarding the limitations of using incentives in crowdsourced studies, particularly with respect to the generalizability of their findings and potential bias due to crowdworker motivations.

***Can better incentive design practices mitigate the limitations associated with their use? How can incentives be designed to better understand ways in which they affect results?***

Can practices be devised to *intentionally design* incentives that can overcome limitations?

### 2.5.7. COMMUNICATION STRATEGIES

Explicitly ensuring that incentive schemes are understood and performance feedback can improve participant understanding and potentially enhance results, but it might also introduce bias or demotivate participants, for example, when they realize they are not doing well enough. More investigation is needed into which strategy can be suitable when. Further, an ethical dilemma lies in communicating one bonus policy and implementing another.

***Is there value in different types of communication strategies for incentives?***

Is it more appropriate to pay for performance as communicated, or pay everyone equally?

### 2.5.8. MISSING INCENTIVE SCHEMES

Lastly, we noticed that some researchers chose not to discuss incentive schemes for their studies. We note that the absence of description doesn't necessarily mean incentive schemes weren't thought about or employed. Researchers possibly might not have considered mentioning them relevant to their specific goals.

***Should researchers be encouraged to consistently report on incentive schemes used in crowdsourced studies, regardless of their perceived relevance? If so, how?***



By addressing the questions raised in this section, we can move towards a more nuanced and context-specific approach to incentive design. The different trends we identified lay the foundation for developing solutions that can help improve the incentive design process.

## 2.6. LIMITATIONS

Here, we discuss some limitations of the thematic analysis we conducted. We also identify potential mitigation strategies.

### 2.6.1. SINGLE-PERSON EXECUTION TEAM

Ideally, a thematic analysis should be conducted by a team of multiple researchers [72]. This allows for different perspectives and mitigates the influence of individual biases during the coding and theme development process. In this study, the analysis was conducted by a single researcher, which potentially limits the breadth and richness of the identified themes. In the future, a larger team of researchers could execute the process to enhance the quality and validity of our findings.

### 2.6.2. UNCAPTURED NOTIONS

Our thematic analysis relies on the information explicitly presented by researchers. Specifically, it is focused on what researchers described in their papers regarding incentive schemes. While we conducted a critical analysis of the research articles to identify underlying meanings, we acknowledge that there are likely unarticulated assumptions or thought processes behind the described incentive schemes that remained uncaptured. For instance, heuristic determination of pay amounts could be attributed to budget constraints. However, more information is needed regarding researchers' reasonings behind the design choices they made, to be able to comment further.

To address this, we could follow-up our findings by conducting interviews with researchers who have experience in designing incentive schemes. Analyzing them could provide valuable insights into the thought processes and motivations of researchers.

# 3

# DESIGNING INCENTIVE SCHEMES FOR HUMAN-AI DECISION-MAKING STUDIES: A CHECKLIST

"*Under conditions of complexity, not only are checklists a help, they are required for success. There must always be room for judgment, but judgment aided — and even enhanced — by procedure.*"

Atul Gawande, The Checklist Manifesto

*In this chapter, we propose a step-by-step guide, the Incentive-Tuning Checklist, to help researchers in designing incentive schemes for human-AI decision-making studies. We first argue for a flexible yet systematic approach to designing incentive schemes, moving beyond the traditional focus on finding a "correct" answer. We also emphasize careful consideration of study goals and desired behaviours from participants. We also demonstrate how to apply the checklist through case studies, highlighting the importance of aligning incentives with study goals, considering trade-offs, justifying and reflecting on decisions, and incorporating participant feedback throughout the design process.*

*We hope that by following the checklist and staying updated on the field's advancements, researchers can design effective and ethical incentive schemes, ultimately improving the quality and validity of human-AI decision-making research.*

## 3.1. TOWARDS RQ2

Building upon the insights obtained from RQ1 and moving towards addressing the research objectives concerning RQ2, this section delves into the complexities of designing incentive schemes for human-AI decision-making studies.

### 3.1.1. INSIGHTS FROM RQ1



The observations made during RQ1, while insightful, raised several questions for the future of incentive design for human-AI decision-making studies. These questions fall into two main categories: *methodological challenges* and *exploratory avenues*. Methodological challenges concern the practical implementation of incentive schemes, such as the *design* of pay amounts and structures. Exploratory avenues encompass broader research directions, including studying the *effects* of manipulation and different communication strategies.

While a comprehensive exploration of all these questions is beyond the scope of this thesis, we prioritize addressing the methodological challenges surrounding incentive design, as we identify that these concerns align with our defined research agenda, particularly RQ2:

> *RQ2: How can incentive schemes be appropriately designed through a standardized process for empirical human-AI decision-making studies?*

The findings of the thematic analysis further solidified the need for standardization in light of the challenges inherent to human-AI decision-making tasks, as we suspected while defining RQ2. Moreover, the results challenge the very notion of "appropriate" incentive design.

Thus, in order to address RQ2 and some of the questions raised in the previous chapter, we take up the task of establishing a standardized solution that can allow researchers to tune "appropriate" incentive schemes for their specific human-AI decision-making studies.

### 3.1.2. EXPLORING "APPROPRIATENESS"

Our pursuit of an "appropriate" incentive design solution for human-AI decision-making studies necessitates a deeper examination of the term itself. What truly constitutes "appropriate" in this context? After careful consideration, we deem that the answer is not straightforward. Appropriate design often depends on various factors, and pinpointing these factors themselves presents a significant methodological challenge. In an ideal scenario, it would be possible to map out all potential experimental factors against all possible incentive design choices, allowing

for optimization based on all specific study conditions. However, this is practically infeasible. This, in turn, raises the question: how to proceed further with RQ2?

### 3.1.3. ADOPTING A NORMATIVE LENS

While the ideal scenario of mapping all factors to all design choices may be unrealistic, it doesn't mean that there aren't effective strategies that can be pursued. Here, we propose a shift in perspective: moving from a purely pragmatic "what we can do" approach to a more *normative* "what we should do" approach. Traditionally, incentive schemes seem to often have been designed with the former in mind, prioritizing feasibility over well-motivated and justified design decisions.

Thus, to crystallize the normative approach we shall be taking from here on, RQ2 is re-worded as:

> *RQ2: How **should** incentive schemes be appropriately designed through a standardized process for empirical human-AI decision-making studies?*

By adopting this normative lens, we move beyond trying to tune appropriate incentive schemes for different experimental factors and/or study conditions. Instead, our goal is now to propose a standardized yet flexible process that can *guide* careful consideration and justification of design decisions for incentive schemes. This process should be such that it allows researchers to *tune* "appropriate" incentive schemes for their studies *themselves*.

### 3.1.4. COMING UP WITH A PROCESS

A process that aims to effectively guide researchers towards an incentive scheme appropriate for their studies must involve asking the *right* questions. These questions should address the several decision to be made during the design process. These can be questions such as, but not limited to, the following: What are the goals of the study? Is the aim to maximize participant effort, encourage specific decision-making processes, or achieve a balance between both? What characteristics of the task can have an effect on the suitability of incentive design? How can the chosen incentive scheme influence the generalizability of study results?

Following this initial questioning phase, the process should encourage exploration of the available design options. It can often turn out that there is not a single "right answer" and several trade-offs may need to be considered. In such cases, rationalization and justification become imperative. Carefully considering the potential implications

of each design choice and how it aligns with the overall study goals should thus be at the forefront of incentive scheme design. This might involve considerations such as, but not limited to, the following: comparing and contrasting different incentive structures (e.g., fixed pay vs. performance-based bonuses) and/or piloting the chosen scheme with a small sample size to test its effectiveness and identify potential issues.

Finally, it's crucial to reflect on the impact of the chosen incentive scheme design on the final analysis. The process must nudge researchers towards reflecting on how the scheme might have influenced study outcomes. Researchers must be upfront about the chosen incentive scheme design as well as the limitations introduced by it, and how they might affect the validity of the findings.

To summarize, we need a solution that:

- systematically tackles each incentive design aspect,
- raises and delves into the right questions,
- guides careful consideration of design choices and trade-offs,
- encourages reflection around the potential implications.

Keeping these requirements in mind, we propose a **checklist**.

#### WHY A CHECKLIST?

Checklists have been described as "informational artifacts" that *conceptualize* the actions and decisions comprising a process [110]. Checklists can act as a clear and concise description of a process by outlining the essential steps involved in it. Further, by highlighting decision points, they can guide users through the process, ensuring they make informed choices and address all necessary tasks. This ability to represent and guide complex processes elevates checklists beyond simple "information lists". They can become a *conceptual model* of the process [110]. Through this lens, checklists can act as a bridge between knowledge (what should be done) and action (how to do it).

Further, checklists have been argued to be powerful tools for overcoming human limitations in conducting complex processes. Critical industries like aviation, manufacturing, and healthcare rely heavily on checklists to ensure safety and quality [111, 112]. For instance, in healthcare, checklists have become a powerful tool for improving patient safety and reducing errors. They act as a memory aid for busy healthcare professionals, ensuring critical steps aren't missed in complex procedures such as surgeries [113, 114].

For our use case, a well-designed checklist can foster a standardized, systematic, and comprehensive approach that can guide research towards designing effective incentive schemes. It can further act as a

safeguard against unintentionally overlooking the important aspects of incentive design.

## 3.2. CHECKLIST DESIGN METHODOLOGY

Existing literature identifies that checklists have a multifaceted nature, highlighting the various properties that they can have [110]. Additionally, there are many valuable recommendations for designing checklists [112, 115, 116]. Informed by this knowledge base and the requirements outlined in the previous section, here we define the key properties of the checklist that we aim to develop.



- The checklist will be of the "*read-do*" type [116]. It will take researchers *through* the process of design. It will further break down the process into *manageable* and *sequential* steps.
- Checklists can have strict rules (criteria) or guiding principles (guidelines) [117]. Our checklist will embody *guidelines* instead of criteria. This approach reflects our goal of empowering researchers rather than dictating a single "correct" approach, ensuring *flexibility* by design.
- Checklist items will be developed through a *literature support* [118] strategy. We will draw upon the key themes identified by the thematic literature review we conducted, as well as the discussion and reflection that followed, to determine the checklist content.
- The checklist will also take an *interrogative* [110] approach. We will implore researchers by raising questions, prompting them to delve deeper into the design process at each step.
- We note that checklists should not be too long and verbose [111, 115]. However, we also want to address the nuances and trade-offs that can emerge during the design process. Hence, we elect that the checklist will be *accompanied* by in-depth discussions on its items and, wherever possible, also provide well-informed suggestions grounded in prior literature.
- Lastly, we explicitly identify the *intended audience* and *study field* of the checklist. The checklist will be built for researchers conducting empirical, crowdsourced studies in human-AI decision-making.

### 3.2.1. DETERMINING THE CONTENT

Here, we delve into identifying the key elements that will form the core content of our checklist.

In the previous chapter, a thematic analysis revealed that researchers manipulate incentives to achieve certain goals. We recognize the

importance of aligning the incentive design with those goals from the very beginning. Therefore, first and foremost, we prioritize this step within the checklist, ensuring that incentive scheme design is informed by research objectives.

Naturally, this should be succeeded by defining the incentive scheme itself. Our thematic analysis also revealed several insights into the components that form an incentive scheme. The next steps thus directly build upon the components we identified. We noticed that for each component there are several decisions to make along with various choices. As we noted in the previous section, it might not be feasible to attempt to identify which choice is optimal when. However, we can identify that the several choices would present *trade-offs*, as different choices can have different impacts. Thus, we develop our suggestions in a way so as to not prescribe the "best" option but rather offer guidance and allow scope for considering trade-offs.

Next, we also saw that researchers conjectured on the effects of incentives and questioned participant motivation, wondering if some results could be attributed to such factors. To address this, we propose an intervention within the experimental process: gathering participant feedback. This would allow for post-hoc reflection on the results from the participants' perspective. This could potentially enhance confidence in the intended effects of incentives or provide valuable context for interpreting unexpected observations.

Lastly, as a good practice we assert that researchers should reflect on the final design decisions and their potential implications. The checklist and accompanying suggestions thus also include pointers that encourage reflection, such as prompting researchers to reason around how the chosen incentive scheme might influence participant behaviour and data interpretation.

### 3.2.2. WHEN TO APPLY

Before outlining the complete checklist and delving deeper into its specifics, it's crucial to identify the ideal timing for its application. The checklist is most beneficial once the experiment design is finalized. This ensures key details, such as those mentioned below, are known beforehand:

- *The experimental pipeline*: This refers to the sequence of steps participants will go through in the experiment. This can include instructions, training tasks, the number of tasks to be completed, and surveys. Understanding this pipeline is essential because it can influence incentive design choices. For example, a lengthy experiment with complex tasks might necessitate a higher pay to maintain participant motivation throughout.

- *Task characteristics*: The specific characteristics of the tasks themselves can also play a crucial role in shaping incentive design. Some of these characteristics as identified by Lai *et al.* [11] are:
  - *Task complexity*: More intricate tasks might require adjustments to incentive structures to encourage participants to invest the necessary effort and attention.
  - *Risk*: Tasks with a high potential for negative consequences (e.g., recidivism) in the real-world might necessitate bonus structures that emulate raised stakes for (often financially-motivated) participants to enhance the applicability of the results.
  - *Subjectivity*: Tasks that involve subjective judgments or interpretation can be challenging to design incentives for. Researchers might need to clearly define evaluation criteria and potentially come up with custom criteria to ensure fairness and consistency in participant evaluation.
  - *Required expertise*: Tasks that require specialized knowledge or skills might warrant higher pay to compensate participants for their expertise. Additionally, bonuses might be warranted for demonstrating exceptional skill or knowledge application within the task.



By having the experiment design and task characteristics be clearly defined before its use, the checklist can become a powerful tool for identifying the relationships between these elements and the various components of an incentive scheme. This can allow researchers to make informed decisions about base pay, bonus structures, and other incentive elements that can ultimately encourage the behaviours they desire in participants within the specific context of their studies.

The next section delves deeper into the checklist itself, outlining how it guides researchers through systematically identifying the purpose of the incentive scheme, analyzing task characteristics and other external factors, and aligning them with incentive design choices.

## 3.3. THE INCENTIVE-TUNING CHECKLIST

This section lays out the *Incentive-Tuning Checklist* itself. This meticulously crafted tool is aimed to serve as a step-by-step guide for researchers to foster *intentional design* by prompting them to consider various factors and carefully tailor incentives to their specific study.

The checklist is presented below:

1. **Identifying the purpose of employing an incentive scheme**: *What is the goal we wish to achieve by incentivizing participants? What are the desired behaviours we expect from participants?*
2. **Coming up with a base pay**: *What is the minimum flat rate participants must be paid to ensure the identified goals are met? Which characteristics of the task can affect this amount?*
3. **Designing a bonus structure**: *Can offering bonuses help reach the identified goals?*
   a) **Coming up with a bonus amount**: *How much total amount should be allocated for bonuses?*
   b) **Deciding the type of bonus(es)**: *What type of bonus is the most suitable for reaching the identified goal?*
   In case of performance-based bonuses:
   c) **Deciding the performance evaluation metrics**: *How to evaluate participant performance? What is the behaviour that we wish to reward?*
   d) **Deciding the policy for mapping rewards to performance**: *How to map performance to rewards?*
4. **Gathering participant feedback**: *What do we wish to understand from the participants' perspective? Can we augment the experiment pipeline to gather such feedback?*
5. **Reflecting on design implications**: *Did the design achieve the desired goals? Were there any unintended consequences?*

Each subsection henceforth discusses an *item* of the checklist, which in turn represents a *step* in the process of designing an incentive scheme. Through these well-defined steps, the checklist aims to ensure that researchers systematically address each crucial component of incentive scheme design. The discussions presented below aim to guide researchers through each step, while providing suggestions wherever possible, to allow them to *tune* their incentive scheme as suitable for their human-AI decision-making study.

### 3.3.1. IDENTIFYING THE PURPOSE

*What is the goal we wish to achieve by incentivizing participants? What are the desired behaviours we expect from participants?*

The first step in designing an incentive scheme is identifying the goal we aim to achieve by incentivizing participants. This is important to facilitate the *intentional design* of our incentive scheme, so that we can ensure that the chosen incentive scheme aligns with our research goals and reflect on our results from a (participant) motivation perspective.

Some possible goals researchers might aim to achieve are:

- *Meet ethical standards*: Ensure fair wages are paid.
- *Improve ecological validity*: Incentivize behaviour that reflects real-world actions to increase the generalizability and validity of results [119].
- *Create or simulate stakes*: Replicate real-world consequences to raise the stakes for the participant.
- *Motivate performance*: Encourage participants to put in their best effort.
- *Encourage careful consideration*: Motivate participants to thoroughly process information and pay attention throughout the task.
- *Encourage specific behaviours*: Reward desired actions that we require participants to engage in, beyond completing or performing "well" on tasks.



### 3.3.2. COMING UP WITH A BASE PAY

*What is the minimum flat rate participants must be paid to ensure the identified goals are met? Which characteristics of the task can affect this amount?*

Researchers must consider their identified goals as well as the characteristics of the task while deciding how much base pay to offer for completing the task. Further, other factors such as the minimum wage standard, platform-recommended rates, and pricing of similar past studies can also be relevant when coming up with an amount. Lastly, pilot studies can be used to assess the sufficiency of the base pay to be offered, and help tune it to better meet participant expectations of fair pay.

At the very least, the identified goals may include

- paying participants fairly
- motivating participants to perform well
- improving the ecological validity of the experiment

and the task characteristics that affect base pay include:

- task complexity
- required expertise

(Note: There can be more goals and task characteristics specific to a study that can be relevant to incentive design. Researchers must make sure to incorporate them into their considerations.)



Keeping the identified goals and characteristics in mind, here are some *suggestions* to help calibrate the base pay amount:

- **Meet the minimum wage standard**: Ensure pay meets ethical and legal standards [43].
- **Consider platform recommended rates**: If the platform chosen to conduct the study on recommends a pay rate, it can be considered as an anchoring point.
  *Reasoning*: If we pay lower than the recommended rate, it is possible that participants will be less motivated to pick up or put in sufficient effort into the task. This is because there can be an anchoring effect [31, 120] on participants' perceptions of fair pay since many (possibly similar) studies are likely to be paying those rates.
- **Consider past pricing of similar studies**: If the experiment design or task is common or similar to studies have been conducted in the past, consider evaluating and possibly aligning with the pricing schemes of those studies.
  *Reasoning*: If participants have performed similar tasks for a higher pay amount (or a seemingly more favourable pay structure), they can be less motivated to perform well on the task, again, due to an anchoring effect on their perceptions of fair pay. Further, if such a practice is adopted across the research community, we can hope to possibly eventually arrive at a standard pricing for a particular kind of task.
- **Consider the complexity of the task**: Assess the complexity of the task and how much cognitive load it induces. Consider adjusting the pay for more complex and time-consuming tasks.
  *Reasoning*: For complex tasks, it is often implied that it would take longer to complete the tasks and participants can just get paid the fixed rate for the amount of time they spent. However, more complex tasks with higher cognitive loads could discourage participation due to perceived low value for time invested [121]. Thus, such tasks may warrant higher pay to attract participants and motivate them to perform well.
  Resources on assessing task complexity and cognitive load: Yang

*et al.* [122], Leppink *et al.* [123], and Klepsch, Schmitz, and Seufert [124]

- **Consider the required expertise**: If there are specific skill requirements from participants, then consider paying higher than one would for no skill requirements.
  *Reasoning*: Tasks requiring specialized skills or in-depth knowledge typically warrant a higher base pay compared to tasks that a layperson can do with minimal training [125]. This reflects the value of expertise and increases the chances of attracting qualified participants who would be motivated to perform well.

- **Leverage pilot studies**: Pilot studies can be utilized to refine the pay amount in the following ways:
  - *Gather self-reported workload and perceptions of fair pay*: Researchers can include questions in the post-task or exit surveys to see if participants felt the offered pay was fair for the workload involved. This can provide a starting point for gauging participant satisfaction with the pay amount.
    In addition to yes/no questions, researchers may also consider including open-ended questions such as, "What aspects of the task felt most demanding?" or "How could the compensation be adjusted to better reflect the effort required?" These can reveal areas where participants perceive a disconnect between workload and pay and can further help identify specific characteristics of the task that might influence participant satisfaction and performance.
  - *Assess cognitive load*: Use or adapt questionnaires like those presented by Leppink *et al.* [123] and Klepsch, Schmitz, and Seufert [124] to measure perceived cognitive load during the task. This can help gauge if the pay offered seems fair, relative to the mental effort required.
  - *Offer different pay rates*: Consider offering different base pay rates to participants while keeping other aspects of the task constant. Track their task completion time and satisfaction at each pay level. This can help identify an appropriate pay rate where participant satisfaction balances out with the expected completion time and performance.

  *Reasoning*: Conducting pilot studies is becoming an increasingly encouraged practice in human-AI research as they provide an opportunity for researchers to identify potential issues with their task design and can help calibrate task parameters for eventually conducting the experiment at a larger scale [126, 127]. Pilot studies have often been used to estimate the time taken to complete a

task in previous studies. This practice can be easily extended to include considerations such as those mentioned.

### 3.3.3. DESIGNING A BONUS STRUCTURE

*Can offering bonuses help reach the identified goals?*



**Possible identified goals**: Improving ecological validity (motivating performance, creating stakes, etc.) and encouraging certain behaviours.

**Possible affecting task characteristic**: Risk.

**Suggestion**: Offer bonuses.
*Reasoning*: While a fair base pay is essential, it might not fully capture the nuances of task complexity or the desired level of performance. Bonuses provide a way to incentivize crowdworkers and encourage specific behaviours, such as performance, that go beyond simply completing the task. Specially in *high-risk* scenarios, where even a single mistake can have significant consequences, bonuses can introduce a sense of raised stakes for the participants as they are often primarily driven by monetary goals [128]. Bonuses can thus incentivize participants to pay closer attention to details and critically evaluate each piece of information before making a decision.

At the same time, we acknowledge that bonuses require nuanced consideration. A strategy such as simply offering large bonuses, may not directly translate to higher quality work [31, 129]. It is thus important to strike a balance between the compensation amount and its distribution, while ensuring alignment with the specific research goals. Hence, we suggest that researchers should critically evaluate the behaviours they aim to incentivize when designing bonus structures. Bearing this in mind, the sub-items below discuss the different aspects of a bonus scheme.

#### COMING UP WITH A BONUS AMOUNT

*How much total amount should be allocated for bonuses?*

**Suggestions**:

- **Consider past pricing of similar studies**: If the experiment design or task is common or similar to studies that have been conducted in the past, assess the bonus payout of those studies.
  *Reasoning*: If participants have performed similar tasks for seemingly more favourable reward structure, they can be less motivated to perform well on a the task, due to an anchoring effect [120] on their perceptions of fair rewards.

- **Leverage pilot studies**: Pilot studies can be utilized to refine the bonus amount as well, similar to as described for base pay, by including questions about participant satisfaction and perceptions of fairness regarding the additional rewards. They can also be used for experimenting with different bonus amounts.

#### DECIDING THE TYPE OF BONUS(ES)

*What type of bonus (performance-based, completion-based, or randomized) is the most suitable for reaching the identified goal?*



**Suggestions**:

- **Randomized bonuses: Use with caution**.
  *Reasoning*: Bonuses that rely on chance, such as "x participants will be randomly picked from the top y% performing participants to be paid a bonus...", can possibly discourage high-performing participants from putting in their best efforts, as they may feel like they are not assured a reward even after performing well. At the same time, it has been shown that they can encourage low-performing participants to engage more deeply. This trade-off should be considered.

- **Completion based bonuses: Limited use**.
  *Reasoning*: Completion-based bonuses can be helpful for encouraging participation in optional tasks (such as surveys), where the main goal is to gather a sufficient amount of data.

- **Performance-based bonuses: Recommended**.
  *Reasoning*: Performance-based bonuses are the most commonly used to motivate desired behaviours. By clearly defining and communicating how participants' performance is rewarded, they can be encouraged to focus on those desired aspects to maximize their bonus potential. The effectiveness of performance-based bonuses relies on appropriately choosing the right performance-evaluation metrics as well as the policy for mapping performance to rewards. These points are addressed below.

#### DECIDING THE PERFORMANCE EVALUATION METRICS

*How to evaluate participant performance? What is the behaviour that we wish to reward?*

**Possible affecting task characteristic**: Decision subjectivity.

**Suggestions**:

- **Consider carefully what behaviour to incentivize**: Performance could mean accuracy or speed, but ensure that the chosen metric truly reflects task goals and does not encourage participants to act in a biased way.
  Performance could also mean more than simply a task metric, it could mean engaging in any kind of desired behaviour. Consider rewarding for such desired behaviours while evaluating performance as well. For example, if explaining reasoning behind decisions is important, offer bonuses for in-depth explanations. If reading the provided information carefully is deemed important, incentivize attention or comprehension checks.

- **Assess the subjectivity of the decision**: If decision subjectivity is high, consider using open-ended questions to assess performance.

- **Leverage pilot studies**: Pilot studies can again be leveraged to gauge participants' perceptions regarding whether they felt their performance was evaluated appropriately or asking them what did the reward structure encourage them to prioritize. This can help assess if the performance metric has the intended effect of rewarding for desired behaviours.

#### DECIDING THE POLICY FOR MAPPING REWARDS TO PERFORMANCE

*How to map performance to rewards?*

**Suggestions**:

- **Clearly define the policy for bonus calculation based on performance evaluation metrics**: There can be several ways to map performance to pay, such as: rewarding over a threshold performance or increasingly rewarding with better performance. Identify what works best for the task considering the goals. We also suggest delving deeper into incentive design research to identify possibly suitable methods of designing optimal policies.

- **Consider the use of penalties with caution**: Penalties can be introduced to create high stakes for high-risk scenarios to improve ecological validity by appealing to the loss averse tendencies of participants [130]. However, be careful to avoid excessive penalties that might discourage participation.

- **Leverage pilot studies**: Pilot studies can again be leveraged to see how participants responded to the policy with which they were rewarded based on their performance.

### 3.3.4. GATHERING PARTICIPANT FEEDBACK

*What do we wish to understand from the participants' perspective? Can we augment the experiment pipeline to gather such feedback?*

Post-task or exit surveys can be useful tools for understanding participant experiences and motivations.
**Suggestion:** The survey can include similar questions about perceived fairness and motivation that were suggested for pilot studies.
*Reasoning*: This practice can allow researchers to compare responses and assess if the pilot study findings regarding participant satisfaction hold true in the main experiment. For example, if the results show a less than expected accuracy rate, survey responses about perceived unfair pay might indicate that participants were less motivated to put in their best effort. This information can be crucial for interpreting the results as well as their generalizability.

Ultimately, this practice can allow us to assess the effectiveness of the chosen incentive scheme. By asking participants questions such as whether they felt the pay was fair and if the bonus structure motivated them, researchers can gauge whether the incentive scheme achieved its goals, such as attracting qualified participants or encouraging desired behaviours. Further, it can aid the next step of reflecting on the design implications by providing the participants' perspective. Such practices could help refine incentive schemes for future research. For example, if participants report that they felt inadequately compensated or unmotivated by the bonus structure, it could be adjusted to better meet participants' expectations for future studies.

### 3.3.5. REFLECTING ON DESIGN IMPLICATIONS

*Did the design achieve the desired goals? Were there any unintended consequences?*

So far, we have suggested that researchers should engage in *intentional design*. We encouraged researchers to understand trade-offs, justify their decisions, and gather feedback from participants, in order to eventually be able to reflect on the implications of their design.

**Suggestion**: To bring the design process to fruition, we encourage researchers to also reflect on the potential effects of their design on their research outcomes.
*Reasoning*: Such a reflection can lead to more robust and reliable research. By carefully reflecting on the potential effects of their decisions on their results, researchers can be more confident in their findings and provide valuable insights into the field. Additionally, it fosters a culture of transparency and accountability, allowing for making

improvements upon existing methods for future research.

## 3.4. APPLYING THE CHECKLIST: CASE STUDIES

In this section, we present two case studies on human-AI decision-making experiments that have been conducted in prior literature to demonstrate how the checklist could be applied in actuality.



### 3.4.1. CASE STUDY I

For the first case study, we have picked the research paper: "How do you feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection" by Lammerts *et al.* [55]. It mentions a simple incentive scheme:

> "Every participant is paid an hourly wage of 9 GBP, exceeding the UK minimum wage at the time of the study."

After taking a closer look at the experiment design and study goals as described in the research article, we present our take on designing the incentive scheme for this study through application of the Incentive-Tuning Checklist.

**Step 1: Identifying the purpose**: We identify that the authors wish for the participants to thoroughly process information and focus on harm evaluation. Additionally, we surmise that the authors have the goals of providing fair compensation, enhancing ecological validity, and simulating real-world stakes.

**Step 2: Coming up with a base pay**: The authors' chosen base pay of £9 per hour exceeds the minimum wage, addressing *fair compensation*. In trying to explore *past-pricing* for studies in the toxicity classification domain, we did not find many relevant articles addressing the similar tasks. Since *no specific skills are required*, the base pay seems appropriate. However, considering the *high number of tasks* (40) per participant, we must consider strategies for maintaining participant motivation throughout and compensating them for the time invested. A simple strategy would be increasing the base pay itself. We note that a *pilot study* was conducted by the authors. We recommend enhancing it to gauge engagement and satisfaction by including questions such as, "Did you feel the need to take any breaks while performing the tasks? If so, how many and for how long?" and "Did you feel the pay was fair compensation for the time and effort required?"

**Step 3: Designing the bonus structure**: We note that the authors did not offer bonuses. Offering bonuses can aid the goal

of improving ecological validity, simulating stakes and encouraging desired behaviours in participants. We recommend that offering *small performance-based bonuses* could be effective.

The authors chose not to incentivize correct answers and indicated that their primary goal is not to encourage accuracy, but evaluating user perceptions of value. They mentioned that they included "lengthy descriptions" of the task instead of rewards to direct participants' focus towards evaluating harm, and included attention checks to filter out inattentive participants.



In order to better pursue this goal, we suggest including *comprehension checks* that focus on processing task information and consider accuracy on them as the evaluation metric for performance-based bonuses. Rewarding participants who pass these checks with a flat bonus could encourage attentiveness towards the "lengthy descriptions" and work within potential budget constraints. Thus, the *performance metric* in this case would be comprehension check accuracy and the *reward mapping policy* would be 100% reward on 100% accuracy.

Further, we note that the study deals with a high-risk scenario (hate speech classification). As the authors highlight the importance of users understanding the consequences of incorrect decisions, levying small penalties emerges as an option. However, to combat participants potentially getting discouraged, there should be sufficient positive rewards as well. But we identified that encouraging task accuracy is not a goal, hence decided not to have task accuracy-based rewards. Thus, we conclude not to use penalties.

**Step 4: Gathering participant feedback**: Especially since we did not take any explicit steps to *create high stakes* by introducing consequences (such as through penalties), we recommend using post-task surveys to get a better understanding of whether participants understood the consequences and stakes of the decision-making scenario. This is a crucial exercise, specially if budget constraints limit implementing all suggestions. Understanding participant perspectives would help in interpreting the results.

**Step 5: Reflecting on design implications**: The current design prioritizes fairness but lacks sufficient strategies to maintain participant motivation throughout the lengthy task list. Additionally, the absence of consequences might lead to underestimating the importance of evaluating harm. While we can't truly know the outcomes of the study in this regard unless it is replicated, potential inconsistencies in how participants evaluate harm could be attributed to such factors.

### 3.4.2. CASE STUDY II

For the second case study, we chose: "Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust

and Reliance in Human-AI Decision-Making" by Salimzadeh, He, and Gadiraju [131]. This describes the incentive scheme in more detail than the previous one:

> "All participants were compensated at the fixed rate of 8 GBP per hour regardless of their performance in the study. Additionally, participants received bonus rewards amounting to 0.2 GBP for each accurate response they provided during the study period. Overall, participants earned an average of 8.44 GBP per hour, well over the wage considered to be 'good' and recommended by the Prolific platform."



**Step 1: Identifying the purpose**: We identify the purpose as ensuring fair compensation for participants, enhancing ecological validity of the experiment, encouraging accurate decision-making in trip-planning tasks, and simulating real-world decision-making.

**Step 2: Coming up with a base pay**: The chosen base pay meets *platform recommendations* and there are not enough *prior studies* using the same task domain or structure for additional insight.

We note that the authors manipulate the complexity of tasks by adjusting the number of features and constraints presented to participants. Considering the complexity levels (low, medium, high) and the cognitive load induced by the tasks, we consider adjusting the base pay based on *task complexity*. Possible calibrated base pay could be the same as described by the authors for low complexity tasks, slightly higher for medium complexity tasks, and significantly higher for high complexity tasks. Appropriately incentivizing participants and addressing the varying task complexity levels through incentives can help the researchers reflect on the effect of participant motivation on performance while analyzing their results. For example, the authors make an observation regarding the decline in performance for medium/high complexity tasks. This could also be attributed to lower participant motivation due to increased cognitive effort. If researchers account for the additional cognitive effort while incentivizing participants, they can assert that they took intentional measures to combat this and thus they can attribute the decline in performance to other factors (such as uncertainty or complexity itself) with more confidence, hence improving the ecological validity of their findings.

At the same time, we notice a *trade-off* in implementing such a pay scheme. Paying participants differently based on the tasks they receive can be perceived as unfair. Some might argue that all participants deserve equal pay for their time, regardless of task complexity.

The authors did not mention a pilot study. A pilot study with targeted feedback questions could be used to validate the intuition behind the

different approaches, gauging participant motivation and perceived fairness with different pay structures. A pilot could also help assess the actual time and effort required for different complexity levels. This can help in objectively calibrating the pay scale. Further, we could include cognitive effort questionnaires as well as open-ended questions such as "Were you motivated to put in genuine effort to perform the task?" and "Was the reward fair and satisfactory as per your expectations?" Based on the responses, the amount and structure can be further tweaked.

**Step 3: Designing the bonus structure**: The authors used *performance-based bonuses*, with the evaluation metric as *accuracy* to encourage accurate trip-planning, addressing the *study goal* of encouraging correct decision-making. Also, since *decision subjectivity* is low, accuracy could be a suitable metric to consider. However, we note that the AI assisting the participant in decision-making is known to have 66.7% accuracy. In this case, it is possible that participants might simply rely on the AI to get a 2/3 chance of a reward. This would undermine the study's goal of measuring genuine human decision-making.

An option to combat this could be rewarding for *accuracy when the AI is wrong*, incentivizing genuine effort and discouraging blind reliance on the AI. This could re-focus participants on the goal of accurate trip-planning instead of maximizing their rewards.

Another option would be to explore rewarding for "accuracy-wid" (final correct decision with initial disagreement with AI) or appropriate reliance. While directly rewarding appropriate reliance might influence behaviour, it may also lead to more confident interpretations of results.

Ultimately, we have a *trade-off* between two choices for the *performance evaluation metric*. When making a decision, we acknowledge the implications of each choice:

1. *Encourage appropriate reliance, use accuracy-wid or reward higher bonuses for correct answers when the AI is wrong*: The implication of this choice would be that we are encouraging what we wish to measure, as the researchers' goal is to measure appropriate reliance itself.

2. *Encourage overall accuracy without influencing reliance*: Here, we need to acknowledge the possibility for overreliance on the AI.

From further delving into the research article, we deem that researchers wish to measure appropriate reliance *in the wild*. Thus, we conclude that *we should not encourage it by rewarding it*. However, rewarding directly for accuracy when the AI has 66.7% accuracy itself, can also mean *encouraging overreliance*. Hence, based on this discussion, we recommend rewarding for accuracy when the AI is incorrect. This addresses the issue of overreliance when simply accuracy is used, and at the same, focuses on encouraging the behaviour of a trip-planner in the real-world whose goal would be to plan an accurate

trip regardless of the AI's decision, thus still allowing us to measure appropriate reliance in the wild.

As for *mapping rewards to performance*, increasing rewards with increasing performance could be a strategy to further motivate participants to make correct decisions, if the budget permits.

Lastly, we observe that the perceived risk is relatively low, so we do not consider the use of negative rewards like penalties.

**Step 4: Gathering participant feedback**: We recommend including questions about participant motivation, cognitive load, and perceptions of pay fairness in the post-task survey as they remain crucial for understanding participant perspectives and interpreting the results.

**Step 5: Reflecting on design implications**: The current incentive scheme prioritizes fairness and encourages accurate decision-making. However, the potential for overreliance on AI can have implications on the ecological validity and must be acknowledged explicitly. We discussed the implications when choosing each of the different evaluation metrics in detail in Step 3. Such discussions should be included when presenting the results of the study.

NOTE

*We acknowledge that our recommendations and final decisions might not present the sole appropriate solution. The primary goal of presenting the case studies is to highlight that through the Incentive-Tuning Checklist we can stimulate a broader discussion on each aspect of the incentive scheme. We aimed to demonstrate how researchers can carefully consider trade-offs and justify their choices when making decisions, guided by the checklist.*

## 3.5. DISCUSSION

It is evident from the detailed discussions accompanying the items of the checklist and its application to the case studies that tuning an incentive scheme is a multifaceted endeavour. This section discusses the key points that researchers must bear in mind throughout this process.

*Stay focused on the purpose.* Always keep the study's goals at the forefront of the design. The ideal incentive scheme should encourage the desired participant behaviours necessary to achieve those goals.

*Navigate trade-offs and make informed decisions.* Recognize that designing incentive schemes often involves trade-offs. For example, in a high-risk task, we might consider implementing penalties to create stakes and encourage careful participation. However, this could also discourage some potential participants. We implore researchers to identify such kinds of trade-offs and prioritize them based on their specific study goals.

*Justify decisions and document them.* Once the trade-offs are evaluated and a decision is made, understand the implications it might have on participant behaviour and overall study results. Further, transparency is key. Throughout the design process, document the choices and the rationale behind them. This documentation serves as a valuable record for future reference and can facilitate improvements in future studies. The application of the checklist itself can serve as this documented record. We further explore this in the next chapter.

*Seek participant feedback.* Don't underestimate the power of explicit feedback. Gather insights from participants about the chosen incentive scheme. Their perspectives can shed light on potential issues or areas for improvement we might have overlooked.

*Embrace reflection.* Reflection is an ongoing process. Take time to reflect on the impact of the chosen incentive scheme on the results. Consider how it might have influenced participant motivation, participant behaviour, and ultimately, the validity and applicability of the findings.



## 3.6. LIMITATIONS AND FUTURE DIRECTIONS

Here, we identify some limitations of the proposed checklist and directions for future improvements.

### 3.6.1. NON-EXHAUSTIVENESS

The research landscape of human-AI decision-making as well as incentive design in crowdsourcing is constantly evolving, and new challenges or unforeseen scenarios might arise. New types of tasks and research questions might emerge, requiring considerations that haven't been captured in the checklist yet. Further, unexpected complexities in the research design, participant behaviour, or platform functionalities might necessitate adapting or going beyond the suggestions we offered.

Acknowledging this inherent limitation, we suggest that researchers should treat the checklist as a guide, not a rigid formula. Researchers should be prepared to adapt and refine the suggestions based on the scenarios that may arise. Further, we believe that the checklist itself should be a living document. As researchers gain experience using it and encounter new challenges, the checklist should be iteratively improved to incorporate best practices and address emerging issues.

### 3.6.2. BARRIERS TO ADOPTION

It is possible that some researchers might perceive the checklist as overly complex or time-consuming to complete, especially for smaller studies with limited resources. Moreover, integrating the checklist and

our recommendations into research workflows might require adjustments and could be met with initial resistance. We suspect this would largely be because researchers might not be convinced of its practical or long-term benefits. To mitigate this, we propose conducting empirical studies that investigate the impact of using the checklist on research outcomes. Two ways of doing so could be:



1. *Retrospective Analysis*: Re-execute existing studies after applying the checklist to design new incentive schemes. We could compare the research outcomes (e.g., data quality, participant performance, participant satisfaction) between the original and the re-runned studies to quantify any potential improvements.

2. *Controlled Experiments*: Conduct new studies with control and experimental conditions. In the control and experimental condition, we could design incentive schemes without and with the application of the checklist, respectively. Then we could measure and compare research outcomes between both conditions.

While the design of such experiments would need to be refined, the resulting data can provide crucial evidence for the checklist's value proposition. This could potentially enable researchers to be more confident in its benefits and more likely to adopt it in their own work.

### 3.6.3. INDIVIDUAL DIFFERENCES

The Incentive-Tuning Checklist offers a structured, systematic approach for designing incentive schemes while allowing flexibility to address different types of studies. However, it's important to acknowledge the potential impact of individual differences among the participants of a study. There's no single "right" incentive scheme that will universally motivate every participant in the same way.

While the checklist focuses on a systematic, step-by-step approach, researchers should also be aware of alternative approaches that cater to individual differences. These approaches include dynamic pricing or adaptive incentives [132, 133]. These methods involve adjusting incentive structures based on factors such as participant skill level, performance history, or even real-time task complexity. They utilize algorithms to continuously adapt incentive structures during the experiment, based on participant behaviour and engagement levels. This allows for a more personalized approach to incentivization. While such approaches hold promise, their adoption is still limited due to factors such as increased design complexity and the need for advanced data analytics capabilities.

In the future, the checklist could be enhanced to incorporate considerations for individual differences. The checklist's framework could potentially be expanded to include steps that guide researchers in

exploring dynamic pricing or even developing basic adaptive incentive structures within the constraints of their specific study design.

### 3.6.4. BIASES

It's important to acknowledge potential cognitive biases that can influence the design process the Incentive-Tuning Checklist outlines. One such bias is the self-interest bias [134]. This bias highlights that participants naturally prioritize their own monetary goals and may be inclined to behave or perform in ways that maximize their rewards.

We do try to address this bias by highlighting the need for caution when designing performance-based bonuses. Unrestricted bonuses can inadvertently exacerbate self-interest bias. However, we don't dismiss bonuses or rewarding for performance. Instead, we acknowledge that self-interest bias can be leveraged to motivate participants. Carefully aligning reward structures with desired behaviours can allow researchers to utilize the self-interest bias to encourage participants to exert effort, focus on desired metrics, and contribute high-quality work. However, there is always the possibility that rewards cause participants to prioritize maximizing their compensation over providing high-quality data. We encourage researchers to integrate quality control measures to mitigate this risk and ensure the validity of their findings.

By remaining vigilant about such potential biases, researchers can design incentive schemes that channel them into a positive force or mitigate them to enhance the overall quality and validity of their results.

## 3.7. ETHICAL CONSIDERATIONS

The Incentive-Tuning Checklist prioritizes the design of incentive schemes that are both effective and ethical. However, it's crucial to acknowledge that some of our suggestions touch upon ethical considerations that researchers should carefully navigate.

One such ethical concern is exploitation. We highlight the importance of ethical wages within our suggestions. We further assert that researchers have a fundamental ethical obligation to ensure fair compensation for participants in their studies.

Further, we discourage the use of excessively high bonuses or punitive penalties. Such practices can pressure participants to prioritize speed or quantity over authenticity and well-being. Researchers should strive to design incentive schemes that offer fair compensation while respecting participant autonomy.

As the research field of human-AI collaboration and decision-making evolves, so too will the ethical considerations surrounding incentive design. The Incentive-Tuning Checklist serves as a foundation, but researchers should stay informed about emerging ethical discussions

in this area. For instance, the potential for individual differences in response to incentives raises new ethical questions about fairness and equity.

By carefully considering these ethical implications, researchers can utilize the checklist to design incentive schemes that are not only effective in motivating participants but also uphold ethical principles and contribute to responsible research practices.

# 4

# DOCUMENTING THE DESIGN OF INCENTIVE SCHEMES: TEMPLATE AND REPOSITORY

"*As for the future, your task is not to foresee it, but to enable it.*"

Antoine de Saint Exupery

*This chapter outlines a clear method for reporting and documenting incentive schemes for human-AI decision-making studies. We draw upon the Incentive-Tuning Checklist and the results of the thematic analysis to highlight items for reporting, aiming to ensure transparency and capturing of the rationales behind incentive scheme design. We further emphasize the importance of researchers documenting any reflections regarding the impact of their incentive design choices.*

*A template is presented to standardize the reporting of incentive schemes within research articles, potentially facilitating easier comparisons across studies. Additionally, a public GitHub repository is established to serve as a central hub for documenting and exploring incentive schemes employed for studies in published research. Finally, we encourage open collaboration by inviting researchers to contribute their own incentive schemes in order to help the repository grow and benefit the research community.*

## 4.1. TOWARDS RQ3

A critical gap in the existing research landscape of human-AI decision-making research is the apparent lack of proper documentation of the incentive schemes that researchers employ for conducting empirical studies. This significantly limits our understanding of the process and its nuances. Furthermore, it poses limitations for future researchers to replicate findings or build upon existing knowledge due to the absence of clear and detailed explanations.

To address this, we formulated our third research question:

> *RQ3: How can the design of incentive schemes be documented through a standardized process to facilitate future research in human-AI decision-making?*



The insights from the thematic analysis in Chapter 2 and the development of the Incentive-Tuning Checklist in Chapter 3 further fortified the need for a standardized solution that allows researchers to effectively and comprehensively report and document the incentive schemes that they design.

### 4.1.1. REPORTING ON CHECKLIST ITEMS

The application of the Incentive-Tuning Checklist can itself serve as the documentation of the incentive design process. Here, we revisit the items of the checklist from the perspective of reporting the incentive design process and the final incentive scheme.

#### 1. PURPOSE

Researchers must clearly identify and state the purpose(s) for which they wish to employ an incentive scheme.

#### 2. BASE PAY

Researchers should explicitly specify the base pay offered to participants for completing the task. They must also explain the rationale behind the chosen pay amount (e.g., platform recommendations, pilot study data, past pricing).

#### 3. BONUS STRUCTURE

Researchers should indicate whether bonuses were used or not, justifying their choice. Researchers should further indicate which type of bonuses were used (performance-based, completion-based, randomized) and why. In case of performance-based bonuses, researchers must describe

the metrics used to evaluate participant performance (e.g., accuracy or task completion time) as well as how performance was translated into pay (e.g., use of penalties, reward mapping policies).

For each aspect, researchers should briefly summarize the key design decision and how it aligns with the identified purpose,

#### 4. FEEDBACK

Researcher should describe the methods used to gather participant feedback (e.g. the specific questions asked within surveys). Additionally, they should highlight any significant or recurring themes emerging from the feedback.



#### 5. REFLECTION

Researchers should go beyond simply reporting the design choices. They should also report any reflection associated with the implications of the design choices on their research findings.

### 4.1.2. CHALLENGES TO REPORTING

While the Incentive-Tuning Checklist provides a systematic and comprehensive overview of items researchers must report, some ambiguities remain regarding the format and data presentation for reporting each element. For example, in Section 2.4.1, we noted inconsistencies in how base pay is reported. Examples include using "fixed amount" vs. "pay rate," and mentioning or omitting the resulting total pay. Another area of ambiguity identified is the use of averages vs. medians to report resultant compensation. While both offer valid summaries of data, a consistent approach across studies is vital.

Inconsistent reporting hinders transparency and replicability. If researchers lack a clear framework for capturing their incentive design choices, it becomes difficult to fully understand the context and potential influences of incentive design. This can hinder the ability to replicate findings and build upon existing knowledge.

## 4.2. A TEMPLATE

A standardized reporting *template*, built upon the foundation provided by the Incentive-Tuning Checklist, can potentially address the challenges identified in the previous section. By providing clear guidelines on reporting and data presentation, the template can promote clarity, consistency, and transparency in reporting the incentive design process and decisions.

The following template can be used to capture details regarding the incentive scheme in the experimental design section of a research article:

> The purpose of our incentive design was to ensure [identify primary goals]. Participants received a resultant pay of [average and median resultant pay amount] based on a base pay rate of [base pay per hour] and [average and median bonus payout].
>
> The base pay was set as [amount] for completing the task to ensure fair compensation, considering [rationale, e.g., platform recommendations / minimum wage / past pricing / pilot study feedback / specific task characteristic]. [Can elaborate further as per choice].
>
> To further [incentivize / ensure / motivate] the [identified goals] a [performance]-based bonus structure was implemented. [Decisions due to specific task characteristics (e.g. use of penalties because of high perceived risk]. Maximum bonus payout was set as [amount]. Participant performance was evaluated based on [evaluation metrics (e.g. accuracy)], calculated as [reward mapping policy]. [Can elaborate further as per choice].
>
> [Optional] Participants received additional bonuses for [task specific consdirations] to encourage [desired behaviours].
>
> [Optional] More details, survey feedback etc.

*Note: It is important to acknowledge that the suggested phrases within the template are not intended to be rigidly adhered to. Their primary function is to illustrate the recommended structure for reporting incentive design decisions. Researchers can adapt these phrases to fit the specific context of their study while maintaining overall clarity and consistency in reporting.*

We notice that the template only addresses the first three items of the checklist. Given the highly context-dependent nature of the descriptions and discussions pertaining to the remaining items, their completion is left to the discretion of the researchers. To this end, we make the following recommendations:

The study design should explicitly describe the methods employed to collect feedback during the experiment (e.g., surveys, open-ended questions, interviews). In the results or observations section, researchers should present the obtained feedback data.

The discussion section should include reflections on the design choices and feedback. This may encompass insights gained from the feedback data, its influence on the overall research findings, and potential areas for improvement in future studies.

## 4.3. A REPOSITORY

The dataset that was built for conducting the literature review, mentioned in Section 2.2.3, is a significant corpus capturing incentive design in existing literature. We identified that by collating and presenting relevant items from within this dataset in a public repository, we could create a valuable resource for future researchers. Hence, we established a public repository on GitHub[1] to promote transparency and collaboration within incentive design for human-AI decision-making studies.

> The repository can be accessed here to view the source code or raise pull requests: GitHub Repository.

> The tabulated incentive scheme data from published articles compiled so far can directly be viewed here: GitHub Pages.

This repository is aimed to serve as a central hub for researchers to share and access incentive design knowledge from past published research. It is open for public access, currently being actively populated with the design decisions extracted from a review of the existing literature conducted in Chapter 2.

### 4.3.1. SOURCE DATA

Each incentive scheme within the repository shall be documented using a standardized JSON[2] file format. This format is used to ensure consistency and allow for easy data presentation, extraction, and analysis. The specifications are captured here.

### 4.3.2. OPEN COLLABORATION

The repository will be open for contributions. Researchers are encouraged to submit pull requests to share the incentive schemes they design by applying the Incentive-Tuning Checklist for their studies in the future and contribute to this valuable resource for the research community. The repository shall be maintained and pull requests shall be reviewed and merged periodically.

[1] www.github.com
[2] www.json.org

## 4.4. DISCUSSION AND LIMITATIONS

By providing a clear outline for reporting structure, format, and data presentation, we aspire that the template promotes clarity, consistency, and transparency in the documentation of incentive schemes and incentive design practices. This, in turn, can strengthen the overall research ecosystem by fostering replicability, facilitating knowledge sharing, and enabling researchers to build upon one another's work. Further, the repository has the potential to become a game-changer for the field. By allowing open access and facilitating standardized documentation, it can accelerate advancements in incentive design and ultimately strengthen the quality and validity of human-AI decision-making research.



We suspect that the primary limitation of these solutions would be adoption. As mentioned for the Incentive-Tuning Checklist in Section 3.6.2, researchers may be hesitant to integrate these tools into their research pipelines due to concerns regarding the time investment associated with creating detailed documentation.

Nevertheless, we attempt to populate the repository with relevant information regarding incentive schemes described in past literature ourselves. This proactive effort ensures that the repository possesses value from its inception, potentially serving as a springboard for researchers to build upon existing knowledge. While researchers may initially be wary of investing time in documentation, we hope that by encountering valuable and readily accessible information within the repository during their design process, they may subsequently see the value in documenting and contributing their own data.

We eventually hope to create an online community around the repository where researchers can share their designs and rationales, ask questions, and provide feedback to one another, thereby fostering a collaborative knowledge base.

# 5

# CONCLUSION

This concluding chapter summarizes the key findings of this thesis on incentive design for human-AI decision-making studies and explores their broader implications.

## 5.1. SUMMARY OF RESEARCH OUTCOMES

We set out to address the challenge of understanding, designing, and documenting incentive schemes for human-AI decision-making studies. In Section 1.4, we detailed three research questions to guide our scientific pursuit. Here, we revisit them one-by-one and discuss how our research findings contribute towards the fulfillment of our research objectives.

> *RQ1: How are incentive schemes currently designed for conducting empirical human-AI decision making studies?*

By means of a thematic analysis of existing literature, we shed light onto the current landscape of incentive design. For instance, we identified that incentive schemes have *components*, varying subsets of which are used by researchers in formulating their incentive schemes. The primary components of incentive schemes were identified as *base pay* and *bonuses*. For these components, different trends were also observed in operationalizing them, such as heuristic determination of pay amounts. For bonuses, several sub-components were also identified, including the amount and type of bonus. For performance-based bonuses, performance evaluation metrics and reward mapping policies were also identified as sub-components. Similarly, several trends regarding the manipulation, impact, and communication of incentive schemes were also identified. Reflecting on such observations raised several questions regarding the design of incentive schemes, setting up future work.

*RQ2: How should incentive schemes be appropriately designed through a standardized process for empirical human-AI decision-making studies?*

For RQ2, we argued for a paradigm shift towards a normative, standardized approach that acknowledges the unique challenges of human-AI decision-making studies. We proposed the Incentive-Tuning Checklist, a useful tool that can guide researchers through designing incentive schemes for their studies. It provides a step-by-step process for incentive design, emphasizing on the alignment of the design with research goals, It addresses the core components of an incentive scheme, as identified through our thematic analysis. We further offered suggestions to help guide researchers towards determining what works best for their study, encouraging them to consider different trade-offs and justify their design choices.



*RQ3: How can the design of incentive schemes be documented through a standardized process to facilitate future research in human-AI decision-making?*

Following the design of incentive schemes, we embarked towards addressing the documentation of incentive schemes. We identified that it is important for researchers to adequately report their chosen incentive schemes in order to aid future efforts within the field of study. To facilitate proper documentation and address RQ3, we generated a template based on the items of the Incentive-Tuning Checklist. The template provides a standardized way for researchers to report their incentive scheme design in their research papers. Further, we created a collaborative public GitHub repository to facilitate easy access to the detailed incentive design decisions for studies within published articles.

## 5.2. IMPLICATIONS

Firstly, the thematic analysis provided valuable insights into the various incentive design approaches employed by researchers and the diverse discussions surrounding the role and impact of incentives. Our exploration also revealed unanswered questions that demand further investigation. We promptly followed up on some of them ourselves. Thus, our first set of outcomes had direct implications for the second step of our work. The resultant themes and insights informed the construction of the Incentive-Tuning Checklist, which addresses our second research question by providing a standardized yet flexible process for incentive

design. While some areas, like incentive manipulation and communication, remained outside this work's scope, they offer exciting directions for future research. By delving deeper into the still unanswered questions, researchers could refine the approach further to create more effective incentive schemes. Such possibilities are discussed in more detail in Section 5.3.

The overview we provide can be a *solid foundation* for researchers to delve deeper into specific aspects of incentive design. Our insights can be a powerful lens for researchers to look into a topic, and use them to kick-start or shape their own research and further contribute to the field's advancement.

Secondly, the Incentive-Tuning Checklist proposed by us offers a valuable foundation for researchers to design effective and *intentional* incentive schemes. By systematically encouraging consideration for the study's goals, incentive design aspects, participant behaviour, and potential trade-offs, the checklist empowers researchers to design incentive schemes that promote desired behaviours. It remains extensible and can in the future be updated to incorporate further advancements that might happen in the field.



The standardized approach that the checklist provides to incentive design not only has the *potential to improve data quality* but it can also *facilitate comparisons* across different research projects, fostering collaboration and accelerating scientific progress in human-AI decision-making. At the same time, the checklist acknowledges that a "one-size-fits-all" approach might not be optimal when designing incentives for every unique study. It offers guidance to researchers while allowing for customization based on specific study requirements and research questions. This, in turn, will contribute to more *robust* human-AI decision-making research, paving the way for advancements in this rapidly evolving field.

Lastly, the template we provide for documenting incentive schemes offers a standardized format, which presents numerous advantages. It ensures that researchers capture all the relevant information following the application of the Incentive-Tuning Checklist. This can make it easier for readers to understand the intent with which incentives were designed and their potential influence on participant behaviour and study findings. The public-access, online repository we set up further allows researchers to access and share detailed information about incentive schemes, including the rationales behind design decisions. The collaborative aspect of the repository offers several benefits. It facilitates *knowledge sharing* as researchers can easily find existing incentive schemes relevant to their work, making it possible to learn from each other's incentive design choices, accelerating progress in the field. It also fosters *transparency* by allowing the community to review the incentive design decisions for different studies.

Ultimately, we believe that that the implications of our findings are

far-reaching. As noted above, they contribute significantly to the field of human-AI decision-making research in several ways.

## 5.3. LIMITATIONS AND FUTURE WORK

Here, we acknowledge some limitations of the work presented in this thesis, that also opens doors for future research.

Firstly, the core assumption of this thesis lies in the understanding that incentives can influence crowdworker behaviour. While this notion is widely supported [28, 34, 35, 135], previous research also highlights instances where incentives may not have a significant effect [67]. This thesis did not delve into the specific conditions under which incentives might be less or more impactful. Therefore, the extent to which the designed incentive schemes would truly influence behaviour cannot be definitively established. Interestingly, while conducting our thematic analysis this very dilemma emerged as a sub-theme, but a deeper exploration of this dynamic fell outside the scope of this work. Future research could explore the nuances of incentive design and their potential effects on study participants under varying conditions.

*Note: We do still emphasize that incentive design remains a crucial part of any crowdsourced study. The tools developed in this thesis to aid this process are still valuable, at least in the context of good research practices. By following a rigorous process that aligns study goals with incentives and engaging in the intentional design and documentation of incentives, researchers can ensure that they conducted their due diligence.*

Secondly, our focus on monetary incentives inherently limits this work to the realm of extrinsic motivation. There is a substantial body of research that acknowledges the importance of intrinsic motivations in driving crowdworker participation as well [61, 136, 137]. Future research could investigate the interplay between intrinsic and extrinsic motivations for crowdworker behaviour, potentially revealing a more holistic picture of what drives them.

Lastly, we briefly comment on the limitations of our methodologies, already discussed in more detail in the previous chapters.

The thematic analysis was conducted by a single researcher. Ideally, multiple coders would be involved to enhance the reliability and trustworthiness of the findings [72]. However, the scope of this thesis limited the coding team to one researcher. Further, qualitative research inherently carries an element of subjectivity. The researcher's background and experiences can influence the interpretation of data. However, we employed a reflexive approach, where researchers actively acknowledge their position and how it might shape the analysis. Thus, we argue that the subjectivity is a feature of the methodology of thematic analysis, rather than a shortcoming of the research itself. However, it

is important to acknowledge that subsequent research outputs developed in this thesis, including the checklist, template, and repository, draw upon the findings of the thematic analysis and therefore inherit its limitations and subjectivity. Future effort that employs a larger coding team and incorporates triangulation with other methodologies could potentially strengthen the foundation upon which these tools are built. Further, to ensure these tools become valuable practical resources, conducting experiments and applying them on real-world research projects is an imperative next step. The checklist and template presented within this thesis can be considered prototype versions, and must be refined by iterative testing and validation.

## 5.4. CONCLUDING REMARKS

In conclusion, this thesis has woven a compelling narrative around the critical role of incentive design in unlocking the potential of human-AI collaboration in decision-making. We began by unraveling the existing research, meticulously examining the landscape of challenges and opportunities (Chapter 2). This understanding provided the foundation for crafting a useful tool - the Incentive-Tuning Checklist - which aims to empower researchers to design appropriate incentive schemes for their studies (Chapter 3). Finally, we provided valuable reporting and documentation tools, driven by the checklist and its theoretical underpinnings, ensuring accessibility and potential for future refinement (Chapter 4). In painting the picture for *understanding*, *designing*, and *documenting* incentive schemes, we have advocated for a more standardized approach to the entire incentive design process. This, in turn, can pave the way for more reliable and generalizable knowledge in the field of human-AI decision-making. Ultimately, this journey aims to empower researchers to develop effective human-AI partnerships, leveraging the strengths of both humans and machines, to achieve quality decision-making outcomes across various domains.

# BIBLIOGRAPHY


[1] Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, X. Liu, Y. Wu, F. Dong, C.-W. Qiu, J. Qiu, K. Hua, W. Su, J. Wu, H. Xu, Y. Han, C. Fu, Z. Yin, M. Liu, R. Roepman, S. Dietmann, M. Virta, F. Kengara, Z. Zhang, L. Zhang, T. Zhao, J. Dai, J. Yang, L. Lan, M. Luo, Z. Liu, T. An, B. Zhang, X. He, S. Cong, X. Liu, W. Zhang, J. P. Lewis, J. M. Tiedje, Q. Wang, Z. An, F. Wang, L. Zhang, T. Huang, C. Lu, Z. Cai, F. Wang, and J. Zhang. "Artificial intelligence: A powerful paradigm for scientific research". In: *The Innovation* 2.4 (2021), p. 100179. issn: 2666-6758. doi: https://doi.org/10.1016/j.xinn.2021.100179.

[2] E. Groff and N. La Vigne. "Forecasting the future of predictive crime mapping". In: *Crime Prevention Studies* 13 (Jan. 2002), pp. 29–58.

[3] S. Dilsizian and E. Siegel. "Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment". In: *Current cardiology reports* 16 (Jan. 2014), p. 441. doi: 10.1007/s11886-013-0441-8.

[4] A. Khandani, A. Kim, and A. Lo. "Consumer Credit-Risk Models Via Machine-Learning Algorithms". In: *Journal of Banking Finance* 34 (Nov. 2010), pp. 2767–2787. doi: 10.1016/j.jbankfin.2010.06.001.

[5] X. Wang and Y. Wang. "Improving Content-based and Hybrid Music Recommendation using Deep Learning". In: Nov. 2014, pp. 627–636. doi: 10.1145/2647868.2654940.

[6] Y. Yang, W. Qian, and H. Zou. "Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models". In: *Journal of Business Economic Statistics* 36 (Aug. 2015). doi: 10.1080/07350015.2016.1200981.

[7] Y. Duan, J. S. Edwards, and Y. K. Dwivedi. "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda". In: *International Journal of Information Management* 48 (2019), pp. 63–71. issn: 0268-4012. doi: https://doi.org/10.1016/j.ijinfomgt.2019.01.021.

[8] V. Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Jan. 2019. isbn: 978-3-030-30370-9. doi: 10.1007/978-3-030-30371-6.

[9] R. Berk. "Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement". In: *Annual Review of Criminology* 4 (Jan. 2021). doi: 10.1146/annurev-criminol-051520-012342.

[10] M. H. Jarrahi. "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making". In: *Business Horizons* 61.4 (2018), pp. 577–586. issn: 0007-6813. doi: https://doi.org/10.1016/j.bushor.2018.03.007.

[11] V. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, and C. Tan. "Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 1369–1385. isbn: 9798400701924.

[12] B. Dietvorst, J. Simmons, and C. Massey. "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them". In: *Management Science* 64 (Mar. 2018), pp. 1155–1170. doi: 10.1287/mnsc.2016.2643.

[13] M. Yin, J. Wortman Vaughan, and H. Wallach. "Understanding the Effect of Accuracy on Trust in Machine Learning Models". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. isbn: 9781450359702. doi: 10.1145/3290605.3300509.

[14] P. K. Kahr, G. Rooks, M. C. Willemsen, and C. C. Snijders. "It Seems Smart, but It Acts Stupid: Development of Trust in AI Advice in a Repeated Legal Decision-Making Task". In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. IUI '23. Sydney, NSW, Australia: Association for Computing Machinery, 2023, pp. 528–539. isbn: 9798400701061. doi: 10.1145/3581641.3584058.

[15] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–14. isbn: 9781450356206. doi: 10.1145/3173574.3173951.

[16] B. Green and Y. Chen. “Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 90–99. isbn: 9781450361255. doi: 10.1145/3287560.3287563.

[17] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy. “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 295–305. isbn: 9781450369367. doi: 10.1145/3351095.3372852.

[18] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. “Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. isbn: 9781450380966. doi: 10.1145/3411764.3445717.

[19] J. Schoeffer, N. Kuehl, and Y. Machowski. ““There Is Not Enough Information”: On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1616–1628. isbn: 9781450393522. doi: 10.1145/3531146.3533218.

[20] C. Rastogi, Y. Zhang, D. Wei, K. R. Varshney, A. Dhurandhar, and R. Tomsett. “Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making”. In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW1 (Apr. 2022). doi: 10.1145/3512930.

[21] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. “To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). doi: 10.1145/3449287.

[22] S. Narayanan, G. Yu, C.-J. Ho, and M. Yin. “How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making?” In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’23. Montréal, QC, Canada: Association for Computing Machinery, 2023, pp. 49–57. isbn: 9798400702310. doi: 10.1145/3600211.3604709.

[23] J. Dressel and H. Farid. “The accuracy, fairness, and limits of predicting recidivism”. In: *Science Advances* 4.1 (2018). doi: 10.1126/sciadv.aao5580.

[24] J. E. Mercado, M. A. Rupp, J. Y. C. Chen, M. J. Barnes, D. Barber, and K. Procci. “Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management”. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58.3 (2016), pp. 401–415. doi: 10.1177/0018720815621206.

[25] A. Springer and S. Whittaker. “Progressive disclosure: empirically motivated approaches to designing effective transparency”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19. Marina del Ray, California: Association for Computing Machinery, 2019, pp. 107–120. isbn: 9781450362726. doi: 10.1145/3301275.3302322.

[26] H.-F. Cheng, R. Wang, Z. Zhang, F. O’Connell, T. Gray, F. M. Harper, and H. Zhu. “Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. isbn: 9781450359702. doi: 10.1145/3290605.3300789.

[27] V. Lai, H. Liu, and C. Tan. “"Why is ’Chicago’ deceptive?" Towards Building Model-Driven Tutorials for Humans”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–13. isbn: 9781450367080. doi: 10.1145/3313831.3376873.

[28] A. Ghezzi, D. Gabelloni, A. Martini, and A. Natalicchio. “Crowdsourcing: A Review and Suggestions for Future Research”. In: *International Journal of Management Reviews* 20.2 (2018), pp. 343–363. doi: https://doi.org/10.1111/ijmr.12135.

[29] F. Neto and C. Saibel Santos. “Understanding crowdsourcing projects: A systematic review of tendencies, workflow, and quality management”. In: *Information Processing and Management* 54 (July 2018), pp. 490–506. doi: 10.1016/j.ipm.2018.03.006.

[30] R. Karachiwalla and F. Pinkow. “Understanding crowdsourcing projects: A review on the key design elements of a crowdsourcing initiative”. In: *Creativity and Innovation Management* 30.3 (2021), pp. 563–584. doi: 10.1111/caim.12454.

[31] W. Mason and D. J. Watts. “Financial incentives and the "performance of crowds"”. In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*. HCOMP ’09. Paris, France: Association for Computing Machinery, 2009, pp. 77–85. isbn: 9781605586724. doi: 10.1145/1600150.1600175.

[32] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh. “Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions”. In: *ACM Comput. Surv.* 51.1 (Jan. 2018). issn: 0360-0300. doi: 10.1145/3148148.

[33] C. Muldoon, M. O’Grady, and G. O’Hare. “A survey of incentive engineering for crowdsourcing”. In: *The Knowledge Engineering Review* 33 (Apr. 2018). doi: 10.1017/S0269888918000061.

[34] B. Bruno, M. Faggini, and A. Parziale. “Motivation, Incentives and Performance: An Interdisciplinary Review”. In: *International Journal of Business and Management* 12 (Nov. 2017), p. 29. doi: 10.5539/ijbm.v12n12p29.

[35] J. R. Terborg and H. E. Miller. “Motivation, behavior, and performance: A closer examination of goal setting and monetary incentives.” In: *Journal of Applied Psychology* 63.1 (1978), pp. 29–39. doi: 10.1037/0021-9010.63.1.29.

[36] M.-C. Yuen, I. King, and K.-S. Leung. “A Survey of Crowdsourcing Systems”. In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 2011, pp. 766–773. doi: 10.1109/PASSAT/SocialCom.2011.203.

[37] B. Morschheuser, J. Hamari, and J. Koivisto. “Gamification in Crowdsourcing: A Review”. In: *2016 49th Hawaii International Conference on System Sciences (HICSS)*. 2016, pp. 4375–4384. doi: 10.1109/HICSS.2016.543.

[38] L. Galli and P. Fraternali. “Achievement Systems Explained”. In: *Trends and Applications of Serious Gaming and Social Media*. Ed. by Y. Baek, R. Ko, and T. Marsh. Singapore: Springer Singapore, 2014, pp. 25–50. isbn: 978-981-4560-26-9. doi: 10.1007/978-981-4560-26-9_3.

[39] D. Karger, S. Oh, and D. Shah. “Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems”. In: *Operations Research* 62 (Oct. 2011). doi: 10.1287/opre.2013.1235.

[40] C.-J. Ho, S. Jabbari, and J. W. Vaughan. “Adaptive Task Assignment for Crowdsourced Classification”. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML’13. Atlanta, GA, USA: JMLR.org, 2013, I–534–I–542.

[41] L. Tran-Thanh, T. D. Huynh, A. Rosenfeld, S. D. Ramchurn, and N. R. Jennings. “BudgetFix: Budget Limited Crowdsourcing for Interdependent Task Allocation with Quality Guarantees”. In: *Proceedings of the 2014 International Conference on Autonomous*

*Agents and Multi-Agent Systems*. AAMAS '14. Paris, France: International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 477–484. isbn: 9781450327381.

[42] V. Williamson. "On the Ethics of Crowdsourced Research". In: *PS: Political Science  Politics* 49 (Jan. 2016), pp. 77–81. doi: 10.1017/S104909651500116X.

[43] S. Standing and C. Standing. "The ethical use of crowdsourcing". In: *Business Ethics: A European Review* 27.1 (2018), pp. 72–80. doi: https://doi.org/10.1111/beer.12173.

[44] A. Kittur, E. H. Chi, and B. Suh. "Crowdsourcing user studies with Mechanical Turk". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. Florence, Italy: Association for Computing Machinery, 2008, pp. 453–456. isbn: 9781605580111. doi: 10.1145/1357054.1357127.

[45] A. Kovashka, O. Russakovsky, L. Fei-Fei, and K. Grauman. "Crowdsourcing in Computer Vision". In: *Foundations and Trends® in Computer Graphics and Vision* 10 (Jan. 2016), pp. 177–243. doi: 10.1561/0600000071.

[46] M. Knauff and A. Wolf. "Editorial: Complex cognition: The science of human reasoning, problem-solving, and decision-making". In: *Cognitive processing* 11 (Mar. 2010), pp. 99–102. doi: 10.1007/s10339-010-0362-z.

[47] G. Paolacci, J. Chandler, and P. G. Ipeirotis. "Running experiments on Amazon Mechanical Turk". In: *Judgment and Decision Making* 5.5 (2010), pp. 411–419. doi: 10.1017/S1930297500002205.

[48] R. Kocielnik, S. Amershi, and P. N. Bennett. "Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–14. isbn: 9781450359702. doi: 10.1145/3290605.3300641.

[49] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan. "Explaining models: an empirical study of how explanations impact fairness judgment". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI '19. Marina del Ray, California: Association for Computing Machinery, 2019, pp. 275–285. isbn: 9781450362726. doi: 10.1145/3301275.3302310.

[50] S. Feng and J. Boyd-Graber. "What can AI do for me? evaluating machine learning interpretations in cooperative play". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI '19. Marina del Ray, California: Association for Computing Machinery, 2019, pp. 229–239. isbn: 9781450362726. doi: 10.1145/3301275.3302265.

[51] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman. “Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 454–464. isbn: 9781450371186.

[52] M. Chromik, M. Eiband, F. Buchner, A. Krüger, and A. Butz. “I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI”. In: *Proceedings of the 26th International Conference on Intelligent User Interfaces*. IUI ’21. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 307–317. isbn: 9781450380171. doi: 10.1145/3397481.3450644.

[53] X. Dai, M. T. Keane, L. Shalloo, E. Ruelle, and R. M. Byrne. “Counterfactual Explanations for Prediction and Diagnosis in XAI”. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’22. Oxford, United Kingdom: Association for Computing Machinery, 2022, pp. 215–226. isbn: 9781450392471. doi: 10.1145/3514094.3534144.

[54] S. Tolmeijer, M. Christen, S. Kandul, M. Kneer, and A. Bernstein. “Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. isbn: 9781450391573. doi: 10.1145/3491102.3517732.

[55] P. Lammerts, P. Lippmann, Y.-C. Hsu, F. Casati, and J. Yang. “How do you feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection”. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’23. Montréal, QC, Canada: Association for Computing Machinery, 2023, pp. 834–844. isbn: 9798400702310. doi: 10.1145/3600211.3604655.

[56] O. Biran and K. McKeown. “Human-centric justification of machine learning predictions”. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI’17. Melbourne, Australia: AAAI Press, 2017, pp. 1461–1467. isbn: 9780999241103.

[57] A. Smith-Renner, R. Fan, M. Birchfield, T. Wu, J. Boyd-Graber, D. S. Weld, and L. Findlater. “No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–13. isbn: 9781450367080. doi: 10.1145/3313831.3376624.

[58] K. Vodrahalli, R. Daneshjou, T. Gerstenberg, and J. Zou. “Do Humans Trust Advice More if it Comes from AI? An Analysis of Human-AI Interactions”. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’22. Oxford, United Kingdom: Association for Computing Machinery, 2022, pp. 763–777. isbn: 9781450392471. doi: 10.1145/3514094.3534150.

[59] V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, and C. Tan. “Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. isbn: 9781450391573. doi: 10.1145/3491102.3501999.

[60] S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, and X. Ma. “Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. isbn: 9781450394215. doi: 10.1145/3544548.3581058.

[61] H. Liang, M.-M. Wang, J.-J. Wang, and Y. Xue. “How intrinsic motivation and extrinsic incentives affect task effort in crowdsourcing contests: A mediated moderation model”. In: *Computers in Human Behavior* 81 (2018), pp. 168–176. issn: 0747-5632. doi: https://doi.org/10.1016/j.chb.2017.11.040.

[62] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna. “Explanations Can Reduce Overreliance on AI Systems During Decision-Making”. In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW1 (Apr. 2023). doi: 10.1145/3579605.

[63] N. Grgić-Hlača, C. Engel, and K. P. Gummadi. “Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). doi: 10.1145/3359280.

[64] N. Grgić-Hlača, C. Castelluccia, and K. P. Gummadi. “Taking Advice from (Dis)Similar Machines: The Impact of Human-Machine Similarity on Machine-Assisted Decision-Making”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 10.1 (2022), pp. 74–88. doi: 10.1609/hcomp.v10i1.21989.

[65] G. He, L. Kuiper, and U. Gadiraju. “Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany:

Association for Computing Machinery, 2023. isbn: 9781450394215. doi: 10.1145/3544548.3581025.

[66] Y. T.-Y. Hou, W.-Y. Lee, and M. Jung. ""Should I Follow the Human, or Follow the Robot?" — Robots in Power Can Have More Influence Than Humans on Decision-Making". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. isbn: 9781450394215. doi: 10.1145/3544548.3581066.

[67] L. S. Treiman, C.-J. Ho, and W. Kool. "Humans Forgo Reward to Instill Fairness into AI". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 11.1 (2023), pp. 152–162. doi: 10.1609/hcomp.v11i1.27556.

[68] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–14. isbn: 9781450356206. doi: 10.1145/3173574.3173951.

[69] *An Evaluation of the Human-Interpretability of Explanation*. 2018.

[70] J. M. Echterhoff, M. Yarmand, and J. McAuley. "AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. isbn: 9781450391573. doi: 10.1145/3491102.3517443.

[71] V. Braun and V. Clarke. "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. doi: 10.1191/1478088706qp063oa.

[72] V. Braun and V. Clarke. *Thematic Analysis: A Practical Guide*. First. 2021. isbn: 9781473953239.

[73] G. Guest, K. MacQueen, and E. Namey. *Applied Thematic Analysis*. SAGE Publications, Inc., 2012. doi: 10.4135/9781483384436.

[74] S. Teslo, E. Jenssen, M. Thurston, M. Mandelid, G. Resaland, A. Chalkley, and H. Tjomsland. "It's the journey, not the arrival that matters – Teachers' perceptions of their practice after participating in a continuing professional development program in physically active learning". In: *Teaching and Teacher Education* 136 (2023), p. 104377. issn: 0742-051X. doi: https://doi.org/10.1016/j.tate.2023.104377.

[75] D. Byrne. “A worked example of Braun and Clarke’s approach to reflexive thematic analysis”. In: *Quality amp; Quantity* 56.3 (2021), pp. 1391–1412. doi: 10.1007/s11135-021-01182-y.

[76] N. Z. Warner, C. Gleeson, P. Fahey, R. Horgan, and A. Groarke. “Experiences of living with Lynch Syndrome: A reflexive thematic analysis”. In: *European Journal of Oncology Nursing* 58 (2022), p. 102117. issn: 1462-3889. doi: https://doi.org/10.1016/j.ejon.2022.102117.

[77] E. Park, D. Ifenthaler, and R. B. Clariana. “Adaptive or adapted to: Sequence and reflexive thematic analysis to understand learners’ self-regulated learning in an adaptive learning analytics dashboard”. In: *British Journal of Educational Technology* 54.1 (2022), pp. 98–125. doi: 10.1111/bjet.13287.

[78] C. Herzog, C. Handke, and E. Hitters. “Analyzing Talk and Text II: Thematic Analysis”. In: *The Palgrave Handbook of Methods for Media Policy Research*. Ed. by H. Van den Bulck, M. Puppis, K. Donders, and L. Van Audenhove. Cham: Springer International Publishing, 2019, pp. 385–401. doi: 10.1007/978-3-030-16065-4_22.

[79] V. Braun and V. Clarke. “Thematic analysis.” In: *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Bio* (2012), pp. 57–71. doi: 10.1037/13620-004.

[80] V. Braun and V. Clarke. “Reflecting on reflexive thematic analysis”. In: *Qualitative Research in Sport, Exercise and Health* 11.4 (2019), pp. 589–597. doi: 10.1080/2159676X.2019.1628806.

[81] D. Pickles, L. King, and I. Belan. “Attitudes of nursing students towards caring for people with HIV/AIDS: thematic literature review”. In: *Journal of Advanced Nursing* 65.11 (2009), pp. 2262–2273. doi: https://doi.org/10.1111/j.1365-2648.2009.05128.x.

[82] Crawford, B. Brown, and Majomi. “Education as an Exit Strategy for Community Mental Health Nurses: A Thematic Analysis of Narratives”. In: *Mental Health Review* 13 (June 2008), p. 8. doi: 10.1108/13619322200800017.

[83] V. Ward, A. House, and S. Hamer. “Developing a Framework for Transferring Knowledge Into Action: A Thematic Analysis of the Literature”. In: *Journal of Health Services Research amp; Policy* 14.3 (2009), pp. 156–164. doi: 10.1258/jhsrp.2009.008120.

[84] K. Broadhurst and A. Harrington. “A Thematic Literature Review: The Importance of Providing Spiritual Care for End-of-Life Patients Who Have Experienced Transcendence Phenomena”. In: *American Journal of Hospice and Palliative Medicine®* 33.9 (2016). PMID: 26169519, pp. 881–893. doi: 10.1177/1049909115595217.

[85] C. M. Hodge and S. P. Narus. “Electronic problem lists: a thematic analysis of a systematic literature review to identify aspects critical to success”. In: *Journal of the American Medical Informatics Association* 25.5 (Mar. 2018), pp. 603–613. issn: 1527-974X. doi: 10.1093/jamia/ocy011.

[86] M. Salm, M. Ali, M. Minihane, and P. Conrad. “Defining global health: findings from a systematic review and thematic analysis of the literature”. In: *BMJ Global Health* 6.6 (2021). doi: 10.1136/bmjgh-2021-005292.

[87] F. Liñán and A. Fayolle. “A systematic literature review on entrepreneurial intentions: citation, thematic analyses, and research agenda”. In: *International Entrepreneurship and Management Journal* 11.4 (2015), pp. 907–933. doi: 10.1007/s11365-015-0356-5.

[88] D. Mallinson, G. Morcol, E. Yoo, S. Azim, E. Levine, and S. Shafi. “Sharing economy: A systematic thematic analysis of the literature”. In: *Information Polity* 25 (Mar. 2020), pp. 1–16. doi: 10.3233/IP-190190.

[89] M. Ali. “Sustainable Entrepreneurship: A Systematic Literature Review with Thematic Analysis”. In: *World Journal of Entrepreneurship Management and Sustainable Development* ahead-of-print (Jan. 2021). doi: 10.1108/WJEMSD-11-2020-0150.

[90] Z. Amin, N. M. Ali, and A. F. Smeaton. “Attention-Based Design and User Decisions on Information Sharing: A Thematic Literature Review”. In: *IEEE Access* 9 (2021), pp. 83285–83297. doi: 10.1109/ACCESS.2021.3087740.

[91] N. Cooper, T. Horne, G. R. Hayes, C. Heldreth, M. Lahav, J. Holbrook, and L. Wilcox. “A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New York, NY, USA: Association for Computing Machinery, 2022. isbn: 9781450391573. doi: 10.1145/3491102.3517716.

[92] L. V. Francisco M. Olmos-Vega Renée E. Stalmeijer and R. Kahlke. “A practical guide to reflexivity in qualitative research: AMEE Guide No. 149”. In: *Medical Teacher* 45.3 (2023). PMID: 35389310, pp. 241–251. doi: 10.1080/0142159X.2022.2057287.

[93] B. Y. Lim, A. K. Dey, and D. Avrahami. “Why and why not explanations improve the intelligibility of context-aware intelligent systems”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’09. Boston, MA, USA: Association for Computing Machinery, 2009, pp. 2119–2128. isbn: 9781605582467. doi: 10.1145/1518701.1519023.

[94] B. Dietvorst, J. Simmons, and C. Massey. “Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err”. In: *SSRN Electronic Journal* (Jan. 2014). doi: 10.2139/ssrn.2466040.

[95] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach. “Manipulating and Measuring Model Interpretability”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. isbn: 9781450380966. doi: 10.1145/3411764.3445315.

[96] O. Biran and K. McKeown. “Human-centric justification of machine learning predictions”. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI’17. Melbourne, Australia: AAAI Press, 2017, pp. 1461–1467. isbn: 9780999241103.

[97] M. Yurrita, T. Draws, A. Balayn, D. Murray-Rust, N. Tintarev, and A. Bozzon. “Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. isbn: 9781450394215. doi: 10.1145/3544548.3581161.

[98] Y. Liu, A. Mittal, D. Yang, and A. Bruckman. “Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. isbn: 9781450391573. doi: 10.1145/3491102.3517731.

[99] G. He, S. Buijsman, and U. Gadiraju. “How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System”. In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW2 (Oct. 2023). doi: 10.1145/3610067.

[100] F. Jahanbakhsh, Y. Katsis, D. Wang, L. Popa, and M. Muller. “Exploring the Use of Personalized AI for Identifying Misinformation on Social Media”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. isbn: 9781450394215. doi: 10.1145/3544548.3581219.

[101] Z. Lu and M. Yin. “Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. isbn: 9781450380966. doi: 10.1145/3411764.3445562.

[102] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze. “Evaluating saliency map explanations for convolutional neural networks: a user study”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 275–285. isbn: 9781450371186. doi: 10.1145/3377325.3377519.

[103] M. A. Gemalmaz and M. Yin. “Understanding Decision Subjects’ Fairness Perceptions and Retention in Repeated Interactions with AI-Based Decision Systems”. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’22. Oxford, United Kingdom: Association for Computing Machinery, 2022, pp. 295–306. isbn: 9781450392471. doi: 10.1145/3514094.3534201.

[104] M. Ghazvininejad, X. Shi, J. Priyadarshi, and K. Knight. “Hafez: an Interactive Poetry Generation System”. In: *Proceedings of ACL 2017, System Demonstrations*. Ed. by M. Bansal and H. Ji. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 43–48.

[105] C.-W. Chiang and M. Yin. “Exploring the Effects of Machine Learning Literacy Interventions on Laypeople’s Reliance on Machine Learning Models”. In: *Proceedings of the 27th International Conference on Intelligent User Interfaces*. IUI ’22. Helsinki, Finland: Association for Computing Machinery, 2022, pp. 148–161. isbn: 9781450391443. doi: 10.1145/3490099.3511121.

[106] G. W. BRIER. “VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY”. In: *Monthly Weather Review* 78.1 (1950), pp. 1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[107] B. Green and Y. Chen. “The Principles and Limits of Algorithm-in-the-Loop Decision Making”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). doi: 10.1145/3359152.

[108] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz. “Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (Oct. 2019), pp. 2–11. doi: 10.1609/hcomp.v7i1.5285.

[109] D. Das and S. Chernova. “Leveraging rationales to improve human task performance”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 510–518. isbn: 9781450371186. doi: 10.1145/3377325.3377512.

[110] H. Reijers, H. Leopold, and J. Recker. “Towards a Science of Checklists.” English. In: *Proceedings of the 50th Annual Hawaii International Conference on System Sciences*. 50th Annual Hawaii International Conference on System Sciences , HICSS ; Conference date: 04-01-2017 Through 07-01-2017. 2017.

[111] A. Gawande. *The Checklist Manifesto: How to Get Things Right*. Metropolitan Books, 2009.

[112] P. Anne Collins McLaughlin. “What Makes a Good Checklist”. In: *PSNet* (2010).

[113] P. Helmiö, K. Blomgren, A. Takala, S.-L. Pauniaho, R. Takala, and T. Ikonen. “Towards better patient safety: WHO Surgical Safety Checklist in otorhinolaryngology”. In: *Clinical Otolaryngology* 36.3 (2011), pp. 242–247. doi: https://doi.org/10.1111/j.1749-4486.2011.02315.x.

[114] A. B. BÖHMER, F. WAPPLER, T. TINSCHMANN, P. KINDERMANN, D. RIXEN, M. BELLENDIR, U. SCHWANKE, B. BOUILLON, and M. U. GERBERSHAGEN. “The implementation of a perioperative checklist increases patients’ perioperative safety and staff satisfaction”. In: *Acta Anaesthesiologica Scandinavica* 56.3 (2012), pp. 332–338. doi: https://doi.org/10.1111/j.1399-6576.2011.02590.x.

[115] A. Ćatić and J. Malmqvist. “Effective method for creating engineering checklists”. In: *Journal of Engineering Design* 24.6 (2013), pp. 453–475. doi: 10.1080/09544828.2013.766824.

[116] E. B. Davidow and C. King. “Developing and Using Checklists in Practice”. In: *Advanced Monitoring and Procedures for Small Animal Emergency and Critical Care*. John Wiley  Sons, Ltd, 2023. Chap. 4, pp. 47–52. isbn: 9781119581154. doi: https://doi.org/10.1002/9781119581154.ch4.

[117] M. Hammersley. “The issue of quality in qualitative research”. In: *International Journal of Research & Method in Education* 30.3 (2007), pp. 287–305. doi: 10.1080/17437270701614782.

[118] A.-K. Dyrvig, K. Kidholm, O. Gerke, and H. Vondeling. “Checklists for external validity: a systematic review”. In: *Journal of Evaluation in Clinical Practice* 20.6 (2014), pp. 857–864. doi: https://doi.org/10.1111/jep.12166.

[119] C. Andrade. “Internal, External, and Ecological Validity in Research Design, Conduct, and Evaluation”. In: *Indian Journal of Psychological Medicine* 40.5 (2018). PMID: 30275631, pp. 498–499. doi: 10.4103/IJPSYM.IJPSYM_334_18.

[120] M. Yin, Y. Chen, and Y.-A. Sun. “The Effects of Performance-Contingent Financial Incentives in Online Labor Markets”. In: *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013* 27 (June 2013), pp. 1191–1197. doi: 10.1609/aaai.v27i1.8461.

[121] A. Skulmowski and K. Xu. “Understanding Cognitive Load in Digital and Online Learning: a New Perspective on Extraneous Cognitive Load”. In: *Educational Psychology Review* 34 (June 2021), pp. 1–26. doi: 10.1007/s10648-021-09624-7.

[122] J. Yang, J. Redi, G. Demartini, and A. Bozzon. “Modeling Task Complexity in Crowdsourcing”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4.1 (Sept. 2016), pp. 249–258. doi: 10.1609/hcomp.v4i1.13283.

[123] J. Leppink, F. Paas, C. P. M. Van der Vleuten, T. Van Gog, and J. J. G. Van Merriënboer. “Development of an instrument for measuring different types of cognitive load”. In: *Behavior Research Methods* 45.4 (2013), pp. 1058–1072. issn: 1554-3528. doi: 10.3758/s13428-013-0334-1.

[124] M. Klepsch, F. Schmitz, and T. Seufert. “Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load”. In: *Frontiers in Psychology* 8 (2017). PubMed-not-MEDLINE, p. 1997. issn: 1664-1078. doi: 10.3389/fpsyg.2017.01997.

[125] A. Felstiner. “Working the Crowd: Employment and Labor Law in the Crowdsourcing Industry”. In: *Berkeley Journal of Employment and Labor Law* 32.1 (2011), pp. 143–203. issn: 10677666, 23781882.

[126] E. van Teijlingen and V. Hundley. “The importance of pilot studies.” In: *Nursing standard (Royal College of Nursing (Great Britain) : 1987)* 16.40 (June 2002), pp. 33–36. issn: 0029-6570. doi: 10.7748/ns2002.06.16.40.33.c3214.

[127] J. Oppenlaender, T. Abbas, and U. Gadiraju. “The State of Pilot Study Reporting in Crowdsourcing: A Reflection on Best Practices and Guidelines”. In: *Proc. ACM Hum.-Comput. Interact.* 8.CSCW1 (Apr. 2024). doi: 10.1145/3641023.

[128] L. Litman, J. Robinson, and C. Rosenzweig. “The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk”. In: *Behavior Research Methods* 47.2 (2015), pp. 519–528. issn: 1554-3528. doi: 10.3758/s13428-014-0483-x.

[129] L. Wang, Y. Yang, and Y. Wang. “Do Higher Incentives Lead to Better Performance? - An Exploratory Study on Software Crowdsourcing”. In: *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. 2019, pp. 1–11. doi: 10.1109/ESEM.2019.8870175.

[130] “Prospect Theory: An Analysis of Decision under Risk”. In: *Econometrica* 47.2 (1979), pp. 263–291. issn: 00129682, 14680262.

[131] S. Salimzadeh, G. He, and U. Gadiraju. “Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. isbn: 9798400703300. doi: 10.1145/3613904.3641905.

[132] X. Miao, H. Peng, Y. Gao, Z. Zhang, and J. Yin. “On Dynamically Pricing Crowdsourcing Tasks”. In: *ACM Trans. Knowl. Discov. Data* 17.2 (Feb. 2023). issn: 1556-4681. doi: 10.1145/3544018.

[133] C.-J. Ho, A. Slivkins, and J. W. Vaughan. “Adaptive contract design for crowdsourcing markets: bandit algorithms for repeated principal-agent problems”. In: *Proceedings of the Fifteenth ACM Conference on Economics and Computation*. EC ’14. Palo Alto, California, USA: Association for Computing Machinery, 2014, pp. 359–376. isbn: 9781450325653. doi: 10.1145/2600057.2602880.

[134] T. Draws, A. Rieger, O. Inel, U. Gadiraju, and N. Tintarev. “A Checklist to Combat Cognitive Biases in Crowdsourcing”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9.1 (Oct. 2021), pp. 48–59. doi: 10.1609/hcomp.v9i1.18939.

[135] N. Kaufmann, T. Schulze, and D. Veit. “More than fun and money. Worker Motivation in Crowdsourcing–A Study on Mechanical Turk”. In: Jan. 2011.

[136] O. Nov, O. Arazy, and D. Anderson. “Scientists@Home: What Drives the Quantity and Quality of Online Citizen Science Participation?” In: *PloS one* 9 (Apr. 2014), e90375. doi: 10.1371/journal.pone.0090375.

[137] F. Cappa, F. Rosso, and D. Hayes. “Monetary and Social Rewards for Crowdsourcing”. In: *Sustainability* 11.10 (2019). issn: 2071-1050. doi: 10.3390/su11102834.

# A

# VARIATIONS OF REFLEXIVE THEMATIC ANALYSIS

The following table briefly summarizes the different variations of reflexive TA, as described by Braun and Clarke [72].

| | | |
|---|---|---|
| *Orientation to data* | *Inductive*: Analysis is data-driven, codes and themes emerge from the data itself. | *Deductive*: Analysis is theory-driven, existing theoretical constructs are used to guide coding and theme development. |
| *Focus of meaning* | *Semantic*: Focuses on the explicit meaning of the data, ensuring themes accurately reflect what researchers have written. | *Latent*: Explores the underlying or implicit meaning within the data. |
| *Qualitative framework* | *Experiential*: Focuses on the lived experiences and perspectives of participants within the data. | *Critical*: Analysis interrogates and unpacks broader meanings and implications around the topic. |
| *Theoretical frameworks* | *Realist, essentialist*: Analysis aims to capture objective truth and reality as expressed in the data. | *Relativist, constructionist*: Analysis examines and deconstructs the realities represented in the data. |

Table A.1.: Table summarizing the different variations of reflexive thematic analysis