Master Thesis

Mitigating Mental Health Misinformation on Instagram: A Systemic and Value-based Approach

Management of Technology

Sophia Zanon Brenck

Delft University of Technology

TU Delft

Master Thesis

Mitigating Mental Health Misinformation on Instagram: A Systemic and Value-based Approach

submitted by

Sophia Zanon Brenck

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Monday June 16th, 2025 at 11:00 AM.

Student number:	5934907	
Project duration:	January 2025 – June 2025	
Thesis committee:	Dr. ir. R. I. J. Dobbe,	TU Delft, Supervisor
	Dr. ir. L. Marin,	TU Delft, Supervisor
	Dr. S. Hinrichs-Krapels,	TU Delft, Chair

Cover Image: Richardson, K. (2025, February 12). The impact of social media misinformation on mental health. Lebanon Valley College. https://www.lvc.edu/news/the-impact-of-social-media-misinformation-on-mental-health/

TU Delft

Abstract

The widespread use of social networks has created new pathways for sharing and engaging with mental health information, particularly in contexts where access to care is limited and stigma remains high. Platforms like Instagram have become informal spaces for support, especially among young people. However, these same platforms also amplify misinformation, often through algorithmic systems that shape user perceptions and behavior in harmful ways.

This thesis investigates how to mitigate the risks of mental health misinformation on Instagram, using the Brazilian context as a case study. Integrating System-Theoretic Process Analysis (STPA) and Value Sensitive Design (VSD) to develop socio-technical interventions that are both system-aware and ethically grounded. The research unfolds in three phases: a conceptual phase to identify hazards and system dynamics; an empirical phase based on semi-structured interviews with Brazilian young adults; and a technical phase that synthesizes system risks and user values into actionable design recommendations.

The findings emphasize the importance of four core values: *Knowledge, Autonomy, Safety* and *Integrity* in shaping how users interpret and respond to mental health content. Participants described how Instagram's algorithm reinforces emotionally charged echo chambers, increasing exposure to harmful misinformation while weakening trust in credible sources. Based on this, four interventions are proposed: (1) echo chamber disruption mechanisms, (2) content trigger warnings, (3) verified institutional accounts, and (4) user-controlled filters for validated content.

By framing misinformation as a socio-technical problem rooted in both platform architecture and user experience, this study offers a novel methodological contribution. It demonstrates how combining system-level safety analysis with value-centered design can support the development of interventions that are not only effective but also aligned with user priorities and sociocultural context.

Preface

Two years ago, I dreamed of this moment, imagining what it would feel like, what it would smell like. Returning to academic life was, above all, a personal decision. I wanted to take on a new challenge, rediscover the joy of learning, and most of all meet incredible people.

During the past two years, I have done exactly that. This journey brought meaningful struggles, not only in navigating study life, but also in adapting to a new country that I now proudly call home. I have made wonderful friends, met inspiring people, and, perhaps most unexpectedly, discovered a learning capacity within myself that I did not know existed.

When it came time to pursue my thesis, I chose a topic that felt both personally meaningful and challenging, something that I hoped could make a small contribution to the world. I was lucky to find a topic that kept me curious and motivated throughout the past five months, making this journey a true pleasure.

I would like to begin by thanking my thesis committee, who challenged me intellectually while offering all the emotional and academic support I needed to bring this work to life. Roel, thank you for your trust, for your guidance before and during the thesis process, and for helping me organize my confusing thoughts. Lavinia, your thoughtful feedback and consistent insights were essential in helping me shape the direction of this research. And Saba, thank you for your powerful reflections and for believing in this work from the beginning. Each of you made this process not only meaningful, but truly rewarding.

To my fellow MOT colleagues, thank you for being patient with me these past two years and for walking this path together. A special thanks to Luis, my Latin-American partner-in-crime who refused to do a single-group project without me. To Camila, thank you for being my safe space and creative element. And, most importantly, thank you to Robin, the greatest gift that MOT gave me. You were my colleague, my partner, and my biggest support. You walked me through the hardest times and I can't wait to see what the future holds for us.

To my *gezin*, the De Boer family, thank you for becoming my anchor. Your love and support turned this new place into a true home. And to all the other friends and family in the Netherlands, thank you for the warmth, the laughter, and the *gezelligheid*.

To my family and friends back in Brazil, you have always been my foundation, my motivation, and my endless source of love. Thank you for softening the *saudade* and for being present, even from thousands of kilometers away.

Finally, to mom, dad, Clara, and Lola, thank you for everything. I would not be here without your love and support. Eu amo vocês!

I am proud to carry the energy, memories, and experiences of these past years into the next phase of life.

Now, it is time to start a new chapter.

Sophia Zanon Brenck
5 June 2025

Summary

The widespread use of social networks has created new opportunities for sharing mental health information, particularly important given the global treatment gap and the barriers many people face in accessing care. Platforms like Instagram have become informal spaces to seek support, information, and connection, especially among young people. In contexts where stigma, limited access to professionals, or institutional distrust is prevalent, social networks often fill a gap left by formal systems.

At the same time, these platforms pose serious challenges in managing the quality and precision of mental health content. Misinformation is common and, when amplified through algorithmic systems, can shape user behavior and perceptions in harmful ways. However, despite its potential impact, mental health misinformation remains underexamined, particularly from the user perspective and within specific sociocultural contexts.

This thesis investigates how to mitigate the risks and harms of misinformation about mental health on Instagram, using the Brazilian context as a case study. Brazil presents a particularly relevant setting due to its high use of social media and its significant mental health burden. The research focuses on young adults and seeks to answer the following question:

What socio-technical interventions can mitigate the risks and harms of mental health misinformation on Instagram, considering the perspectives of young adults in Brazil?

To address this, the study combines system-theoretic analysis with value-sensitive design (VSD). VSD helped identify the ethical principles and user priorities that shape how individuals engage with mental health content, while system analysis mapped the interactions between users, algorithms, and platform mechanisms to understand how harm arises.

The research unfolded in three phases. First, a conceptual phase established a system-level understanding of the problem using stakeholder mapping, causal diagrams, and an initial STPA (System-Theoretic Process Analysis) to identify hazards and safety constraints. Second, an empirical phase collected information through semistructured interviews with Brazilian young adults, incorporating fictional scenarios to explore their perceptions and values of risk. Finally, a technical phase synthesized these findings into a refined STPA, integrating user values into the development of systemic design recommendations.

This process led to the priority of four core user values: *knowledge*, *autonomy*, *safety*, and *integrity*. Users reported how Instagram's algorithm often reinforces emotionally charged echo chambers, influencing their mental well-being and trust in content. Based on this, four system-level interventions were proposed: (1) a mechanism to break echo chambers, (2) trigger warnings for mental health content, (3) verified accounts for trusted institutions, and (4) a user-controlled filter for validated information.

The study concludes that the risks of misinformation are not just technical, but deep embedded in how users navigate content through personal, emotional, and social filters. Although misinformation is often treated as a regulatory or moderation issue, participants rarely mentioned moderation directly, suggesting a disconnect

between institutional responses and user experiences. Even health professionals interviewed shared how platform incentives sometimes pressured them to oversimplify information, unintentionally contributing to the problem.

By integrating System-Theoretic Process Analysis (STPA) with Value Sensitive Design (VSD), this thesis proposes a methodological framework for developing value-aligned interventions against mental health misinformation on Instagram. Frame misinformation as a socio-technical issue shaped by platform architecture, algorithmic behavior, user interaction, and cultural norms. Future research should examine the practical implementation and regulatory feasibility of these interventions, as well as their long-term impact on user trust. In general, the study offers a novel approach that connects system-level safety analysis with user values to inform ethically and technically grounded solutions.

Contents

List of Figures

List of Tables

1

Introduction

In this opening chapter, we explore the broader context and relevance of this research by outlining key problems related to mental health misinformation on social media. We begin by describing the global mental health problem and the growing role of platforms such as Instagram in shaping how people access and engage with mental health information. Brazil is introduced as a case study due to its high rates of social media use and significant mental health challenges. The chapter then presents the research objective and questions, clarifying the study's focus and motivations. Finally, it outlines the structure of the thesis and the methodological approach taken to develop value-sensitive systems-level interventions.

1.1. Problem formulation

Mental illnesses are a leading cause of disability worldwide, yet many individuals continue to face major barriers to accessing essential mental health care (Naslund & Deng, 2021). Contributing factors include stigma and discrimination (Thornicroft, 2008), as well as structural challenges such as geographic distance and long wait times (De Hert et al., 2011), all of which reinforce the global mental health treatment gap. As a result, social media platforms have become an important tool for sharing and seeking mental health-related information (Chen & Wang, 2021), with more 97% people with severe mental health conditions using social media daily (Birnbaum et al., 2017). Platforms like Instagram, Facebook, and Twitter provide spaces for people to share, seek advice, and build supportive communities (Naslund et al., 2016). The use of social networks for health purposes offers an opportunity to increase health literacy, self-efficacy, and adherence to treatment among populations (Suarez-Lledo & Alvarez-Galvez, 2021).

However, the same factors that make social media an easily accessible source of information also make it an efficient tool for creating and broadcasting false information (Bodaghi et al., 2024). Although misinformation is often shared or produced without malicious intent (Bodaghi et al., 2024), it can still distort public perceptions of mental health and influence treatment-seeking behaviors (Starvaggi et al., 2024). While extensive research has addressed misinformation in areas such as vaccines and other health topics, mental health misinformation remains significantly underexplored (Suarez-Lledo & Alvarez-Galvez, 2021).

The relevance of this issue becomes evident when examining the prevalence of misinformation in mental health content on social networks. More specifically, a study with 500 social media videos labeled #mentalhealthadvice or #mentalhealthtips found that 31% contained scientifically inaccurate information and 14% were classified as potentially harmful (PlushCare, 2025). According to World Health Organization (WHO)

(2022), the consequences of misinformation on social networks include such negative effects as *'an increase in the erroneous interpretation of scientific knowledge, polarization of opinion, increased fear and panic, or decreased access to health care.'*

In September 2022, a British court determined that Instagram posts played a role in the suicide of 14-year-old Molly Russell (The Guardian, 2022). According to data obtained by her family, she had saved, shared, or liked 2,100 posts related to depression, self-harm, or suicide in the six months leading up to her death. The Center for Countering Digital Hate (2022) study by the Center for Countering Digital Hate (CCDH) found that TikTok's algorithm exposes new teen accounts to self-harm and eating disorder content in minutes, with suicide-related content appearing in 2.6 minutes. In addition, accounts that engage with body image and mental health content were shown similar material every 39 seconds.

1.2. Brazil as a case study

Brazil is a large country with about 210 million people (by 2025), and social networks are a huge part of Brazilian daily life. Instagram alone has 134 million active users, representing 62% of the population (Statista, 2024). At the same time, mental health issues are a growing concern. Brazil ranks fifth worldwide in the prevalence of depression and leads to anxiety disorders (Souza, 2017). By 2021, anxiety alone affected around 9% of the population (Statista, 2022). More recently, in March 2025, the government reported an increase in 400% work leave requests due to anxiety and depression since 2020, increasing from 91,600 to 472,000 in 2024 (InfoMoney, 2024).

The country also has one of the largest free healthcare systems in the world, which guarantees universal medical access as a constitutional right. However, the system continues to face persistent challenges, including unequal access to healthcare care, variations in quality, and disparities in care levels (Roman, 2023). Following worldwide phenomena, geographic barriers, poverty, and long waiting times also impact access to care (De Hert et al., 2011), in addition to stigma, discrimination, and social isolation (Thornicroft, 2008). As noted in Filho (1999), these limitations have contributed to the growth of private healthcare providers since the 1960s, offering an alternative for those seeking faster or higher-quality medical services.

There is growing evidence of health misinformation that spreads rapidly and poses significant risks to public health. Research indicates that false or misleading health information is widespread in Brazil, some cases even involving healthcare professionals as sources or distributors (Marocolo et al., 2021). In this thesis, mental health is considered a subset of general health; however, it also presents specific challenges, such as stigma, underdiagnosis, and limited access to treatment. We then assume that the spread of mental health misinformation follows patterns similar to general health misinformation, while also being shaped by the unique vulnerabilities of the mental health field.

Given the scale of misinformation and its impact, it is essential to understand how Brazil's digital governance structures address this issue, particularly in relation to online content moderation, platform accountability, and user rights. The following section outlines the key Internet regulations that shape Brazil's approach to online governance and provide a basis for the system analysis developed in this thesis.

1.2.1. Contextualization of Brazil's Regulatory landscape

To contextualize the Brazilian landscape, an initial review of existing and emerging legislation related to mental health misinformation on social media was carried out. This step was essential to frame the topic and inform subsequent analysis.

1.2.1.1. Marco Civil da Internet (the "Internet Constitution")

The Marco Civil da Internet, passed in 2014, is one of the main laws regulating internet use in Brazil and is often called the country's "Internet Constitution". It sets guidelines for how the Internet should be governed, covering topics such as net neutrality, user privacy, freedom of speech, and the responsibility of service providers (Presidência da República Federativa do Brasil, 2014). The law requires Internet providers to treat all data equally without favoring or restricting access based on content, origin, or destination. It also states that providers are not responsible for user-generated content, but must follow court orders to remove certain posts when necessary. However, some critics argue that it does not do enough to tackle misinformation or hold platforms accountable for harmful content, which would justify the need for new regulations such as the Fake News Bill (Tech Policy Press Staff, 2023).

1.2.1.2. Lei Geral de Proteção de Dados (LGPD), inspired by General Data Protection Regulation (GDPR)

The Lei Geral de Proteço de Dados (LGPD), approved in 2018 and fully in effect since 2020, is Brazil's main law on data protection, inspired by the GDPR of the European Union. It sets the rules for how personal data is collected, processed and stored by both public and private organizations, ensuring that people have rights over their own information (G. F. do Brasil, 2018). The law requires companies to obtain clear user consent before handling personal data and to put security measures in place to protect it. The National Data Protection Authority (ANPD) is responsible for ensuring that the law is followed.

1.2.1.3. Proposed Brazil's fake news bill

Also called the Brazilian Law on Freedom, Responsibility, and Transparency on the Internet, this proposed law aims to fight the spread of false information, including health-related misinformation, on social media and messaging platforms (Silveira et al., 2025). The bill proposes measures such as mandatory user identification, content verification mechanisms, and holding digital platforms responsible for disseminating false information.

In February 2025, the Brazilian "Fake News Bill" was not enacted into law. The proposal has sparked debate, with strong opinions from different groups, including tech companies, government representatives, and civil organizations. Opponents worry that it could limit free speech and affect user privacy, while supporters believe that it is necessary to control the spread of false information online (Anthony Boadle, 2023).

In this specific proposed law, misinformation is defined as:

"Content that is, in whole or in part, unequivocally false or misleading, verifiable, placed out of context, manipulated, or fabricated, with the potential to cause individual or collective harm, except in cases of humorous intent or parody" (Brazilian Fake News Bill, 2020, Art. 4, as cited in S. F. do Brasil, 2020).

1.3. Research Motivation and the Socio-Technical Landscape

This thesis is motivated by the growing use of social networks to disseminate mental health information and the growing concern over the spread of misinformation in these digital environments. Although topics like vaccine misinformation have been widely studied, mental health misinformation remains significantly underexplored, particularly from a socio-technical perspective. However, platforms like Instagram clearly function as socio-technical systems, where user behavior, platform design, and institutional structures interact to shape how information is produced, shared, and perceived.

Figure 1.1 shows the scope of this research through a layered view of the system. The outer circle represents socio-technical systems, where social and technical elements influence each other. Inside that, the mental

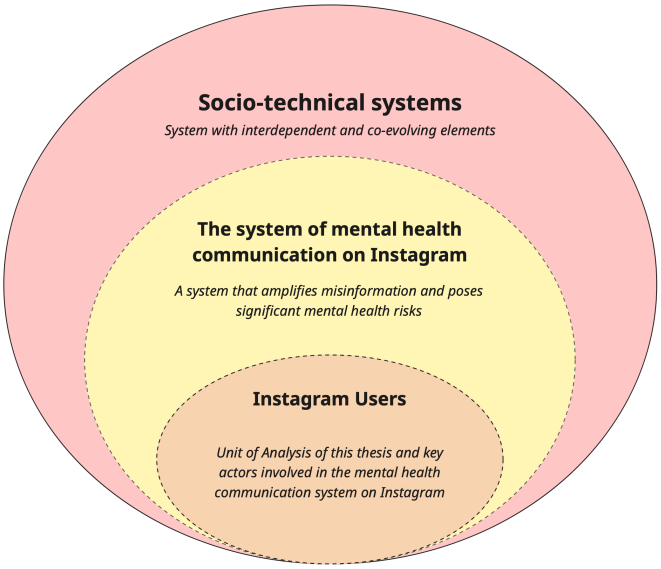

Figure 1.1: Scope of the Study within a Socio-technical System

health communication system on Instagram is defined as a space where content, users, and algorithms interact, potentially promoting misinformation and creating risks. In the center are Instagram users, the unit of observation of this study, and those who are both affected by and contribute to this system. This structure reflects the research approach and objective in the next section.

1.4. Research Objective and Questions

This thesis aims to develop socio-technical interventions and user-centered interventions to mitigate the risks and harms of misinformation about mental health on Instagram. Rather than focusing solely on safety or content moderation, the objective is to design recommendations that reduce harm while respecting the motivations, perceptions, and values of users.

The study uses system analysis to map how misinformation risks emerge within the socio-technical structure of Instagram and identify points of intervention. Value-oriented semi-structured interviews complement this by exploring how users perceive these risks and what values shape their engagement with mental health content. These insights guide the design of interventions that are both technically grounded and socially responsive.

The desired research output is a set of socio-technical interventions designed to reduce risks and harms while also respecting the means, aims, and wants of the users. To achieve this, the Master's thesis seeks to answer the following research question:

> **What socio-technical interventions can mitigate the risks and harms of mental health misinformation on Instagram, considering the perspectives of young adults in Brazil?**

In this study, socio-technical interventions are not treated as isolated solutions but as outcomes that emerge from a broader understanding of the socio-technical system in which they are embedded. The risks and harms associated with mental health misinformation on Instagram are not solely the result of individual user behavior or algorithmic shortcomings; rather, they arise from a complex interplay of platform design, user practices, institutional policies, cultural dynamics, and social norms. Within this framework, socio-technical

interventions are understood as targeted design or system-level changes informed by the interdependencies among these elements.

1.4.1. Research subquestions

In order to design effective interventions that can reduce the spread and impact of misinformation, one needs to first understand how harms emerge within this complex system. This includes analyzing the nature of mental health misinformation, mapping out its pathways of influence, and identifying how users perceive and respond to such content. In addition, designing interventions that are both ethical and effective requires attention to the values that users bring to their engagement with mental health information online.

Therefore, to answer the main research question above, we aim to also address the following subquestions:

RQ1: How do risks related to mental health misinformation emerge within social media platforms analyzed as socio-technical systems?

This question seeks to identify the specific risks and harms associated with social media misinformation about mental health and to examine the systemic factors that contribute to their emergence. The goal is to map the structural dynamics that shape the spread and impact of misinformation within these platforms.

RQ2: What values shape the meaning, goals, and wants of users when engaging with mental health information on social media?

This question aims to explore the underlying values and associated reasons behind the participation of users in mental health information in general. Rather than focusing solely on misinformation, it considers the broader information system, acknowledging that users may not always be able to assess the credibility of the content they see online.

RQ3: How do young adults in Brazil perceive the risks and harms of misinformation about mental health?

The objective of Question 3 is to assess the awareness of users of the risks and harms associated with misinformation about mental health. Perception is the first level and an important element in creating awareness of the situation around a risk (Endsley, 1995). As highlighted by Morita and Burns (2011), people with strong awareness can navigate their environments more effectively, accurately interpret information, and take appropriate actions to mitigate potential risks. Understanding user perceptions of the issue allows the development of more targeted recommendations while also revealing their priorities and concerns.

RQ4: How can insights from perception, values, and existing interventions be operationalized into systemic recommendations for Instagram?

Question 4 serves as the methodological bridge of this thesis. It brings together the insights from the previous subquestions and forms the final connection to the main research question.

1.5. Thesis Outline

This thesis is structured into seven chapters, each contributing to a system-theoretic and value-sensitive understanding of how to mitigate mental health misinformation on Instagram. Chapter 1 introduces the research topic, problem statement, objectives, and questions. Chapter 2 defines key concepts, reviews the literature, and highlights the main research gap. Chapter 3 outlines the research design and methods.

Chapter 4 begins the conceptual analysis, identifying stakeholders and risks, and applies STPA to define initial loss scenarios and constraints. Chapter 5 brings the empirical results and presents the findings of the user interviews on risk perception. Chapter 6 integrates these insights to refine the STPA model and propose value-aligned systemic interventions.

Chapter 7 concludes with answers to the research questions, key contributions, study limitations, and future directions. Table 1.1 summarizes the structure of the chapter.

Table 1.1: Overview of Thesis Chapters and Content

Chapter	Content
1. Introduction	Problem statement, research objectives, and research questions.
2. Literature Review	Defines key constructs, synthesizes relevant literature, and identifies the research gap.
3. Methodology	Research strategy, methodology, and methods; justification of choices and outline of research phases.
4. Conceptual	Analyzes how mental health misinformation creates risks; identifies stakeholders; Applies STPA to define loss scenarios, hazards, and initial safety constraints.
5. Empirical	Interview findings on user perceptions and values related to mental health content.
6. Technical	Refines STPA model and proposes value-sensitive, risk-informed systemic interventions.
7. Discussion and Conclusion	Reflects on contributions, synthesizes findings, discusses limitations, and outlines future directions.

2

Literature Review

To ground this thesis in existing academic knowledge, a systematic review of the literature was conducted to identify relevant concepts, current debates, and research gaps related to misinformation about mental health on social media. The review aimed not only to map the landscape of existing studies, but also to inform the conceptual framework and guide the development of the interventions. This chapter outlines the review process, including search strategies, database selection, and inclusion/exclusion criteria.

2.1. Literature Review Process

Our literature review process followed a structured sequence of steps aimed at identifying key concepts, relevant articles, and the research gap that would help define the scope of this thesis. We started by formulating a targeted search string, which served as the foundation for our database queries across selected platforms.

Search String Used in Database

```
("mental health") AND ("misinformation" OR "fake news") AND ("social media" OR
"Instagram" OR "algorithm")
```

At first, we deliberately limited our search to studies focused on understanding the problem space, rather than exploring solutions. This approach allowed us to map the scope and characteristics of mental health misinformation on social media, particularly how it is defined, perceived, and amplified. In the later stages of the literature review process, system analysis was incorporated as a methodological lens to address this problem. This shift helped to clarify the research gap, which is discussed in more detail at the end of this chapter.

To maintain a focus on everyday patterns of misinformation and reduce context-specific bias, we excluded articles that focused primarily on the COVID-19 pandemic. Many of these studies addressed general health misinformation, and mental health appeared only as a secondary concern. Vaccine-related misinformation was also prevalent in this period and was beyond the scope of this thesis. Table 2.1 summarizes the database search process and the exclusion of literature related to pandemics.

Table 2.1: Overview of Literature Search Across Databases

Database	Search Terms Used	Hits	Exclusion Criteria	Screened
Scopus	(”mental health”) AND (”misinformation” OR ”fake news”) AND (”social media” OR ”Instagram” OR ”algorithm”)	166	Pandemic-related articles	66
Google Scholar	”mental health misinformation” AND social media AND ”solutions” OR ”mitigation”, 2020 and above	71	All found documents were screened	56
PubMed	(”mental health”) AND (”misinformation” OR ”disinformation” OR ”fake news”) AND (”Instagram” OR ”social media”)	131	NOT (vaccine OR COVID-19)	53

From the selected articles, we identified a set of foundational studies that helped shape the direction of our research. In particular, while searching PubMed and Google Scholar, we found that the study by Suarez-Lledo and Alvarez-Galvez (2021) was frequently cited and served as a key reference in multiple sources. We used this study as a starting point to build the visual graph shown in Figure 2.1, which was generated using the Connected Papers platform (Connected Papers, 2025). This helped us identify the most frequently co-cited articles related to Suarez-Lledo and Alvarez-Galvez (2021), a process that also supported the selection of other key initial articles for our review.

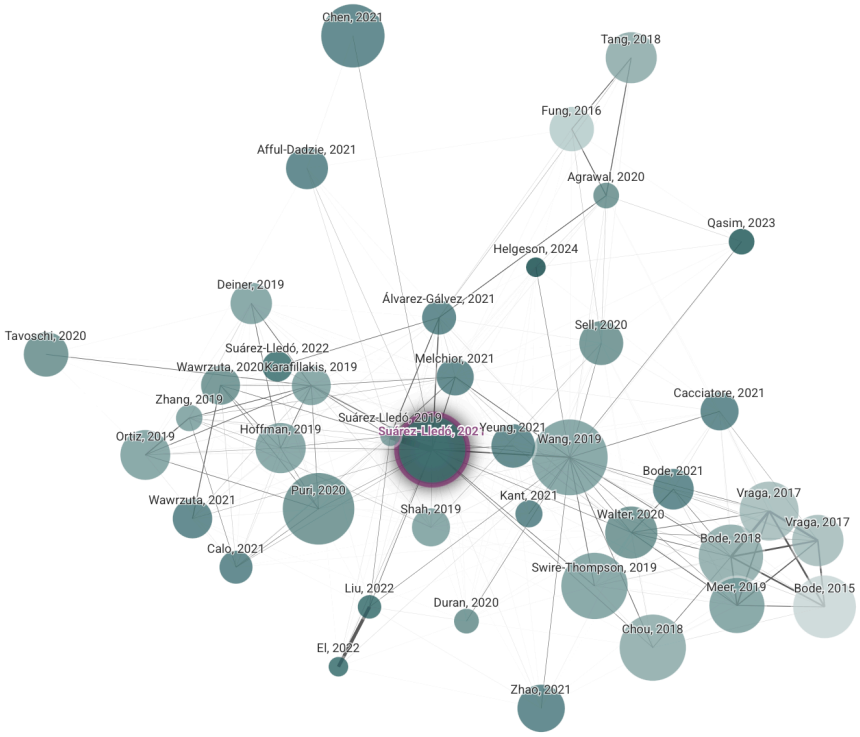

Figure 2.1: Visual graph generated using Connected Papers, based on the study by Suarez-Lledo and Alvarez-Galvez (2021)

The initial selected articles were then organized into categories according to what is shown in table 2.2. It is important to note that not all articles shown in the initial graph were selected for in-depth review at that point. As the thesis progressed, additional concepts were explored and new literature was incorporated accordingly.

Table 2.2: Overview of Literature by Category

Category	Related Article	Source
Misinformation on Social Media: The Nature, Spread, and Impact of False Content	Mental health misinformation on social media: Review and future directions	(Starvaggi et al., 2024)
	Prevalence of Health Misinformation on Social Media: Systematic Review	(Suarez-Lledo & Alvarez-Galvez, 2021)
	Systematic Literature Review on the Spread of Health-related Misinformation on Social Media	(Wang et al., 2019)
	Misinformation on Instagram: The Impact of Trusted Endorsements on Message Credibility	(Mena et al., 2020)
	Why people accept mental health-related misinformation: Role of social media metrics in users' information processing	(S. Zhang, 2024)
The Behavior of Seeking and Sharing Mental Health Content: How Users Interact with the Content and the Motivations for These Behaviors	Social Media Use as Self-Therapy or Alternative Mental Help-Seeking Behavior	(Gere et al., 2020)
	The future of mental health care: Peer-to-peer support and social media	(Naslund et al., 2016)
	Exploring opportunities to support mental health care using social media	(Naslund et al., 2019)
	Online Health Information Seeking Behavior: A Systematic Review	(X. Jia et al., 2021)
	Users' health information sharing behavior in social media: An integrated model	(C. Jia & Qi, 2024)
Platform Design and Features: Instagram's Technical and Design Elements, Exploring Opportunities for Redesign	Exploring Digital Affordances in Online Mental Health Resources	(Harris, 2024)
	What Drives Health Information Exchange on Social Media?	(Wu et al., 2024)
	The Needs–Affordances–Features Perspective for the Use of Social Media	(Karahanna et al., 2018)
	Design Opportunities for Mental Health Peer Support Technologies	(O'Leary et al., 2017)
	Integrating Social Values into Software Design Patterns	(Hussain et al., 2018)
	A Literature Review on Detecting, Verifying, and Mitigating Online Misinformation	(Bodaghi et al., 2024a)
	The Ineffectiveness of Fact-Checking Labels on News Memes and Articles	(Oeldorf-Hirsch et al., 2020)
	Discerning Individual Preferences for Identifying and Flagging Misinformation on Social Media	(Barman et al., 2024)
Solution Space: Strategies for Addressing Mental Health Misinformation on Social Media	A Literature Review on Detecting, Verifying, and Mitigating Online Misinformation	(Bodaghi et al., 2024a)
	The Ineffectiveness of Fact-Checking Labels on News Memes and Articles	(Oeldorf-Hirsch et al., 2020)
	Discerning Individual Preferences for Identifying and Flagging Misinformation on Social Media	(Barman et al., 2024)
	Combating misinformation online: Re-imagining social media for policy-making	(Kyza, 2020)

2.2. Key Concepts

This section defines the core concepts underpinning the study: mental health misinformation, recommendation algorithms, echo chambers, and the mechanisms through which platforms may intervene. Together, these concepts inform the system analysis and interventions proposed in later chapters.

2.2.1. Mental Health Misinformation

The World Health Organization defines mental health as a state of well-being that enables people to cope with stress, realize their abilities, work productively, and contribute to their community Organization (2022). Misinformation, in turn, refers to false or misleading content shared with no intention of harm, distinct from disinformation, which involves deliberate deception (Wardle, 2017).

In the context of health, Suarez-Lledo and Alvarez-Galvez (2021) describes misinformation as 'health-related claims that are false, misleading, or based on anecdotal evidence due to lack of scientific support'. Combining this with studies specific to mental health such as S. Zhang et al. (2024), this thesis defines mental health misinformation as *a subgroup of claims related to mental health that are not supported by existing scientific knowledge*.

2.2.2. Recommendation Algorithms

Recommendation algorithms are algorithms designed to personalize and rank content based on user prior behavior, preferences, and engagement patterns (Helberger, 2020). On platforms like Instagram, they determine what users see in their feeds, on the *Explore page*, and on *Reels* by optimizing for metrics such as likes, shares, watch time and comments. These algorithms function as gatekeepers of information, significantly influencing user exposure and attention (Gillespie, 2018).

In the context of health-related content, recommendation systems can inadvertently amplify misinformation by prioritizing posts that trigger emotional responses or generate high engagement, regardless of precision (Cinelli et al., 2021; S. Zhang et al., 2024). This dynamic increases the visibility of simplified or anecdotal mental health content while reducing exposure to diverse or scientifically grounded perspectives. As such, recommendation algorithms are not neutral tools, but active mediators of what becomes credible and visible in digital environments (Cotter, 2019).

2.2.3. Echo Chambers

The concept of echo chambers, as defined by Nguyen (2020), refers to environments in which dissenting voices are not merely absent but actively discredited. Users are conditioned to distrust external sources, which results in communities that are resistant to opposing viewpoints. Echo chambers emerge when users are consistently exposed to content that reinforces their existing beliefs, emotional needs, or personal identities. On platforms like Instagram, this dynamic is amplified by algorithmically personalized recommendations and social validation from peers and influencers. Within such environments, users can develop greater trust in unverified content, particularly when presented by emotionally resonant or relatable figures (Tang et al., 2024; S. Zhang et al., 2024).

A study by Fernández et al. (2021) that analyzed the impact of recommendation algorithms found that even users with initial low exposure to misinformation can, by repeatedly accepting algorithmic suggestions, quickly enter reinforcement cycles that increase their exposure. This insight is critical for our analysis, as it highlights

that echo chambers of misinformation may not stem solely from user behavior, but can also emerge from the inherent design and feedback loops of the recommendation algorithms themselves.

2.2.4. User Perception of misinformation

User perception plays a critical role in how misinformation spreads and how it can be mitigated. A study by Tang et al. (2024) found that users generally define mental health misinformation as content lacking scientific support, but often view it as 'gray', meaning it can be unverified but still perceived as helpful or emotionally supportive. Users also express mixed opinions on the intervention of the platform. Most prefer labels or warnings over content deletion and emphasize the importance of transparency in how content is evaluated and flagged (Tang et al., 2024).

2.2.5. Content Moderation

Content moderation refers to the processes through which platforms define and enforce what is considered acceptable user-generated content. These decisions are shaped by platform policies, cultural norms, and legal frameworks (Roberts, 2019). However, moderation practices often operate in a 'black box', largely invisible to the public (Witt et al., 2019).

On Instagram, where much of the content is visual, human moderators play a key role in interpreting and applying platform guidelines. Their decisions are influenced by cultural context and personal judgment, making moderation uneven and sometimes opaque (Roberts, 2019).

2.2.6. Trigger warnings

Trigger warnings are preemptive alerts that alert people to content that could cause emotional or psychological distress, particularly for those with a history of trauma (Rothman, 2023). In the realm of misinformation prevention, trigger warnings function similarly to content labels or warnings on social media platforms. These labels aim to inform users about the questionable nature of certain content, thereby reducing its spread and impact (Martel & Rand, 2023).

2.3. Knowledge Gap

Upon reviewing the literature, we found that while the prevalence and harms of misinformation are well documented in several health domains, especially in areas such as infectious diseases and vaccine hesitancy. Research focused specifically on mental health misinformation remains relatively limited (Starvaggi et al., 2024). This is especially concerning given the growing public discourse on mental health and the increasing role of social media in shaping how people understand and manage psychological well-being.

In addition to this broad thematic gap, there is limited understanding of how users perceive and interact with mental health misinformation on platforms such as Instagram. As highlighted in Tang et al. (2024), user opinions on what qualifies as misinformation and how it should be addressed vary widely. Although some users prefer content labeling and contextual cues, others distrust moderation mechanisms altogether. These findings underscore the need to capture user perspectives, not only on the perceived harms of misinformation, but also on what types of intervention are considered legitimate or desirable. Without this understanding, well-planned interventions can risk being ineffective or even counterproductive.

The literature on misinformation mitigation is also inconclusive. For example, Barman et al. (2024) suggests

that detailed fact check labels can improve user understanding, while Oeldorf-Hirsch et al. (2020) reports that platform credibility may outweigh fact-check accuracy in shaping trust. These contrasting findings point to an unresolved debate on what intervention mechanisms are effective and socially accepted, especially in sensitive domains such as mental health.

From a geographical perspective, most studies are concentrated in high-income countries such as the United States, the United Kingdom, and Australia. However, as noted by Naslund et al. (2020), mental health challenges and digital engagement are growing rapidly in low- and middle-income countries (LMICs), such as Brazil, where this study is located. The combination of high social media use, underresourced mental health systems, and regulatory uncertainty makes LMICs a particularly relevant but underexplored context.

In addition, the review of the literature revealed a gap in the connection between the nature of the content and the methodological approaches used to study it. As shown in Table 2.3, very few studies explicitly adopt a system-oriented approach to understanding or mitigating mental health misinformation. Using targeted searches that combine terms such as mental health misinformation, Instagram, and system analysis, we found only two relevant results in Google Scholar, one in Scopus, and none in PubMed. This strongly suggests that the integration of system analysis into the study of mental health misinformation is still largely absent from academic research.

Table 2.3: Search Results for System Analysis and Mental Health Misinformation

Database	Search Terms Used	Results
Google Scholar	”mental health misinformation” AND Instagram AND ”system analysis”	2
Scopus	”mental health” AND ”misinformation” AND (Instagram OR ”social media”) AND (“system analysis” OR “systems thinking” OR “sociotechnical”)	1
PubMed	”mental health” AND ”misinformation” AND social media AND (”system analysis”)	0

Furthermore, when replacing system analysis with Value Sensitive Design (VSD) in the search structure presented in Table 2.3, none of the databases returned any relevant results. This absence highlights a significant research gap. VSD has not yet been applied meaningfully to the study of mental health misinformation, particularly within platform ecosystems like Instagram. This gap suggests a lack of methodological integration between value-based frameworks and the problem of misinformation about mental health on social networks.

2.3.1. Methodological gap

One of the key gaps identified in the literature is the absence of studies that address the problem of mental health misinformation through the lens of system analysis or value-sensitive design. Beyond the thematic focus, there is also a notable methodological gap: these two approaches, particularly value-sensitive design and system-theoretic process analysis (STPA), are rarely applied in combination in peer-reviewed research.

Table 2.4 presents the results of the literature search for articles referencing both Value Sensitive Design (VSD) and System Analysis, as well as VSD and the System-Theoretic Accident Model and Processes (STAMP) or its analytical tool, System-Theoretic Process Analysis (STPA). Although several results were recovered under the broader search for ’system analysis’, this was not the case for STPA specifically. Of the nine results found combining STPA and VSD, closer inspection revealed that the two frameworks were treated separately in nearly all cases, with no meaningful methodological integration between them.

In our screening, we perceived that typically VSD is used to identify or surface stakeholder values, while STPA is applied to assess safety-related risks. For example, Rao et al. (2022) refer to STPA and VSD as 'leading frameworks' to address challenges in cyber-physical social systems but do not explore how they might be combined in practice. Similarly, Rismani et al. (2023) use VSD to examine value tensions and STPA to analyze safety issues, yet apply them independently. Martelaro et al. (2022) also emphasize the usefulness of both frameworks for ethical AI engineering but treat them as complementary rather than integrated. These findings underscore a methodological gap in the literature: Despite growing interest in both VSD and STPA, their combined applications remain underdeveloped and unexplored.

Table 2.4: Search Results Combining Value Sensitive Design with System Analysis or STPA

Database	Search Terms Used	Results
System Analysis		
Google Scholar	”system analysis” AND ”value-sensitive design”	155
Scopus	(”system analysis” OR ”systems thinking”) AND (”value sensitive design”)	20
STPA or STAMP		
Google Scholar	”STPA” ”value sensitive design”	9
Scopus	(”STPA” OR ”STAMP”) AND (”value sensitive design” OR ”value-sensitive design” OR VSD)	0

Finally, Gerdes and Frandsen (2023) identified a standing-alone gap related to VSD. He conduced a systematic review of nearly three decades of VSD research identified a significant lack of technical involvement and implementation across VSD projects. Despite repeated calls from foundational scholars to integrate ethical considerations into design and to develop concrete technical solutions (Batya Friedman & Borning, 2013; Friedman & Hendry, 2019; van der Wilt et al., 2015), most VSD studies remain focused on conceptual and empirical inquiries, offering minimal technical reflection. More specifically, only a small number of studies result in the development of functional systems, prototypes, or implementable design interventions.

2.3.2. Summary of main research gap

In summary, the main research gap that this thesis aims to address is the lack of integrated approaches that combine system-level analysis with value-centered design to tackle misinformation about mental health on social networks. Although existing studies often examine either technical risks or user values in isolation, this thesis bridges that divide by applying and connecting System-Theoretic Process Analysis (STPA) and Value Sensitive Design (VSD) within a unified methodological framework. In doing so, it responds to the underexplored nature of mental health misinformation, particularly in low- and middle-income contexts like Brazil, and proposes a novel approach for developing interventions that are technically robust and ethically aligned.

3

Methodology

In this chapter, we outline the research strategy used to address the main question of this thesis. We begin by explaining the combined use of system theory and value-sensitive design (VSD) as a methodological foundation for identifying systemic risks and proposing interventions based on users' values. We then introduce the three phases of the research (conceptual, empirical, and technical), along with the methods and tools used to structure the research process and guide the development of the final recommendations.

3.1. Theoretical Frameowrk

This thesis is grounded in sociotechnical systems theory, viewing health communication on Instagram as a system shaped by interdependent and evolving elements, technologies, markets, policies, and cultural meanings (Fuenfschilling & Truffer, 2014). While the whole system is the focus, users are the unit of analysis.

The theoretical framework and methodological motivation of this study align with the scope illustrated in Figure 1.1. Grounded in socio-technical systems theory, the research investigates how misinformation operates within the mental health communication system on Instagram. To support this analysis, System Analysis is applied to map key interdependencies and mechanisms across the platform. More specifically, the study seeks to understand how misinformation is generated and spread and to identify the main sources of harm and risk within this complex system.

Bringing focus to the study's unit of analysis, the user, Value Sensitive Design (VSD) complements the system analysis by identifying what shapes user engagement. This methodological approach was incorporated to fully understand how users perceive risk and interact with the system based on their underlying values. Together, these approaches enable the development of interventions that are informed by the dynamics of the system and aligned with ethical considerations. The following sections outline how each method was applied in the study.

3.2. Methodological Framework

This thesis combines system-theoretic process analysis (STPA) with value-sensitive design (VSD) to address the main research question, bridging a methodological gap, as these approaches have rarely been systematically integrated. As Gerdes and Frandsen (2023) notes, VSD often remains at the conceptual level, limiting its impact on technical design. STPA helps address this by translating complex system dynamics into actionable

interventions.

Although VSD provides a strong user-centered lens, it can overlook structural constraints such as platform governance, economic incentives, or regulatory environments. Integrating it with STPA allows for the development of interventions that are both ethically grounded and systemically viable.

Systems theory offers a structured approach to mapping the complexity of misinformation. Furthermore, as highlighted by Vincent and Amalberti (2012), preventing harm in healthcare is essential to improve patient safety, which is a fundamental aspect of overall quality of care. Since safety is an essential component of the system (N. G. Leveson, 2011), this methodology is justified as it focuses on minimizing risks and harm, ultimately contributing to a safer information environment.

Meanwhile, value-sensitive design (VSD) will allow us to integrate the actor's and stakeholder's moral psychology into the system. As mentioned in Friedman and Hendry (2019), this methodology can complement traditional system design methodologies in security contexts by identifying key value tensions and ways to address them.

3.3. Value Sensitive Design (VSD)

According to Friedman and Hendry (2019), all technologies reflect and influence human values. Given that perspective, Friedman developed Value Sensitive Design (VSD), which integrates these values into the development of technology to ensure ethical and user-centered design (Friedman & Hendry, 2019). Since misinformation is not just a technical issue but also an ethical one, VSD provides a structured approach to designing solutions that align with user values.

VSD follows a three-part methodology, each with key objectives and investigation goals (Batya Friedman & Borning, 2013): *conceptual, empirical, and technical*. First, the *conceptual investigation* identifies the stakeholders involved and the values they hold, examining how these values relate to each other. In this context, values are not defined by economic value, but by the principles and ethical standards that shape the way users interact with technology (Simpson & Weiner, 1989). Second, the *empirical investigation* explores how these values manifest in practice, testing initial assumptions, and gathering insights from user experiences. Lastly, the *technical investigation* analyzes how existing technological features influence these values and explores possible design modifications to better align the system with user needs and ethical considerations. In social networks, the value-sensitive design approach helps assess social norms, user behavior, and complex value considerations within the platform itself (Friedman & Hendry, 2019; W. Hussain & Whittle, 2018).

3.4. System Theory and System Analysis

System theory was a response to the limitations of classic analysis techniques in dealing with increasingly complex systems (N. G. Leveson, 2011). As explained in N. G. Leveson (2011), this approach emphasizes the analysis of the system as a whole rather than the isolation of individual components. It recognizes that certain system properties can only be fully understood by considering the entire system, integrating both social and technical aspects.

Systems analysis applies scientific methods to analyze large and complex systems (Enserink et al., 2022). The purpose of this project is to map, analyze and structure these systems using an open, explicit, and empirically grounded approach. The key advantage of systems analysis is its ability to structure complex policy fields, making assumptions explicit and improving communication among actors (Enserink et al., 2022). Although it

may not offer definitive solutions, it helps eliminate ineffective alternatives.

3.5. Research Phases

This study integrates systems analysis and Value-Sensitive Design (VSD) to develop systemic recommendations to address misinformation about mental health on Instagram. The research is divided into three key phases, based on the goal proposed by Batya Friedman and Borning (2013): Conceptual, Empirical, and Technical. Each phase builds on previous insights. Table 3.1 presents an overview of the research design, detailing the objective of each phase, the desired results, the methods and techniques.

Table 3.1: Overview of Research Phases, Goals, Outputs, and Techniques

Research Phase	Conceptual	Empirical	Technical
Goal	Identify and analyze the risks and harms of mental health misinformation, understand how they emerge, and develop the system that will hold the final recommendations.	Explore how young adults perceive risks related to mental health misinformation and identify the values that shape their engagement with mental health content.	Develop systemic design recommendations informed by user insights, system boundaries, and mechanisms to mitigate risks and preserve user values.
Desired Outputs	- List of stakeholders and their values from the literature - List of risks and harms associated with mental health misinformation on social media - Explanation of how these risks emerge - Identified loss scenarios	- List of values identified in interviews that shape or justify young adults' engagement with mental health information online - Participants' risk perception and awareness levels in relation to potential harms in the identified loss scenarios	- List of systemic recommendations for mitigating misinformation based on user values and system safety - A normative system design proposal incorporating value-sensitive considerations and risk perception into safety control structures
Method	Literature Review	Value-oriented semi-structured interviews	Literature Review combined with inputs from interviews
Techniques	Stakeholder Analysis, Causal Diagram, STPA	Inductive Content Analysis and Thematic Analysis	STPA (second iteration)

3.6. Phase 1: Conceptual

The conceptual investigation will identify key stakeholders, the values they uphold, and how these values may interconnect or conflict within the context of mental health misinformation. This phase also includes a comprehensive literature review to map the risks and harms associated with such misinformation. Based on this analysis, we will develop causal diagrams to visualize how misinformation spreads and its potential consequences. The findings of stakeholder analysis, risk and harm mapping, and causal diagrams, together with the information from the literature review (Chapter 2) will collectively inform the development of the initial analysis of the systematic process (STPA).

During this conceptual phase, we will perform the first iteration of STPA to identify loss scenarios, which will

serve as the basis for the design of the interview protocol. The identification of losses, hazards, and constraints at the system level will be based on data from the literature review and the evaluation of how risks emerge within the system.

3.6.1. Stakeholder Analysis

The stakeholder analysis was conducted following the principles proposed by Friedman and Hendry (2019), to incorporate their values in the analysis. For this thesis, stakeholders are defined as individuals or entities significantly affected by the problem of mental health misinformation, either as those impacted by it or as those with the ability to intervene. In addition, stakeholders were classified according to their level of influence on decision making within the problem space.

In alignment with the Value Sensitive Design framework, stakeholders were identified in relation to their interaction with the mental health misinformation system. Their role is determined by their responsibilities and actions within the system. For example, platform users can be divided into those seeking mental health information and those sharing it. While these roles may overlap in practice, they represent distinct interactions with the system and require separate analysis.

Thus, stakeholder analysis focuses on the role of each group in the problem, their interests, expected outcomes, and potential challenges they may face in addressing or mitigating mental health misinformation.

3.6.2. Causal Diagrams

The Causal Diagram visualizes the cause-and-effect relationships among key factors contributing to the problem, illustrating how these elements interact and influence one another. This helps clarify the underlying dynamics of the system (Enserink et al., 2022). These diagrams were developed to support a structured understanding of the problem context and to serve as a foundational input for the System-Theoretic Process Analysis (STPA) presented in Chapter 4. In particular, the feedback loops identified in the Causal Diagrams were used to inform the analysis of risk emergence and to guide the hazard identification process in the subsequent phase of conceptual analysis.

3.6.3. System-Theoretic Process Analysis (STPA)

STPA (System-Theoretic Process Analysis) is a modern hazard analysis method based on an expanded model of accident causation that moves beyond traditional failure-based approaches (N. G. Leveson, 2011). In the context of this research, *accidents* refer to the emergence and occurrence of harm. STPA is built upon STAMP (System-Theoretic Accident Model and Processes), a theory that extends conventional models of causality (typically focused on linear chains of component failures) to include complex system processes and unsafe interactions among components.

This system theory views the system as a whole, emphasizing emergent properties. Emergent properties are outcomes that arise from interactions between components, not individually from components (N. Leveson & Thomas, 2018). These properties can only be fully understood by considering both technical and social factors.

The execution and application of STPA involves two main steps: identifying inadequate control actions that could lead to hazardous states and determining how these hazards might occur. Hazards are state systems that together with a set of worst-case environmental conditions will lead to accidents and thus losses.

The various components of a system are organized within a hierarchical control structure, as illustrated in Figure 3.1. This structure is made up of control loops in which a controller makes decisions to achieve specific goals and issues control actions to manage a process and enforce constraints on its behavior.

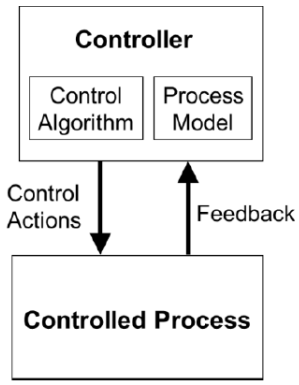

Figure 3.1: Generic control loop (N. Leveson & Thomas, 2018)

The controlled process can be any entity subject to control, such as a physical process or even another controller. The control algorithm reflects the decision making logic of the controller, determining which control actions to take in order to achieve the objectives of the system (N. Leveson & Thomas, 2018).

Each controller also relies on a process model, which represents its internal understanding or beliefs about the system (N. Leveson & Thomas, 2018). This model may include assumptions about the state of the controlled process, other components of the system, or the surrounding environment. Process models are continually updated, in part through feedback, which provides information about the current state and behavior of the controlled process.

According to N. Leveson and Thomas (2018) four conditions are required for a controller that is controlling a process. Those conditions will guide the development of our STPA analysis, the selection of hazards, and the design of the control structure itself. They are:

- **Goal Condition:** The controller must have a clear goal and enforce the safety restrictions associated with it.
- **Action Condition:** The controller must be able to influence the state of the controlled process through appropriate control actions.
- **Model Condition:** The controller must have or include a model of the controlled process to inform its decisions.
- **Observability Condition:** The controller must be able to monitor or receive feedback on the state of the controlled process in order to adjust its actions accordingly.

Identifying these conditions and assessing whether they are met is essential, as their failure can result in inadequate control, increasing the risk of accidents. Accidents often occur when the controller relies on an incomplete or inaccurate process model, leading to inappropriate control actions. This can happen, for example, when feedback is missing or delayed, preventing the controller from correctly interpreting the state of the system.

In this research, the motivation for applying STPA is to develop recommendations based on identified inade-

quate system scenarios. The purpose of describing loss scenarios is to ultimately propose changes that help mitigate hazards and prevent harm. To incorporate the perspectives of users, the scenarios initially identified will be explored during interviews to understand how participants perceive and respond to them. The general research flow is explained and illustrated in further detail in Figure 3.4.

3.7. Phase 2: Empirical

The second phase builds on identified risks, harms, and scenarios to explore user perspectives, expectations, and values. We will conduct value-oriented semi-structured interviews to understand how young adults perceive mental health misinformation on Instagram and the values shaping their interactions. Using a value-sensitive design approach, we will analyze the data to translate user preferences into values. The desired output includes an assessment of the awareness of young adults of misinformation and their perception of its potential effects on mental well-being. In addition, it aims to identify and compile a list of values that influence or justify their engagement with online mental health information.

Semi-structured interview questions are used to explore stakeholders' perspectives, beliefs, and values regarding a technology. These questions often focus on the evaluative judgments of the stakeholders (for example, whether they perceive the technology to be acceptable or not) and the reasoning behind their views (for example, why they hold those opinions) (Friedman & Hendry, 2019).

Since the goal of this research is to assess the risk perceptions of users, understand their level of risk awareness, and design appropriate interventions, risk perception was integrated directly into the interview structure. To achieve this, we used fictional scenarios to explore how participants interpret and respond to imagined situations, an approach inspired by Tang et al. (2024). This method allowed us to access the perceptions of the participants in a more contextualized and reflective manner.

3.7.1. Interviews Design

Semi-structured interviews provide a structured yet flexible approach to data collection, ensuring that key topics are covered while allowing participants to elaborate on the issues they find most relevant (Czeskis et al., 2010). Friedman et al. (2006) emphasize the importance of asking the respondents *why they hold certain opinions*, noting that the initial answers are often surface level. Follow-up questions that probe deeper tend to uncover the reasoning behind these views, revealing underlying values. This strategy will be applied throughout the interview process to explore the feelings of the participants, allowing us to access the value systems that inform their perspectives.

Each interview followed a three-part structure: (i) an engagement exploration section, (ii) a scenario-based reflection section, and (iii) a comparison and final reflection segment focused on potential solutions. The conversation started by discussing the general participation of participants in mental health content on Instagram before gradually introducing the issue of misinformation. This sequencing was intentional: many users may not initially recognize whether the content they see is accurate or misleading. In addition, actively seeking mental health-related content can itself increase exposure to misinformation, making it important to understand how users navigate and evaluate such information.

Table 3.2 represents the interview design summary that will be explored in more detail in this chapter.

Interview Part	Goal	Question Design Strategy
Part 1: Engagement Exploration	Understand user engagement with mental health content and reactions to misinformation.	Semi-structured questions based on daily use and perceived problems (Köhler et al., 2022; van der Wilt et al., 2015). Reflective probes like “Do you think it is all right?” and “Why?” (Friedman, 1997; Friedman et al., 2006).
Part 2: Scenario Reflection	Explore values and risk perceptions via fictional, relatable stories.	Reaction-focused prompts like “What do you feel about this?” (Tang et al., 2024), ”Do you think it is all right?” and “Why?” (Friedman et al., 2006).
Part 3: Comparison and Solutions	Compare perceived risks and explore user-suggested interventions.	Comparative and reflective prompts like “Which is riskier?” and “Why?” (Friedman et al., 2006). Seek for judgments to understand solution preferences (van der Wilt et al., 2015).

Table 3.2: Concise Summary of Interview Design

3.7.1.1. Part 1: Engagement with the mental health information system on Instagram

Part 1 of the interview aims to understand the participation of users in mental health information in general and assess how they feel about non-credible content using their experience. This section of the interview was inspired by the work of Köhler et al. (2022), who applied a value-sensitive design approach to uncover user values within a healthcare context. Their semi-structured interviews focused on participants’ daily routines and interactions, emphasizing the importance of asking why to reveal the underlying values behind their responses. We have split part 1 into two main categories: (i) engagement with mental health content and (ii) encountering non-creditable mental health information.

The first category aimed to explore why users engage with mental health content on Instagram and how they perceive its sharing across the platform. Drawing on van der Wilt et al. (2015), who emphasize the starting of interviews by identifying what participants find problematic, this section served as an entry point to surface general concerns. Techniques from Friedman et al. (2006), such as consistently asking whether something felt ’all right or not all right’, were used to reveal the underlying values.

The second category introduced the notion of non-credible information and examined how users recognize, interpret, and respond to misleading mental health content. This part also aimed to uncover factors that influence trust and decision-making. To support these reflections, we also incorporate question formats inspired by Friedman (1997), including ethical prompts such as ’Would it be all right...?’ to elicit deeper emotional and normative responses.

3.7.1.2. Part 2: Reflection on the actions of characters in fictional scenarios

Our goal in this phase of the interviews is to assess values and risk perception using the experiential model, which focuses on intuitive judgments at the intestinal level of young adults (Ferrer & Klein, 2015). To do this, we use scenario-based situations designed to elicit reactions and reflections from participants about the actions of fictional characters. This approach is inspired by the methodology used by Tang et al. (2024), which enables participants to engage in sensitive situations without having to directly refer to their own personal health history.

Two fictional scenarios are presented below, designed to be used in upcoming interviews. These were based on the STPA output performed in Chapter 4, the loss scenarios. The aim is to create a more personal and relatable context, helping participants connect with the situation and reflect on their own experiences or perceptions. To achieve this, we draw inspiration from Tang et al. (2024) and use the ChatGPT language model [1] to generate fictional narratives. A custom prompt was developed to guide the model, describing the intended

[1] ChatGPT is a conversational AI model developed by OpenAI based on the GPT-4 architecture https://chat.openai.com

goals, the core storyline, and key elements derived from the loss scenarios identified previously. The resulting narratives were then reviewed and refined by the researcher to minimize potential biases introduced by the model and to remove any elements that might unintentionally lead interviewees toward specific conclusions. The two final scenarios are presented below.

3.7.1.3. Scenario 1: Recommendation element

Scenario 1 was developed based on Loss Scenario 1 (Table 4.4). The main objective is to explore the perceptions of users of the recommendation algorithm, specifically in a situation where a user is repeatedly exposed to similar misinformed content and gradually begins to believe it.

User Scenario: Ana

Ana (21 years old) is a university student studying communications. She's been feeling increasingly stressed due to academic pressure and has experienced some episodes of anxiety, though she hasn't sought professional help yet. In her free time, Ana follows several influencers on wellness and mental health on Instagram.

One evening, while scrolling through Instagram during a particularly stressful moment, Ana finds a reel from a popular mental health page. The video suggests that anxiety can be 'cured naturally' by fasting and drinking lemon water every day. The reel has thousands of likes, shares, and comments filled with personal success stories.

Because Ana sees this type of content repeatedly, she begins to adopt the suggested practices instead of considering professional mental health support. Over time, her condition worsens, but the platform continues to show similar content, reinforcing her belief that professional help may not be necessary.

3.7.1.4. Scenario 2: Moderation element

Scenario 2 was developed based on Loss Scenario 1 (Table 4.5). The main objective is to explore the perceptions of users of the moderation algorithm, specifically focusing on the challenges and gaps in its effectiveness, which may contribute to the continued circulation of misinformed content online.

User Scenario: Gabriel

Gabriel (24 years old) recently started a new job in a design company. He feels that he has always struggled with focus: zoning out during long conversations and frequently putting off tasks until the last minute. To unwind between projects, he often follows Instagram.

One evening, a reel caught his attention: 5 signs you might have ADHD and you don't know it. The video feels surprisingly relatable, and Gabriel watches it all the way through.

In the weeks that follow, his feed becomes filled with similar content: creators sharing personal stories about living with ADHD, describing symptoms, and coping strategies. The posts are engaging and easy to connect with, often using humor, metaphors, and nonclinical language that feels more like a conversation than a diagnosis.

Over time, he starts to wonder: "Could I have ADHD too?" Without a formal diagnosis, he begins to adopt some of the tips and techniques he has seen: testing new habits, tools, and routines based on what others have shared online.

3.7.1.5. Part 3: Scenario Comparison and closing questions

After conducting the mock-up and training interview (Interview 1), we identified an opportunity to gather deeper insights by encouraging participants to compare the two scenarios directly. More specifically, our goal was

to understand which scenario they perceived as more risky or problematic and why. To support this, we incorporate comparison questions to assess what participants considered realistic and how they evaluated the relative severity of each scenario. Following the technique proposed by Friedman et al. (2006), we consistently asked participants why they felt a certain way, allowing us to uncover the reasoning behind their judgments.

In addition, drawing on van der Wilt et al. (2015), we asked participants what strategies they believed could resolve the problems presented to assess their solutions. This helped us better understand not only why they found certain situations concerning, but also how they envisioned a viable solution space. Although this was approached through open-ended questioning, it provided valuable input for the technical phase and informed the development of the final design recommendations.

3.7.2. Participant selection

Participants were selected using purpose-sampling to ensure relevance to the study focus. Selection criteria included: (i) being between 18 and 35 years old and (ii) using Instagram at least on a daily basis. We deliberately chose not to target users already engaged in mental health content in order to explore the perceptions of those who are less exposed to it on Instagram. Using scenario-based prompts, we were still able to gather meaningful insights from these participants, analyzing their reactions to hypothetical situations even if they do not actively interact with the mental health-related content on the platform.

The sample selection followed a nonprobability approach, as generalizability was not the main objective of this study. Initially, four participants were selected through convenience sampling (including the pilot interview) due to the time-sensitive nature of the research. To mitigate the risk of sampling bias and avoid limiting the pool to individuals directly connected to the lead researcher, snowball sampling was also used. Through this method, initial participants recommended others for the study, who then recommended other participants, helping to reach people from more diverse and potentially marginalized backgrounds. We acknowledge that this method still carries a risk of sampling bias, which will be discussed in more detail in the limitations section of the study.

All participants were based in Brazil to ensure contextual alignment with the focus of the study.

3.7.3. Interview Process

All interviews were held in Microsoft Teams, where both audio and transcription features were utilized to document the sessions in the Portuguese language of the local community. Following each interview, the autogenerated transcripts were cross-checked with the original recordings to verify their precision and ensure that the statements of the participants were faithfully captured.

Before starting the interview, participants were given the opportunity to ask questions about the consent form and research objectives. The interviews started only after all the questions had been fully answered and the consent form was signed correctly.

Given the semi-structured nature of the interviews, the phrasing and flow of questions varied slightly between sessions. In some cases, participants provided brief responses, prompting the use of follow-up questions such as 'Why?' to encourage deeper reflection. In other cases, the participants offered more detailed explanations without prompting. The sequence of questions also occasionally relied on prior responses. For example, if a participant stated that they had never encountered mental health information on Instagram, we would follow up with "Why do you think that happens?" rather than proceeding with questions about how they felt about

such content or whether they had engaged with it. This adaptive approach was particularly relevant for the questions from the first phase of the interview.

3.7.4. Assessing Data Saturation

Redundancy was the primary indicator used to determine that data saturation had been reached. In particular, with regard to participant risk perceptions and the values identified, it became evident that additional interviews were no longer contributing new conceptual insights. Although later participants offered unique personal examples or context-specific anecdotes, they did not introduce new themes or alter existing analytical categories. As such, the data collected were considered sufficient to support the objectives of the study.

	1 Interview 1 PO...	2 Interview 2 PO...	3 Interview 3 PO...	4 Interview 4 PO...	5 Interview 5 PO...	6 Interview 6 PO...	7 Interview 7 PO...	8 Interview 8 PO...	9 Interview 9 PO...	10 Interview 10...	Totals
	48	73	61	51	51	68	27	40	45	32	
Value Autono... 50	7	1	6	8	6	9	2	3	4	4	50
Value Integrity 55	5	10	6	6	6	4	1	3	12	2	55
Value Knowle... 93	5	20	14	15	4	7	1	7	10	10	93
Value Safety 54	13	4	6	1	11	9	2	4	2	2	54
Totals	30	35	32	30	27	29	6	17	28	18	252

Figure 3.2: Distribution of Value-Related Codes Across Interview Transcripts

Although the coding was performed after the interviews, Table 3.2 exemplifies the researcher's perception of saturation. The core value-related codes, *autonomy, integrity, knowledge, and safety*, emerged consistently among the participants, with the four values identified in the fourth interview. In subsequent interviews, no new value categories appeared; instead, later participants reinforced the same themes with additional examples. In particular, the number of coded segments decreased in the final interviews (e.g., interviews 7-10), indicating diminishing returns in new conceptual contributions.

3.7.5. Data Analysis

3.7.5.1. Initial reflections on data collected

During the interviews, it became clear that Research Questions 2 and 3 were not addressed neatly in separate sections, as originally anticipated. Although the interview design assumed that the first part would primarily uncover the values of the participants and the second part would focus on their perceptions of risk, the actual responses revealed a more fluid and overlapping dynamic. Participants often expressed their values while responding to fictional scenarios, and their discussions about Instagram use also shed light on how they perceived various risks.

In addition, additional themes emerged organically throughout the interviews. For example,participants frequently referred to the role of images and visual content in contributing to potential harm on Instagram, an aspect not explicitly covered in the original interview guide. These emerging topics were incorporated into the inductive coding process. To enrich the analysis, quotes that did not directly address the research questions were also coded, offering complementary information about the broader context of user experiences and perspectives.

3.7.5.2. Coding process

The code analysis focused on uncovering users' values and mapping their risk perceptions related to misinformation about mental health on Instagram. This process was supported by ATLAS.ti software and applied open, axial and selective coding, following the methodological steps described by Williams and Moser (2019). A visual summary of the coding process is presented in Table 3.3.

The qualitative data analysis coding process was carried out in two main phases using Atlas.ti, which supported the organization and management of codes, memos and comments throughout the analysis. The first phase involved inductive open coding, where the researcher worked closely with the interview transcripts to identify recurring themes and topics. This exploratory stage resulted in the creation of approximately 200 initial codes. Comments were added to each code to clarify the context and preserve the original meaning of the participants' statements. Following this, some codes were refined or merged to reduce redundancy and better reflect overarching patterns in the data. Code groups were also introduced to group related codes into broader thematic categories, facilitating a more structured overview.

In the second phase, the coding became more targeted, aligned with the study subquestions. For Research Question 2 (Values), the initial round of open coding provided a foundation for identifying values embedded in participants' narratives, such as *autonomy*, *safety*, *integrity*, and *knowledge*, along with less prominent values like *equality* and *empathy*, which were not prioritized due to limited depth. This was followed by a deductive coding phase in which the previously identified values guided a more focused search for associated norms and patterns. The objective was to refine the values-related themes by capturing both explicitly stated values and the underlying principles that shape user perceptions and behaviors.

For Research Question 3 (Risk perception), the analysis followed a slightly different approach. Although the initial identification of risk-related codes emerged inductively during open coding, the subsequent phase involved deductive grouping. These codes were organized into three overarching risk categories, aligned with the harms identified during the first iteration of STPA. This allowed for a more systematic and theory-informed analysis of how participants perceive and articulate potential risks related to misinformation exposure on Instagram.

Phase	Type of Coding	Related Research Question	Actions Performed
Phase 1: Exploratory	Inductive Open Coding	RQ2 and RQ3 (exploratory phase)	Read transcripts line-by-line to generate approximately 200 initial codes based on recurring themes. Added comments to clarify context. Refined and merged overlapping codes. Grouped related codes into broader thematic categories.
Phase 2: Targeted (Values)	Deductive Coding	RQ2 (Values)	Used value candidates identified in Phase 1 to guide a deductive search for related norms and behaviors. Refined themes to reflect both explicit values and underlying principles.
Phase 2: Targeted (Risk Perception)	Deductive Grouping	RQ3 (Risk Perception)	Organized risk-related codes from Phase 1 into three overarching risk categories, based on the predefined harms identified in the initial STPA iteration. Supported a systematic understanding of how users perceive and express risk.

Table 3.3: Qualitative Coding Process Across Research Phases

3.8. Phase 3: Technical

This final phase integrates insights from both the conceptual and empirical stages to develop systemic recommendations to address misinformation about mental health on Instagram. The goal is to ensure that the proposed interventions are not only technically sound, but also aligned with the values, motivations, and expectations of the users. The outcome is a set of actionable, value-driven recommendations that balance user safety with the dynamics of interaction with the platform.

A new iteration of the STPA analysis builds on the initial model by incorporating findings from the interviews. Losses, hazards, and system constraints are refined to reflect participants' risk perceptions, and hazards are

prioritized based on the values and concerns expressed by users. A revised set of loss scenarios then guides the final design of interventions, resulting in a normative safety control structure and a consolidated list of recommendations. This process, along with the connections between all phases of research, is visualized in Figure 3.4.

To structure the intervention design, an iterative framework was developed to align the four Unsafe Control Actions (UCAs) with the value and norm tensions identified in the study (summarized in Table 5.2). This framework is presented in Figure 3.3.

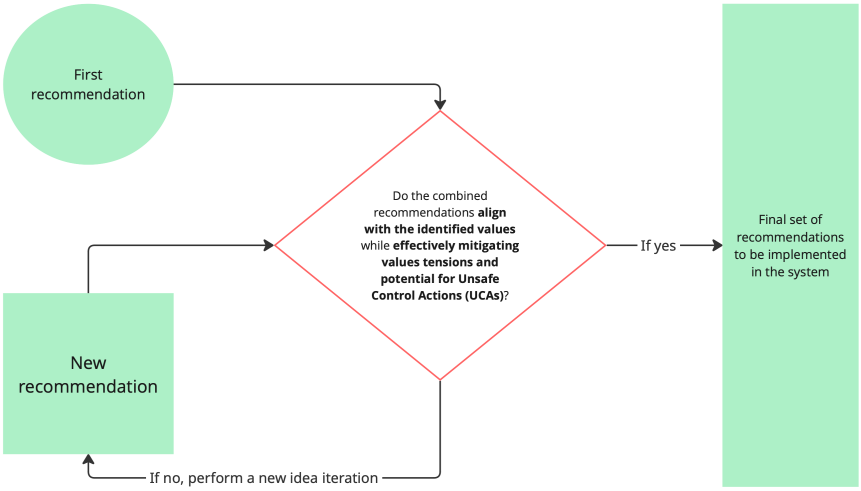

Figure 3.3: Iterative recommendation framework

The recommendations emerged through a stepwise process: initial concepts were drawn from interview insights and were iteratively refined to ensure alignment with core human values and effective mitigation of identified risks.

3.9. Research flow

To provide a clearer visualization of the research flow and the interaction between methods, we developed Figure 3.4. This representation outlines the intended outcomes of each phase and illustrates how they inform the next. In general terms, the literature review provides input for the initial STPA process, which then leads to the development of fictional scenarios used during the interviews. The insights gathered from these interviews are subsequently fed into a second iteration of STPA: a refined version of the original analysis, now incorporating empirical findings.

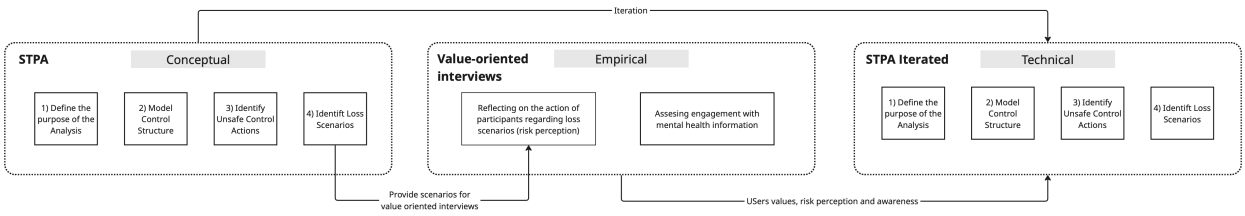

Figure 3.4: Research flow and interaction betweeen methods

4

Conceptual Phase

This chapter begins the conceptual phase of the research, in which we identify the systemic conditions underlying the spread of mental health misinformation on Instagram. The goal is to build a foundational understanding of how risks and harms arise by analyzing key stakeholders and the values they support. Drawing from an extensive review of the literature (Chapter 2), we map the key risks associated with misinformation and develop causal diagrams to illustrate how these risks are generated and amplified within the system.

This phase also includes the first iteration of Systems-Theoretic Process Analysis (STPA), which enables the identification of loss scenarios, system-level hazards, and initial safety constraints. These elements will serve as the basis for the design of the interview protocol in the following empirical phase.

4.1. Identifying the Risks and Harms

Risk can be understood as an unwanted event that may or may not occur (Rosa, 1998), while harm refers to an unwanted outcome involving a non-trivial level of damage or loss (Roeser et al., 2012). In this context, we define risk as the situation or condition that has the potential to lead to harm, whereas harm represents the actual negative consequence resulting from such a risk. According to Roeser et al. (2012), being safe from particular harm implies that the risk of that harm has been eliminated. As mentioned above, the goal of this research is to develop socio-technical recommendations to mitigate the main risks and harms associated with misinformation about mental health. To inform these recommendations, we began by identifying relevant risks and harms through a literature review of existing problems, effects, and consequences related to general health misinformation online.

Although there is still limited evidence on the long-term effects of false or misleading health messages distributed on social networks (Suarez-Lledo & Alvarez-Galvez, 2021), some risks associated with exposure to health misinformation have already been identified. According to Y. Wang et al. (2019), the most serious consequences include misinterpretation of scientific evidence, negative impacts on mental health, misallocation of health resources, and increased hesitancy in vaccination (which is not the focus of this study).

When it comes to understanding how people use this information to make health decisions, Janz and Becker (1984) argue that decisions are based on what people know about a health problem, such as its severity and their likelihood of being affected, as well as the perceived safety and effectiveness of recommended actions. If misinformation distorts what people understand about a health diagnosis or available treatments, it can lead

to unsafe or ineffective health decisions.

Furthermore, misinformation not only creates false beliefs about diagnostics or treatments, but can also reduce a person's willingness to seek or receive protective care (Nan et al., 2022). In addition to causing confusion and eroding trust in healthcare professionals, unfiltered exposure to misinformation can delay or prevent effective treatment, in some cases even putting lives at risk (Y. Wang et al., 2019).

Table 4.1 presents an organized view of the risks and harms identified in the literature. It is important to emphasize that these risks are not mutually exclusive; instead, they are interconnected and are often presented in the literature as a group of problems. Setting boundaries between those risks was thus a difficult task. However, it was possible to identify a key pattern in which risks and harms tend to fall into two primary categories: (i) those that affect individual health conditions and (ii) those that lead to broader social consequences. Thus, moving forward without analysis, we will dedicate ourselves to finding how individual and societal risks and harms emerge and spread.

Table 4.1: Risks and Harms of Mental Health Misinformation identified in the literature

Level	Risk	Harm
Individual	Self-diagnosis, wrong treatment, or self-treatment (Murthy, 2021; Starvaggi et al., 2024)	Worsening health, emotional distress, and confusion (Y. Wang et al., 2019)
	Delaying or avoiding professional treatment (Murthy, 2021)	Delay in proper care, emotional distress, or confusion (Y. Wang et al., 2019)
	Exposure to harmful content promoting self-harm or dangerous behaviors (Oksanen et al., 2016)	Self-injury, eating disorder, physical and emotional distress (Oksanen et al., 2016)
Societal	Increased use of unsafe or illegal "alternative" treatments on a public level (Suarez-Lledo & Alvarez-Galvez, 2021)	Collective health problems, public confusion (Oksanen et al., 2016; Southwell et al., 2019)
	Declining public trust in medical professionals and evidence-based care (Murthy, 2021; Southwell et al., 2019)	Reduced adherence to health advice, loss of trust in public authorities (Southwell et al., 2019; Y. Wang et al., 2019)
	Increased public stigma and reduced empathy for affected individuals (Saha & Banerjee, 2024)	Higher rates of self-harm, public discrimination (Saha & Banerjee, 2024)

Individual risks occur at the personal level, affecting users through exposure to unproven treatments or behaviors that discourage seeking professional help (Murthy, 2021). Health misinformation on social media has also been linked to increased risks of self-harm and eating disorders, worsening psychological distress (Oksanen et al., 2016).

Societal risks emerge when large populations are exposed to similar misinformation, leading to the widespread adoption of harmful practices. For instance, collective promotion of false treatments can normalize unsafe alternatives (Suarez-Lledo & Alvarez-Galvez, 2021). Indirect effects include growing mistrust in science and credible medical sources (Southwell et al., 2019), and even incidents of violence against public health workers communicating evolving health measures (Murthy, 2021).

Our review of the literature revealed that many societal risks associated with misinformation about mental health are derived from amplified individual risks. Understanding how misinformation spreads is essential to

identify systemic solutions, as introduced in the feedback loop in Chapter 1. Bodaghi et al. (2024) notes that false information spreads more rapidly and persistently than truth. This is not only due to user behavior, but is significantly intensified by social media algorithms, which exploit psychological biases to prioritize emotionally charged or morally provocative content (McLoughlin & Brady, 2024). Together, these dynamics create a feedback loop that fuels the spread of misinformation.

4.2. Stakeholders Analysis

To develop the list of stakeholders, a literature review was conducted to identify relevant actors involved or affected by the issue. The findings of the stakeholder analysis are presented in Figure 4.2. The role of each stakeholder aimed to explain their act in the system. The issue of interest was then developed according to the findings of the previous literature review. The divergence between the existing and the expected situation highlights the key obstacles preventing stakeholders from achieving their objectives. The Causes section identifies the root factors that contribute to these challenges, while the Solution section explores potential strategies to effectively mitigate these issues.

The information presented in the desired and existing situation causes and possible solutions columns are derived from assumptions grounded in the literature. These assumptions are aligned with the identified Values and Objectives to ensure consistency. Additionally, the solutions proposed for each stakeholder are tailored to their specific Values and Objectives to maintain relevance and feasibility.

4.2.1. Values Conflict

This section examines the value conflicts that arise from the differing priorities of stakeholders with respect to mental health misinformation on social media. By mapping these conflicts, we can better evaluate the dynamic of the system. The values previously outlined in the stakeholder analysis are integrated here to identify synergies and develop more realistic, well-fitted solutions. According to Friedman and Hendry (2019), the term 'value conflict' implies that different values can come into tension or opposition, but there is the potential to find solutions that balance or accommodate these competing values.

Figure 4.1: Values Conflict Diagram

Table 4.2: Stakeholders List

Actor	Role	Issue of Interest	Desired Situation / Objectives	Existing Situation	Causes	Possible Solution
Users	Seek mental health information on social media	Knowledge exchange, sense of well-being (Akhther & Sopory, 2022; Jia et al., 2021)	Easy and quick access to reliable information; Safety	Exposure to misinformation and harmful content	Lack of regulation and content moderation	Better content curation
Users, influencers, and health professionals sharing misinformation	Share mental health misinformation on social media	Engagement, freedom of expression, user growth (Marocolo et al., 2021)	Ability to speak freely and spread their beliefs	Spread of unverified or misleading information (Marocolo et al., 2021)	Lack of accountability	Encouraging ethical content creation
Presidency of Republic	Oversees national policy-making	Dignity, Progress and Individual rights (of Brazil, n.d.)	Balanced approach between regulation and digital freedoms	Difficulty enforcing regulations	Conflicting interests between health policies and digital platforms	Strengthening cooperation between entities to address misinformation effectively
Ministry of Health (MS)	Oversees national public health policies (Vijaykumar et al., 2022)	Public health accessible to all (da Saúde, 2020)	Social media accelerates the spread of evidence-based mental-health information (Bodaghi et al., 2024)	Proliferation of misinformation undermining public health efforts	Lack of resources to monitor online content	Collaboration with social media platforms to promote verified health content
National Health Surveillance Agency (ANVISA)	Regulates health-related products and content (of Brazil, 2025)	Consumer protection, medical safety, and transparency (of Brazil, 2025)	Online promotion of only regulated health products and treatments	Presence of misleading health claims and treatments	Insufficient monitoring of digital advertisements; Lack of enforcement of existing policies	Stricter penalties for misleading promotions
Brazilian Unified Health System (SUS) & Mental Health Reference Centers (CAPS)	Provides public mental health services and information (da Saúde, 2020)	Universal and free access to healthcare (da Saúde, 2020)	Population trusts public mental health services	Misinformation leading to mistrust and underutilization of services (Murthy, 2021; Southwell et al., 2019)	Spread of false information about public health services	Collaboration with fact-checking organizations
National Data Protection Authority (ANPD)	Ensures compliance with the General Data Protection Law (LGPD)	Digital privacy, data protection (G. F. do Brasil, 2018)	Protection of user health data	Platforms may collect and misuse sensitive mental health data	Weak enforcement of LGPD in digital spaces	Strengthening enforcement mechanisms

Continued on next page

Table continued from previous page

Actor	Role	Issue of Interest	Desired Situation / Objectives	Existing Situation	Causes	Possible Solution
Brazilian Internet Steering Committee (CGI.br)	Regulates internet governance and digital rights	Democracy, privacy, data protection, network neutrality (Bioni, 2021)	Balance between free speech and the need to curb misinformation	Challenges in regulating content while ensuring digital rights	Complex regulatory environment	Encouraging transparency in content algorithms
Medical Associations	Regulate professionals and advise on best mental health practices	Medical ethics, evidence-based treatments, professional integrity	Ensuring accurate mental health information is shared	Some professionals disseminate unverified information	Lack of continuous education and oversight	Stronger disciplinary actions against professionals spreading false information
Instagram Shareholders	Drive companies strategies and focus	Community building, user engagement, profit (Lemert, 2022)	High engagement, growing user numbers and advertisement on the platform	High user engagement but facing criticism and pressure over misinformation	Profit-driven algorithms that prioritize engagement over accuracy (Fernández et al., 2021)	Labeling misleading content.
Content Moderators	Enforce platform policies, flag harmful content	Ensuring compliance with policies (K. Wang et al., n.d.)	Effective moderation that reduces misinformation	Struggle to keep up with the volume of misinformation	Limited resources, unclear moderation guidelines	More robust moderation teams, clearer enforcement policies
Algorithm Developers	Design and refine recommendation algorithms	Platform efficiency, engagement optimization	Algorithms perform effectively	Algorithm prioritize engagement over accuracy (Fernández et al., 2021)	Resource allocation challenges	Adjust recommendation systems to tackle misleading content
Instagram Management Team	Oversees platform policies and strategy	Engagement, profit, reach (Fernández et al., 2021)	Balance between engagement and credibility	Facing regulatory scrutiny and public criticism over misinformation	Business model favors engagement over accuracy (Fernández et al., 2021)	Introduce accountability measures, improve transparency in policies
Private Companies and Advertisers	Use the platform to reach target audiences	User engagement, brand visibility, revenue (Domenico et al., 2021)	High visibility and user interaction	Risk of association with controversial or harmful content	Prioritize reach over content quality	Stronger brand safety policies to prevent ad placement on harmful content

4.3. Causal Diagram

After conducting a literature review and identifying the associated risks and harms, we will begin by analyzing three different layers of the problem. First, as mentioned in Chapter 2, the social media platform itself plays a central role in creating feedback loops and amplifying the reach and propagation of misinformation. Second, at the individual level, users experience a personal set of risks and harms. Finally, on a broader social level, these individual risks, when amplified by the platform, can escalate into widespread concerns that affect entire communities.

To build the diagrams, we first identified the factors for each of the systems and then the connections from the literature that support each link in the diagram. Separate diagrams were created for each of the three layers, with supporting justifications detailed in Appendix A.

To better understand how risks and harms emerge within the system, the key feedback loops identified in the causal diagrams are examined in more detail below. These loops offer critical insights into the underlying system dynamics and were foundational for informing the subsequent hazard analysis conducted using the STPA (System-Theoretic Process Analysis) framework. As N. Leveson and Thomas (2018) emphasizes, the identification of hazards requires the recognition of system states that can potentially lead to losses. In this context, causal loop diagrams serve as a valuable tool for visually mapping the systemic pathways through which harms can emerge.

4.3.1. Individual Level

The individual diagram represents how risks and harms manifest on an individual level. In other words, what are the potential causes and effects of mental health misinformation on social media on a personal level. It starts with the individual consuming mental health information (and eventually encountering misinformation). For harm to occur, the individual must partially or fully believe the false information, with susceptibility playing a key role. Susceptibility in the context of misinformation is influenced by cognitive factors, motivation to reason accurately, familiarity with the content, and social-affective influences (Nan et al., 2022). Short-term psychological effects include anxiety or emotional confusion caused by exposure to contradictory information. In more severe or chronic cases, long-term mental health deterioration can occur and in the worst cases lead to self-harm or suicide (Oksanen et al., 2016).

4.3.1.1. R1: Misinformation Reinforcement Loop – Short-Term

This loop captures how people's exposure to mental health misinformation increases their belief in false claims. This belief leads to behavior change, such as self-diagnosis or inappropriate treatment, which causes short-term psychological effects. These effects then increase the person's emotional vulnerability and susceptibility to future misinformation, strengthening belief, and sustaining the misinformation cycle on a personal level.

4.3.1.2. R2: Misinformation Reinforcement Loop – Long-Term

This loop builds on R1 but extends it over time. Misinformed behavior leads to short-term psychological distress and eventually to long-term deterioration of mental health. This decline makes individuals more likely to encounter and consume even more misinformation, creating a reinforcement loop where worsening mental health directly feeds back into increased exposure to and belief in misinformation.

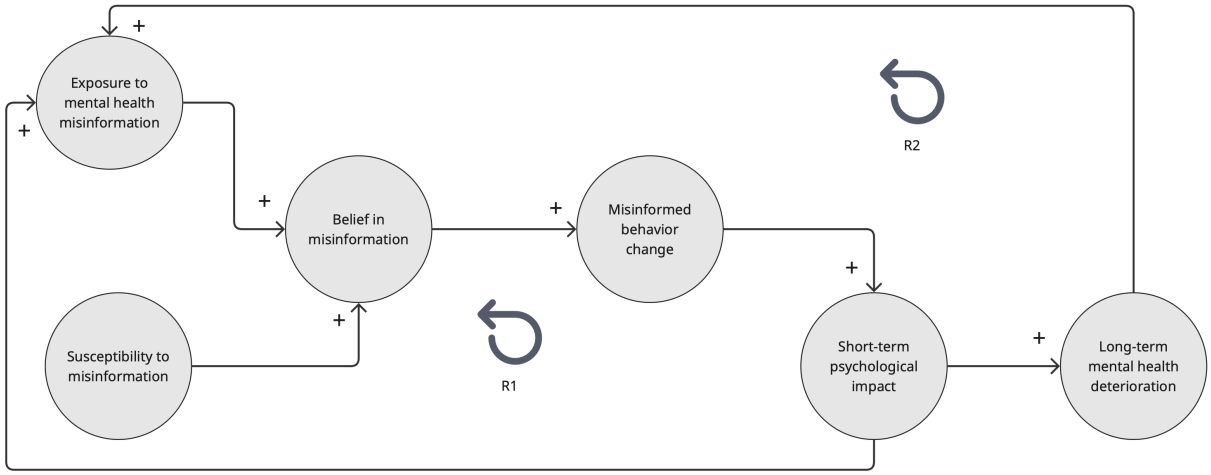

Figure 4.2: Causal Diagram of the Individual level

4.3.2. Societal Level

The societal diagram illustrates how risks and harms emerge at a collective level. Examines the cause-and-effect relationships of mental health misinformation when consumed by the public. Two key entry points are users sharing misinformation, which increases its visibility on the platform, and influencers, opinion leaders, and financial incentives that drive its spread. Furthermore, the literature review highlights evidence that health-related misinformation is financially incentivized, further amplifying its reach (Au et al., 2021; Guardian, 2025).

As with individual-level risks, the belief in misinformation plays a critical role in amplifying its harmful impact. Increased exposure to false content leads to a less informed public, contributing to poor health decisions. As individuals seek more health-related information, they are often exposed to misinformation, further strengthening the cycle. Research also shows that false information tends to generate more engagement than verified content, accelerating its spread across platforms (Murthy, 2021).

Once the public believes the misinformation, two possible consequences can emerge: public health consequences and societal and cultural consequences. Public health consequences include the overload of health systems, as people who adhere to incorrect treatment or no treatment at all can lead to worsening conditions and require emergency interventions (Suarez-Lledo & Alvarez-Galvez, 2021). It also pushes individuals towards the use of unsafe and alternative treatments, further escalating health risks (Suarez-Lledo & Alvarez-Galvez, 2021). On a broader level, social and cultural consequences include conspiratorial thinking and polarization (Murthy, 2021), which can lead to greater misinformation engagement and create another feedback loop.

Especially in the Brazilian context, where public healthcare is a constitutional right funded by the government, the social and health consequences of misinformation can increase costs and place additional strain on an already burdened system. The corresponding feedback loops were identified to structure the analysis and inform the subsequent STPA.

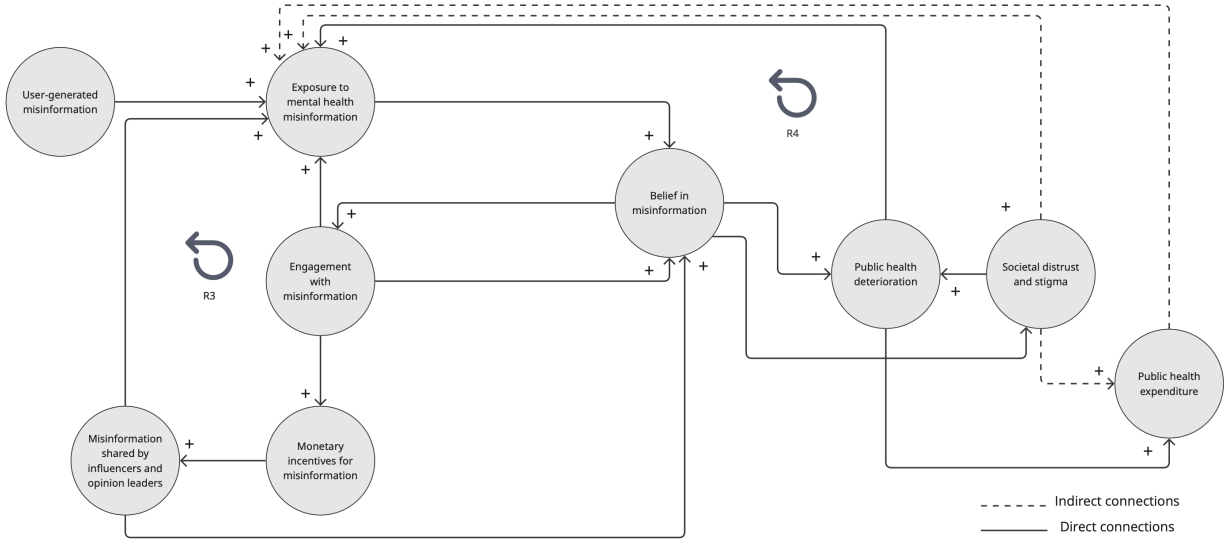

Figure 4.3: Causal Diagram of the Societal level

4.3.2.1. R3: Incentivized Misinformation Loop

This loop illustrates the economic drivers behind the spread of misinformation. As users engage with misleading content, platforms prioritize it due to its high engagement potential. This incentivizes influencers and opinion leaders to produce and share more of it, even with no intention of generating harm. As misinformation circulates more widely, participation increases, reinforcing its profitability and sustaining the cycle.

4.3.2.2. R4: Harm Reinforcement Loop

This loop shows how belief in misinformation can escalate broader social and public health issues. As more people accept false information, health outcomes worsen, leading to increased stigma and increasing distrust in institutions. This puts greater pressure on the public health system, driving up costs, and reducing its ability to respond effectively. In turn, this weaker response leaves the public more vulnerable to harmful misinformation, reinforcing the cycle.

4.3.3. Instagram level

The Instagram diagram illustrates how risks and harms emerge on the platform. This analysis focuses on the role of algorithmic mechanisms and content policies in the spread of false information. Two key entry points drive the dissemination of health-related content amplification: advertising and influencers and self-initiated searches (Tang et al., 2024).

The literature review also shows that false information generates higher engagement than verified content, which makes it more likely to be promoted by Instagram recommendation systems (Murthy, 2021). This creates a reinforcing feedback loop, where increased engagement leads to better advertising revenue, increasing platform profits, and further reducing the incentive to moderate misinformation (Domenico et al., 2021; Lemert, 2022). In addition, algorithm-driven amplification plays a central role. Since engagement-based ranking prioritizes content that draws users' attention, misinformation is often amplified faster and more widely than verified content, increasing user exposure (Fernández et al., 2021).

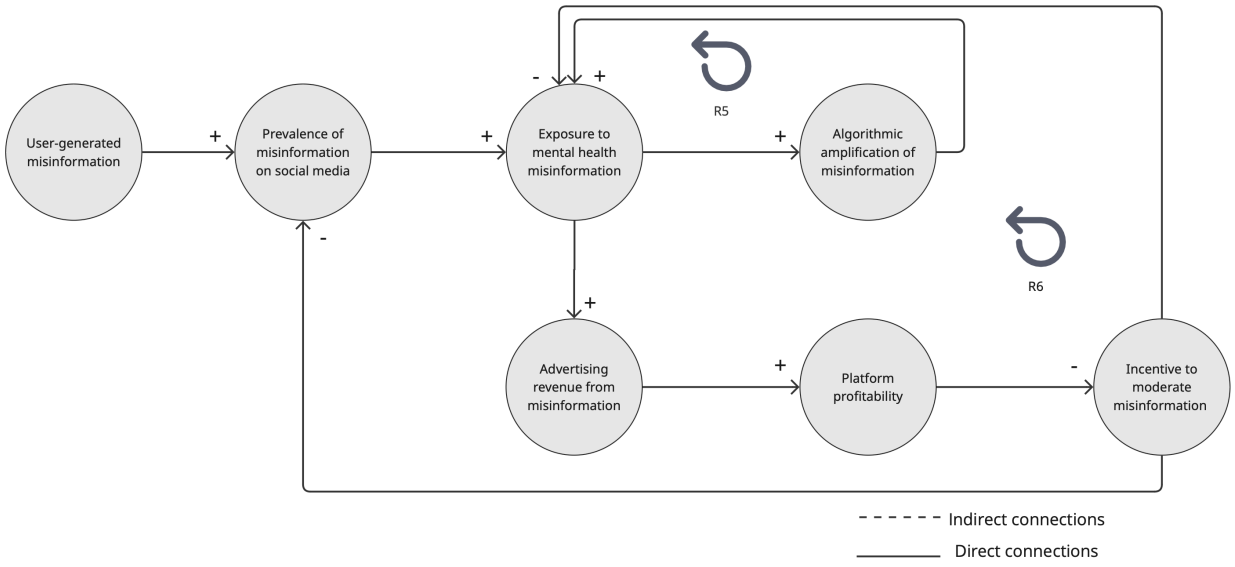

Figure 4.4: Causal Diagram of the Instagram Level

4.3.3.1. R5: Algorithmic Amplification Loop

This loop explains how platform algorithms amplify engaging content, even when it is misinformation. As users engage in mental health misinformation, algorithms interpret this as a signal of interest and promote similar content. This leads to even more exposure and reinforces what users see in their feeds, creating a self-reinforcing cycle driven by attention and engagement.

4.3.3.2. R6: Profit-Driven Moderation Erosion Loop

This loop illustrates how platform profitability can undermine content moderation. As misinformation attracts high engagement, it generates more advertising revenue, boosting platform profits. This creates little incentive to moderate harmful content, especially if moderation risks reducing profit. As a result, misinformation stays online longer, increasing exposure, and reinforcing the cycle of revenue-driven inaction.

4.3.4. Interpretation of the causal diagrams

Interpreting a causal diagram involves examining the factors and relationships that define the problem. The diagrams constructed at the individual, societal and platform levels illustrate how misinformation about mental health spreads, the risks it poses, and the mechanisms that reinforce its persistence. Across the three diagrams, several reinforcement feedback loops can be identified, each contributing to the amplification of misinformation and the emergence of harm.

At the individual level, loops such as the short- and long-term misinformation reinforcement cycles (R1 and R2) reveal how initial exposure to mental health misinformation increases belief in false claims, which in turn leads to changes in misinformed behavior. These behaviors can result in short-term psychological distress and long-term mental health deterioration. Specifically, these conditions increase the individual's emotional vulnerability and susceptibility to further misinformation, creating a self-inforcing dynamic in which harm and exposure mutually intensify.

At the societal level, the incentivized misinformation loop (R3) and the harm reinforcement loop (R4) show

how user engagement not only increases visibility of misinformation, but also makes it financially rewarding. The more people interact with misleading content, the more profitable it becomes for influencers and opinion leaders to continue to spread it. This increases both the reach and the perceived credibility of misinformation. As Marocolo et al. (2021) notes, this is especially concerning in Brazil, where even licensed health professionals sometimes act as influencers and share non-scientific claims. This cycle strengthens public belief in misinformation and weakens trust in health institutions.

At the platform level, the algorithmic amplification loop (R5) and the profit-driven moderation erosion loop (R6) further deepen the systemic nature of the problem. Recommendation algorithms are optimized to prioritize content with high engagement, and since misinformation often generates strong emotional reactions, it is consistently promoted. This increases user retention and advertising revenue, which in turn reduces the incentive of the platform to moderate harmful content (Lemert, 2022). As a result, misinformation is repeatedly amplified not because of its truthfulness but because of its profitability, leading to a structural misalignment between platform incentives and public health outcomes.

4.4. System-Theoretic Process Analysis

In this section, we develop the first System-Theoretic Process Analysis (STPA), which will later be iterated by incorporating input from the interviews. The STPA method relies on four key steps: (i) Define the Purpose of Analysis, (ii) Model the Control Structure, (iii) Identify Unsafe Control Actions, and (iv) Identify Loss Scenarios. All four steps are developed in this chapter using the literature and inputs from stakeholders' analysis and causal diagrams. The identified loss scenarios will serve as input to explore risk perception during the interview phase. The entire method will then be iterated in subsequent chapters to integrate the perceptions and values of the users into the analysis and final recommendations.

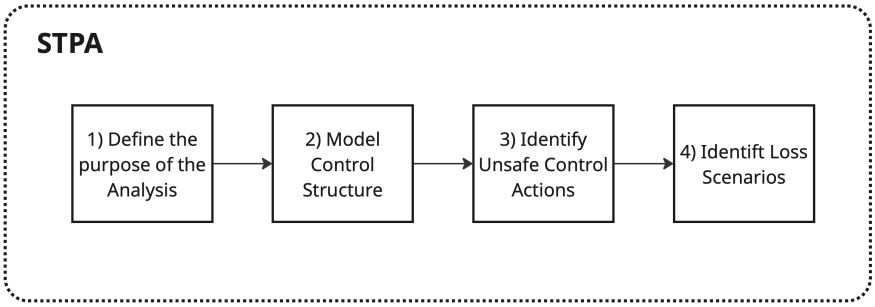

Figure 4.5: Steps of the System-Theoretic Process Analysis (STPA)

4.4.1. Define the Purpose of Analysis

An important part of STPA is to define the controllers and the controlled process. In our case, there could be numerous potential controllers that might expand the scope of this research beyond what is manageable or intended. Therefore, defining a specific set of controllers is necessary to narrow down the boundaries of the system and keep the analysis manageable.

Three main controllers or actors were selected to delimit this study figure 4.6. First, users represent those affected by it, consuming or sharing, and producing mental health content. Users are identified here as controllers because they control the content and engage with it, having some action on the controlled process.

Second, algorithms are non-human controllers capable of interfering in the process. As discussed earlier and supported by the literature, recommendation algorithms play a key role in spreading misinformation by strengthening feedback loops. Furthermore, algorithms are responsible for automated moderation on platforms such as Instagram (Willcox, 2025).

Finally, we include human controllers within the Instagram corporation, specifically developers and content moderators. These actors have a direct impact on how the platform operates and responds to content. For the purpose of this phase of the study, we have chosen to exclude higher-level managers and broader stakeholders within Instagram to maintain a focus on the functional aspects of the system.

By delimiting the limits of the system and identifying the risks and potential harms related to the proposed problem, the controlled process is defined as *Flow and visibility of misinformed mental health content on Instagram*. This formulation was intentionally chosen to capture not only the presence of misinformation on the platform but also the way it circulates, gains visibility, and is consumed by users.

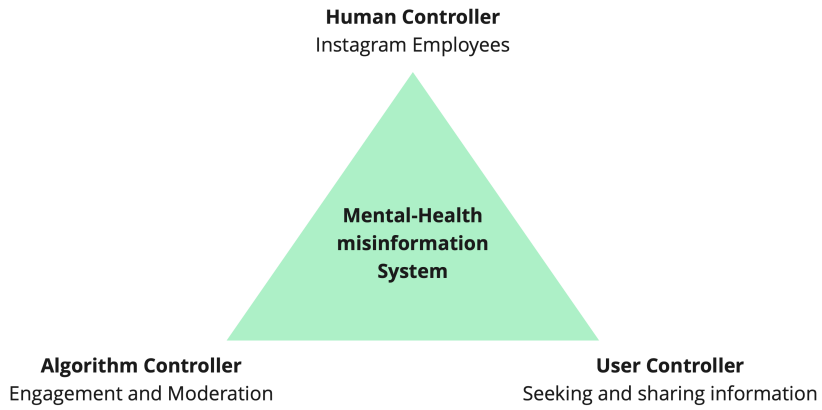

Figure 4.6: Controllers and system boundaries

4.4.1.1. Conditions

Controller	Goal Condition	Action Condition	Model Condition	Observability Condition
Instagram Team	Maintain platform engagement, user trust, and regulatory compliance	Design platform features, moderation systems, and policies (Witt et al., 2019)	Rely on internal data, reporting systems, and external inputs (Gorwa & Binns, 2020)	Receive feedback via user reports and platform metrics
Algorithm	Engagement optimization while applying further requirements (e.g., safety)	Recommends and moderates content based on engagement and predefined rules (Gorwa & Binns, 2020)	Operates based on machine learning models trained on historical data (Mishra & Sinha, 2021)	Detects patterns in content and user behavior to refine outputs (Y. Zhang et al., 2020)
User	Knowledge exchange, emotional support, and freedom of expression	Likes, shares, comments, and reports content	Builds mental models based on consumed content	Receives feedback primarily through likes, comments, and other platform reactions

Table 4.3: STPA Controllers Conditions in the Context of Mental Health Misinformation on Instagram

4.4.1.2. Identifying losses

According to N. Leveson and Thomas (2018), a loss refers to the deprivation of something valued by the stakeholders. After reviewing the literature on the risks and harms associated with mental health misinformation, we chose to focus on one loss to allow a more in-depth application of the STPA methodology. This loss was selected at the individual level for two primary reasons: (i) our unit of analysis is the user, making it more appropriate to concentrate on a loss that directly affects them; and (ii) we assume that individual-level losses, when experienced at scale, can accumulate and result in broader social consequences, thus enhancing the relevance and depth of this study. Based on these considerations, the following loss was selected as the focal point of this research:

> **Key Loss L1:** Decrease in mental well-being of users due to repeated exposure to misleading or distressing mental health content on Instagram.

This loss reflects several individual-level harms listed in Table 4.1 and specifically incorporates the feedback loop effect identified in the platform recommendation algorithm mechanisms. It was inspired more directly by the work of Fernández et al. (2021), who defends that ”recommendation algorithms have been pointed out as one of the major culprits of misinformation spreading in the digital sphere”. Most importantly for this thesis, recommendation algorithms filter the content users see, reinforcing a biased filter bubble in which individuals are primarily exposed to information similar to what they have previously liked or engaged with (Fernández et al., 2021). Although this concept was originally developed to explain general misinformation, it is applied here in the context of misinformation related to mental health.

4.4.1.3. Hazards

Following the identification of focal loss for this study, the next step involved defining the associated hazards. According to N. Leveson and Thomas (2018), a hazard is a state of the system or a set of conditions that, when combined with the worst-case environmental circumstances, can lead to an unacceptable loss. In this study, hazards were derived from the causal diagrams developed during the analysis phase. Specifically, they were informed by selected feedback loops that illustrate the systemic pathways through which these hazardous conditions can emerge. The corresponding loops linked to each identified hazard are described in the following.

- **H-1** Individuals act on unverified mental health information without professional validation, leading to inappropriate self-diagnosis or treatment. (Loop R1)
- **H-2** The recommendation algorithm prioritizes high-engagement content regardless of accuracy, increasing exposure to harmful misinformation. (Loop R5)
- **H-3** Financial incentives undermine content moderation, allowing harmful misinformation to remain online. (Loop R6)

The first hazard (H-1), linked to loop R1, reflects a failure of the model condition of the users. Individuals often build flawed mental health models based on online content. Without professional oversight or corrective feedback, they lack the observability needed to assess whether their self-diagnosis or self-treatment is appropriate, reinforcing a cycle in which misinformation-driven behaviors lead to further harm and increased susceptibility.

The second hazard (H-2), associated with loop R5, illustrates a mismatch of the goal conditions in the system.

The recommendation algorithm is designed to maximize engagement rather than promote safe or accurate content. This undermines safety goals and introduces a model condition, as the algorithm cannot reliably distinguish between credible and misleading mental health information, ultimately reinforcing exposure through algorithmic amplification.

The third hazard (H-3), driven by loop R6, reveals multiple control structure problems. An action condition appears when the moderation system fails to act or enforces ineffective measures. Furthermore, an observability condition exists, as both automated systems and human moderators, often under time pressure (Willcox, 2025), miss context-dependent harms. These systemic weaknesses are reinforced by platform-level incentives that deprioritize moderation when it conflicts with revenue generation.

4.4.1.4. System-level constraint

A system-level constraint defines conditions or behaviors that the system must maintain to prevent hazards and, ultimately, to avoid associated losses (N. Leveson & Thomas, 2018). Constraints may also specify how the system should act to minimize losses if a hazard occurs. In this study, a constraint was identified for each hazard.

- **SC-1 [H-1]** If individuals engage in self-diagnosis or treatment, the system must provide corrective information and guidance.
- **SC-2 [H-2]** Instagram's algorithm shall not prioritize or promote content unless verified by reliable health sources.
- **SC-3 [H-3]** The content moderation system shall detect and suppress harmful mental health misinformation within a predefined time frame.

4.4.2. Model the Control structure

The control structure model was developed to help define the boundary of the system. A hierarchical control structure represents the system through interconnected feedback control loops. When effectively designed, such a structure enforces constraints on the behavior of the overall system (N. Leveson & Thomas, 2018).

Figure 4.7 represents a hierarchical control structure that governs defined controlled process on Instagram, particularly within the system boundaries delimited previously. This diagram is a representation of the interaction between key controllers through control actions, feedback loops, and the central process being managed.

The development of the control structure followed a step-by-step logical reasoning process. First, we identified two primary controllers at the Instagram level that play a central role in content moderation. As highlighted by Willcox (2025), content management on Instagram is a hybrid process involving human judgment and automated decision making.

Second, we incorporated users into the control structure, both those who share content and those who seek information, acknowledging that these actions can occur simultaneously or in a sequence. To bridge the interaction between human moderators and algorithmic systems, we introduced algorithm developers as additional controllers, responsible for translating moderation objectives into code and shaping the behavior of automated moderation tools.

The content moderation process and the interaction between different controllers in our model were informed by the work of Willcox (2025). It is important to note that this represents just one method through which

content can be filtered on Instagram. In some cases, content moderation can be outsourced to users or outside institutions, who are encouraged to flag posts that they consider inappropriate.

To briefly explain the control structure, the information shared by users first passes through the content moderation algorithm, which determines whether it should be filtered or allowed based on platform policies. If the content is approved, it is then passed to the engagement algorithm, which decides whether to recommend it to other users, based on individual preferences and the moderation status of the content.

Above these algorithms, human controllers, such as content moderators and algorithm developers, work in coordination with Instagram management. Moderators apply content policies to guide moderation practices, while developers design and adjust algorithms according to the goals of the platform. Instagram management oversees the entire system by setting internal targets and using feedback from both moderators and developers to refine policies and strategic decisions.

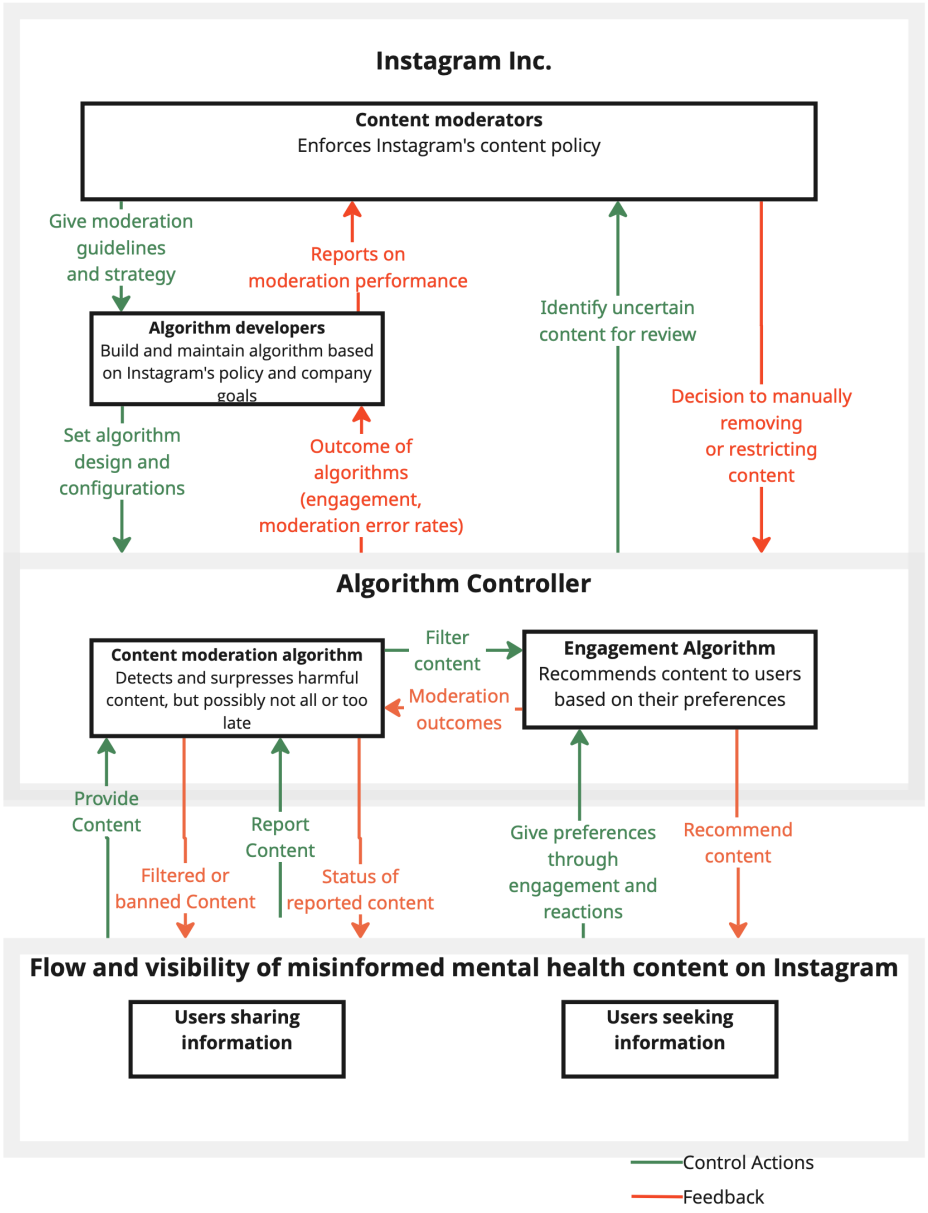

Figure 4.7: Model Control Structure

4.4.2.1. Safety Control Structure

We developed a Safety Control Structure to map out the actions, feedback loops, and control mechanisms between stakeholders. Although the Model Control Structure (Figure 4.7) will be the main focus of the recommendations in this study, the safety control structure provides a valuable context. It illustrates how the functional system is embedded within a larger environment by revealing the formal relationships between entities and stakeholders.

The Safety Control Structure was developed as an extension of the Model Control Structure. Additional controllers, control actions, and feedback loops were derived from our stakeholder analysis and literature review. For instance, insights from the Brazilian regulatory landscape, discussed in Chapter 1, informed the inclusion of control actions and feedback loops involving governmental entities and their interactions. Further non-governmental entities were incorporated based on the assumptions and findings outlined in the stakeholder analysis (table 4.2).

The idea behind a functional and safe control structure is that a downward reference channel ensures that the lower levels receive the necessary safety constraints, while an upward measuring channel provides feedback on their effectiveness (N. G. Leveson, 2011). This identified current dynamics of mental health misinformation in Brazil is represented in Figure 4.8. Although it can be a dense and complex representation, it is possible to take some insights from this representative structure.

One key observation is the lack of direct connections (safety restrictions or feedback loops) between Instagram and health-related government entities. Although some organizations focus on identifying and advocating against misinformation, and others enforce privacy laws, there are no clear, structured actions specifically targeting health misinformation. This gap in regulation leaves mental health misinformation largely unchecked within the system.

4.4.3. Identify Unsafe Control Actions

An Unsafe Control Action (UCA) is a control action that, under specific contextual conditions and worst-case environmental circumstances, can lead to a hazard (N. Leveson & Thomas, 2018). Identifying these control actions is a crucial step, as it lays the foundation for developing loss scenarios and formulating final recommendations.

We developed unsafe control actions (UCA) following the approach proposed by N. Leveson and Thomas (2018). In the model control structure (Figure 4.7), we examined the four ways in which a control action can be unsafe: (i) not providing the control action when it is needed leads to a hazard, (ii) providing the control action causes a hazard, (iii) providing a potentially safe control action, but at the wrong time (too early, too late) or in the wrong sequence, and (iv) allowing the control action to continue for too long or stopping it too soon (relevant for continuous control actions rather than discrete ones). The list of all UCAs identified related to the structured model control can be found in table B.1.

The following unsafe control actions (UCA) were prioritized for deeper analysis based on two main criteria: (i) their connection to multiple hazards and (ii) their relevance to the scope of this thesis and its bottom-up analytical approach. These UCAs were selected to guide the interview phase, with a focus on user perceptions of unsafe actions that occur within both human and algorithmic control domains.

The UCAs selected to move forward on loss scenarios are as follows:

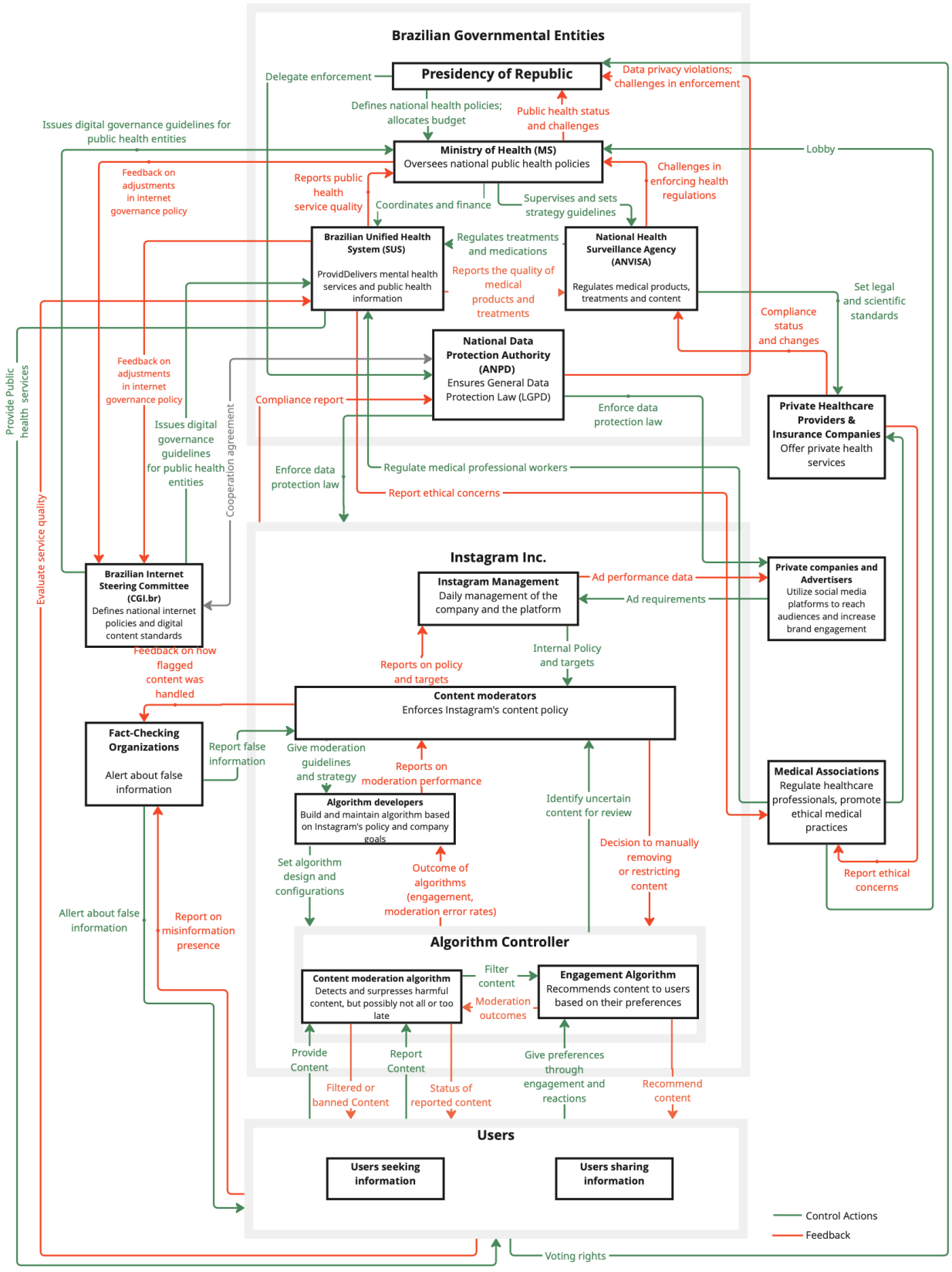

Figure 4.8: Brief Safety Control Structure

- **UCA-6:** The algorithm is configured to prioritize content based on engagement, including misinformation **[H-2]**.
- **UCA-13:** Harmful mental health misinformation is not filtered from feeds **[H-1][H-3]**.

Selecting the two UCAs above allows us to explore both user perception on the feedback loop around content recommendation (UCA-6 and H-2) and also explore their thoughts on content moderation and the potential harms around the lack of filtered content (UCA-13 and H-3).

4.4.4. Identify Loss Scenarios

4.4.4.1. Loss Scenario 1

Table 4.4 below presents a detailed loss scenario for one of the prioritized Unsafe Control Actions (UCA-6), focusing on how Instagram's engagement-driven algorithm can amplify mental health misinformation.

UCA	UCA-6: The algorithm is configured to prioritize content based on engagement, including misinformation
Hazards	H-2: Algorithmic amplification of harmful content
Scenario Type	(a) Why would Unsafe Control Actions occur?
Scenario	The Engagement Algorithm within Instagram's Algorithm Controller is designed to optimize for user engagement metrics such as likes, shares, and watch time. This approach is based on the assumption that high engagement is an indicator of content quality. However, it cannot distinguish between verified mental health information and harmful misinformation. As noted by Willcox (2025), the algorithm's moderation capabilities are trained primarily on normative and previously known content, which limits its sensitivity to medical context and prevents it from identifying posts that contain unverified self-diagnosis or treatment advice.
Consequence	Consequently, the algorithm may promote misleading mental health content that performs well in terms of engagement—such as posts offering "cures" for anxiety without medical backed. This amplifies misinformation and contributes to both unsafe self-diagnosis behaviors (H-1) and the broader algorithmic spread of harmful content (H-2).

Table 4.4: Loss Scenario for UCA-6

4.4.4.2. Loss Scenario 2

Table 4.5 presents a loss scenario for UCA-13, which focuses on the failure of platform moderation systems to effectively filter harmful mental health misinformation. It describes how technical and procedural limitations in content detection can lead to moderation gaps.

UCA	UCA-13: Harmful mental health misinformation is not filtered from feeds
Hazards	H-3: Gaps in content moderation
Scenario Type	(b) Why would control actions be improperly executed or not executed?
Scenario	The algorithmic content moderation system is expected to filter out harmful mental health misinformation. However, due to limitations in its training data and detection mechanisms, certain misleading posts remain undetected and are not flagged for removal. The filtering process relies heavily on keyword-based models and previously flagged content patterns, but new posts use euphemisms, emojis, and non-standard language that bypass these filters. Additionally, the system does not receive timely updates or input from external mental health experts or fact-checkers to refine its detection capabilities.
Consequence	This leads to the ongoing visibility and spread of mental health misinformation (H-2), contributes to unsafe self-diagnosis and treatment behaviors among users (H-1), and reflects a significant gap in the moderation system's effectiveness (H-4), increasing the risk of psychological harm (L-1).

Table 4.5: Loss Scenario for UCA-13

5

Empirical Phase

In this chapter, we delve deeper into the empirical phase of the research, focusing on the planning and execution of the interviews. The primary goal of these interviews was to capture the perceptions and values of the participants about online mental health information systems. participation with mental health information.

5.1. Participant's information

Table 5.1 provides an overview of the demographic and contextual background of the ten participants interviewed for this study. Participants ranged in age from 20 to 33 years and included a balanced representation of genders and professional backgrounds, such as public administration, law, education, and healthcare. Most of the participants had at least a bachelor's degree, two having completed or are currently pursuing a Ph.D. Instagram usage was predominantly daily, reflecting a high level of familiarity with the platform. The levels of engagement with mental health content on Instagram varied, with several participants reporting active interaction with such content, while others only encountered it passively or expressed little interest. This diversity in backgrounds and participation patterns allowed for a broad exploration of perspectives on mental health misinformation and platform dynamics.

Table 5.1: Participant Profile Overview

Interview	Age	Gender	Occupation	Education	Instagram Use	Mental Health Engagement
1	31	Female	Public Administration	Bachelor's	Daily	Occasionally sees
2	28	Female	Project Manager	Bachelor's	Daily	Occasionally sees
3	33	Male	Middle School Teacher	Bachelor's	Daily	Actively engages
4	30	Male	PhD Student (Phys. Ed.)	PhD in progress	Daily	Occasionally sees
5	24	Male	Lawyer	Bachelor's	Not currently using	Actively engages *(when in use)*
6	28	Female	Community Manager	Bachelor's	Daily	Actively engages
7	20	Male	Undergraduate Student	—	Daily	Not interested
8	32	Female	University Professor of Architecture	PhD	Daily	Actively engages
9	30	Female	Medical Doctor	Master's	Daily	Actively engages
10	31	Male	Public Administration	Bachelor's	Daily	Actively engages

5.2. Interview Findings

5.2.1. Values

Value	Description	Norm	Norm Description
Knowledge	Knowledge is traditionally defined as 'justified true belief' (Pritchard, 2008). It's prized because it gives people a solid basis for acting confidently and rationally in the world.	Accessibility	Social media ensures that reliable and diverse mental health content is easily accessible to all users.
		Scientific Accuracy	Mental health information must be evidence-based, citing credible scientific or medical sources.
		Immediacy	Users value short time and efficient solutions.
		Reflexivity	Users should use knowledge found on social media to regularly reflect on their mental health habits.
		Specificity	Shared mental health content should be contextually clear and not overly generalized.
		Understandability	Content should be presented in a language and format that is easily understandable to a non-expert audience.
Autonomy	Autonomy refers to the capacity for self-governance, enabling individuals to make decisions based on their own values and reasoning (Christman, 2009).	Explorability	Users should be able to explore related mental health topics and guidance freely.
		Controllability	Users must have control over the content they see and how they interact with it.
		Freedom of Speech	Users should be able to express their own mental health experiences without censorship.
		Self-resolveness	Users should be able to make personal health decisions based on online content, as long as those decisions do not replace professional diagnosis or treatment.
Safety	Safety can be understood as a material, emotional, and mental state in which relevant values or valuables will probably be preserved over a desired period. This probability must be supported by strong, credible knowledge, making safety not just the absence of harm but a justified expectation of continued well-being (Vandeskog, 2024).	Accountability	Platforms and content creators must be accountable for the impact of the content they share, especially if it poses harm.
		Confidentiality	Users' personal data must be protected from unauthorized access or misuse.
		Non-exploitation of Vulnerability	Content should not exploit emotionally vulnerable users for engagement, marketing, or influence.
		Social and Emotional Support	Social media should foster environments where individuals feel safe, supported, and treated with empathy.
		Transparency	Content should be presented with transparency on source of information and intention of message.
Integrity	Refers to a quality of a person's character (Cox et al., 2021). Integrity involves adherence to moral and ethical principles, ensuring consistency between one's actions, values, and beliefs.	Authenticity	Mental health content should reflect genuine experiences and not be fabricated or manipulated for influence.
		Professionalism	Mental health content should be presented with seriousness and professionalism.
		Social and Emotional Responsibility	Content creators should consider how their posts may emotionally affect others and take care to prevent harm.
		Self-honesty	Users should critically reflect on their own mental health experiences and avoid adopting labels or narratives that do not align with their actual needs or realities.

Table 5.2: Values and Norms Related to Mental Health Content on Social Media

The coding process described in Chapter 3 led to the identification of four distinct values that are directly related to our research question and topic. Specifically, values were derived through the grouping of related codes that reflect the shared underlying concerns, motivations, or priorities expressed by participants. Then each value category was analyzed according to the norms that support it. Table 5.2 presents the complete list of identified values and norms. The identified values and their definitions were developed based on the literature. The identified norms were taken from the interviews, and supporting quotation fragments can be found in Appendix D.

5.2.2. Values compared to participant's groups

In this section, we analyze how encoded values and associated norms are distributed between groups of participants according to key demographic characteristics, as described in Table 5.1. This comparative analysis focuses particularly on gender and educational level, to identify possible patterns in the way participants relate to the core values of *Knowledge*, *Autonomy*, *Integrity*, and *Safety*.

5.2.2.1. Analysis by gender

The first comparison focused on gender to explore how the values were distributed between male and female participants. In general, interviews with female participants produced a higher number of coded segments. Although the distribution of codes related to the values of *Knowledge* and *Autonomy* was generally similar between sexes, notable differences emerged in the values of *Integrity* and *Safety*, which appeared more prominently in the responses of female participants. Figure 5.1 presents these findings.

	Female 5 274	Male 5 222	Totals
Value Autonomy 50	24	26	50
Value Integrity 55	34	21	55
Value Knowledge 93	49	44	93
Value Safety 54	32	22	54
Totals	139	113	252

Figure 5.1: Distribution of User Values by Gender

To further explore patterns within the values of *Integrity* and *Safety*, we analyzed which specific norms were addressed more frequently by each gender. As shown in Figure 5.2, the norms of *Professionalism* and *Social and Emotional Responsibility* were more commonly emphasized by female participants, while *Self-honesty* was mentioned more often by male participants. Within the value of *Safety*, female participants expressed greater concern with *Transparency* and *Confidentiality*, whereas male participants focused more on the importance of avoiding the *Exploitation of Vulnerability*.

	Female 5 274	Male 5 222	Totals
Value Integrity: Authenticity 15	7	8	15
Value Integrity: Professionalism 16	12	4	16
Value Integrity: Self-honesty 9	3	6	9
Value Integrity: Social and Emotional Responsibility 17	13	4	17
Value Safety: Accountability 3	3		3
Value Safety: Confidentiality 7	5	2	7
Value Safety: Non-exploitation of vulnerability 18	8	10	18
Value Safety: Social and emotional support 12	6	6	12
Value Safety: Transparency 16	11	5	16
Totals	**68**	**45**	**113**

Figure 5.2: Norms on *Integrity* and *Safety* by Gender

Although the interview data did not provide conclusive evidence of a direct link between values and gender, the existing literature suggests a gender dimension to the value of *Safety*. As noted in Farr and Forsyth (2018), women tend to be more concerned about their personal safety and are more likely to take precautionary measures than men.

5.2.2.2. Analysis by educational level

Another relevant dimension of the analysis was the potential influence of the educational level on the values expressed by the participants. Although only 30% of the sample had a Master's degree or higher, the data still revealed notable patterns, as shown in Figure 5.3. It is important to note that two of the three participants in this group also had an academic background in health sciences, which may have influenced their value orientation and perceptions of mental health information.

	Educational Level: Bachelor 6 333	Educational Level: Master or above 3 136	Totals
Value Autonomy 50	33	15	48
Value Integrity 55	33	21	54
Value Knowledge 93	60	32	92
Value Safety 54	45	7	52
Totals	**171**	**75**	**246**

Figure 5.3: Distribution of User Values by Educational Level

A more noticeable difference appeared in the value of *Safety*. This can be attributed to the fact that participants in higher education, particularly those with a background in health sciences, when asked about how they assess credibility often mention that they rely on their training to distinguish between misinformation and reliable information. Although they expressed concern for the safety of others exposed to misinformation, they generally felt more confident and protected themselves due to their academic knowledge. The extract below illustrates this idea of concern for other's safety.

Interview 9

"Yeah, he's not seeing the context of things, and it is not Gabriel's fault, because Gabriel does not have the technical knowledge to understand that these are different things."
(Interview 9, participant, translated from Portuguese)

5.2.3. Values tension

Throughout the interview process, it became evident that some participants expressed conflicting perspectives, which were also reflected in the coding and value analysis stages. Figure 5.4 presents some of those tensions and relations between values and norms.

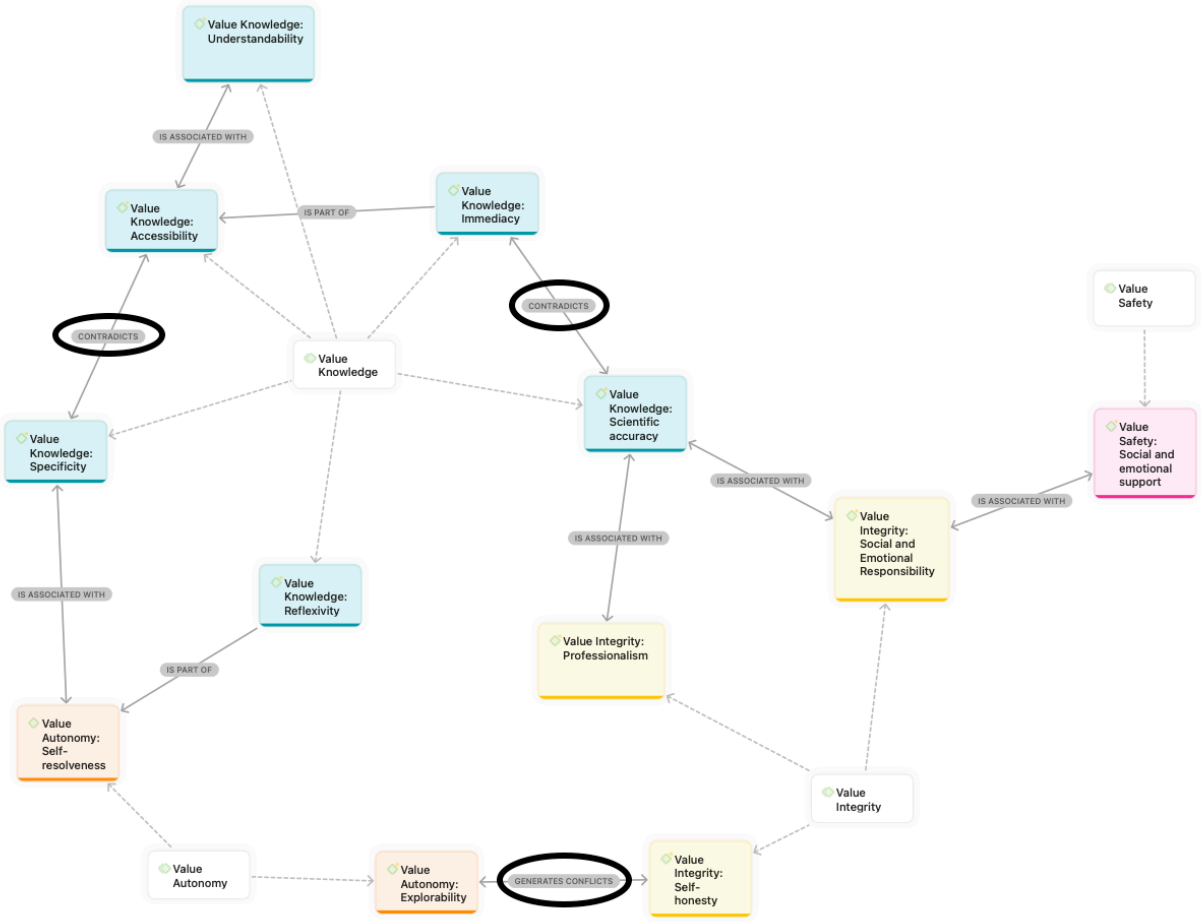

Figure 5.4: Values and Norms relations

To illustrate the opposing ideas and values, representative excerpts were selected and presented below according to the specific tension identified.

5.2.3.1. Immediacy vs Scientific Accuracy

The identified conflict is that participants prioritized the practical benefits of easily accessible mental health content on Instagram, even when it lacked clinical precision. They valued content that felt relatable and offered support at the moment. However, some interviews also highlighted impulsivity and immediacy as

problematic aspects, often seen as consequences of the use of social media. Although participants expressed a preference for accurate and evidence-based information, some still sought overly simplified advice, which could mislead users and potentially cause harm.

The following excerpts from the interviews further illustrate this dynamic. In Interview 8, the participant directly associated this sense of immediacy with the use of social media. This insight helped to establish a connection between Instagram's visual design and impulsive user behavior.

Interview 8

"I really notice this: students try to solve things, to want to solve them as quickly as possible. And the Internet gives a strong sense of immediacy, of everything being very fast. And that creates the feeling that they can solve things on their own and without any support."
(Interview 8, participant, translated from Portuguese)

Furthermore, the participant in Interview 2 emphasized that the visual nature of video content creates a sense of continuous flow and the need to continue moving forward. We interpret this as contributing to the development of an immediate, reactive behavior among users.

Interview 2

"That's kind of what social networks make you do all the time, right? If you like it, watch it; if you don't, just move on, I think."
(Interview 2, participant, translated from Portuguese)

The following excerpt further illustrates the perception of the participants of the consequences of these tensions between immediacy and precision on Instagram. The interviewee highlights that the content with the widest reach often prioritizes entertainment over information accuracy. As a result, essential mental health messages risk being overshadowed by simplified or even misleading narratives.

Interview 10

"The content that gains the most reach is often the one that relies more on entertainment techniques rather than truly appropriate information, right? What I mean is that sometimes an influencer ends up reaching more people than a doctor."
(Interview 10, participant, translated from Portuguese)

The oversimplification of mental health solutions on Instagram can foster the misconception that complex psychological conditions can be resolved through quick fixes or singular interventions. As illustrated in the excerpt below, one participant critiques this tendency by pointing out how users begin to seek'miracle solutions', such as assuming that medication alone can fully address the challenges of ADHD. This perception reflects how overly reductive narratives, often amplified by social media platforms, can distort public understanding of mental health care.

Interview 9

"And then? Then people start looking for miracle solutions, right? As if the treatment for ADHD, which involves medication, could solve everything in the person's life.
(Interview 9, participant, translated from Portuguese)

5.2.3.2. Accessibility vs Specificity

In some interviews, participants emphasized the benefits of social networks in making information more accessible, noting that it helps reach audiences who might otherwise lack access to mental health content. However, others raised concerns about the amount of generalized information on the platform, pointing out that the content is often presented in ways that can be misleading. This tension is reflected in the interviews in which social networks are recognized as both a valuable source of information and a platform that can distort complex issues. The following excerpts illustrate this conflicting perception through the own experiences of the participants.

Starting with the accessibility theme, many participants emphasized the importance of social networks in providing information to groups that might otherwise lack access due to economic restrictions or social barriers. The excerpt below illustrates a participant's recognition of how difficult it can be to access professional help, specifically within the context of the Brazilian mental health system.

Interview 9

"And another problem is that, although I believe professional help is essential, it is not accessible. Many people will not be able to see a psychologist or psychiatrist to care for their mental health. It is just not accessible, not because of lack of will, but because of the financial cost.
I think mental health care is still very financially inaccessible."
(Interview 9, participant, translated from Portuguese)

Other participants also emphasized the role of social networks in their learning journeys. The excerpt from Participant 6 below illustrates how discovering mental health information on Instagram can lead to further exploration through other media sources, such as podcasts and books. Although Instagram is recognized as a valuable starting point for accessing knowledge, it is not necessarily considered a reliable source on its own.

Interview 6

"So a lot of what I know about mental health—and various other topics on social media—came from there.
Like a book I found interesting, and then I went and bought it, same with podcasts.
It all started on social networks."
(Interview 6, participant, translated from Portuguese)

Interview 8

"Well, I think it is a source of knowledge, so I end up understanding and learning some things there. However, I always try to be critical, and whenever I find information that seems odd,
I look for a second source to understand if it is really trustworthy or not."
(Interview 8, participant, translated from Portuguese)

Some participants said they do not fully trust Instagram because the content is often too general or simplified. They felt that mental health topics are sometimes shown in short emotional posts that are easy to understand but do not explain things sufficiently.

Interview 10

"I think with social media we start oversimplifying a lot of things, you know?
Like, in 5 minutes she solved it. [The user] shared an experience that solved her problem.
And then you think—maybe I can solve mine too, right?"
(Interview 10, participant, translated from Portuguese)

5.2.3.3. Explorability vs Self-honesty

Although this tension was not explicitly stated in the data set, it emerged in specific interviews in which participants expressed that exploring alternatives was acceptable, but also noted that some individuals seemed to lose themselves in diagnoses and information. The idea that people are allowed to explore health alternatives that they see on social media is expressed in the fragments below.

Interview 6

"But I think it's appropriate to try some of the tips we see,
as long as they are not explicitly or implicitly harmful to us as individuals."
(Interview 6, participant, translated from Portuguese)

Interview 2

"So, I think it is nice that it has opened up the possibility for him to name a feeling he has,
and from there, investigate and understand it."
(Interview 2, participant, translated from Portuguese)

However, some participants expressed concerns about the risks of over-identification with mental health content. We labeled this tension *self-honesty*. The concern is that repeated exposure or strong personal identification with certain topics may lead to a distorted sense of self or the formation of a false identity. The following excerpt illustrates this perspective.

Interview 4

"So, many times you end up not being yourself, and often you use the tools or strategies you saw on social networks as a shack, when in fact you could continue improving one thing or another, but that does not necessarily mean using every routine someone with ADHD follows." "Now I have ADHD and that's it, the problem is solved. That explains why I can't study anymore."
(Interview 4, participant, translated from Portuguese)

In a more striking example of the tension around *self-honesty*, one interviewee shared an experience in which a friend had self-identified with ADHD symptoms. However, after seeking professional help, it was revealed that his symptoms were actually related to fatigue and emotional exhaustion. Rather than feeling relieved, the friend was dissatisfied with this outcome, as he had strongly identified with the diagnosis of ADHD and the narrative it provided.

Interview 3

"And then he found out that his problem was actually stress, and the psychologist told him that he probably has a lot of unresolved trauma, which is why he was repressing so many things; none of it had anything to do with ADHD.
And he was outraged. Really angry. He came to us completely furious."
(Interview 3, participant, translated from Portuguese)

5.2.4. Summary of Values Tensions

To summarize the value tensions discussed above, Table 5.3 organizes the key themes identified throughout the analysis.

Value Tension	Conflicting Values	Description
Immediacy vs Scientific Accuracy	*Immediacy:* users seek quick, relatable content *Scientific Accuracy:* users value evidence-based content	Participants appreciated content that was easily digestible and tips that were somewhat quick, even when it lacked clinical accuracy. However, some expressed concern that such content could promote impulsive behavior and oversimplified understandings of mental health.
Accessibility vs Specificity	*Accessibility:* users value open and inclusive access to information *Specificity:* users want detailed, context-sensitive explanations	Social media was seen as an important tool for democratizing mental health knowledge. At the same time, several participants were critical of overly general content, which they believed lacked the nuance necessary to support informed decision-making.
Explorability vs Self-Honesty	*Explorability:* users see value in trying out alternative solutions *Self-Honesty:* users recognize the need to critically assess their motivations and reactions	Some participants supported exploring alternative practices they found online. Others cautioned that this behavior could lead to over-identification with certain labels or routines, distorting self-perception.

Table 5.3: Summary of Value Tensions Identified in User Interviews

5.2.5. Risk Perception

The primary objective of this part of the analysis was to assess whether the risks perceived directly by users were aligned with those previously identified in Chapter 4, which were later translated into system-level hazards in the STPA framework. To do so, we first reinterpret the three previously defined hazards as concrete user-facing risks. We then examine the interview data to identify whether and how these specific risks appear. Table 5.4 translated the hazards into risks to facilitate comparison and the following analysis. The interview fragments used in part of the analysis are presented in Appendix E.

Hazard	Associated Risk
H-1 Individuals act on unverified mental health information without professional validation, leading to inappropriate self-diagnosis or treatment	Self-medication and self-diagnosis: Risk from self-medication and self-diagnosis
H-2 The recommendation algorithm prioritizes high-engagement content regardless of accuracy, increasing exposure to harmful misinformation	Mass misinformation spread: Risk from most engaged content not being scientifically accurate, amplifying the possibility of spreading misinformation
H-3 Financial incentives undermine content moderation, allowing harmful misinformation to remain online	Lack of content moderation: Risk from harmful or misleading content remaining visible on the platform

Table 5.4: Hazards and Their Associated Risks in the Context of Mental Health Misinformation

5.2.5.1. Risk One: Self-medication and self-diagnosis

The nuances of risk perception became clearer when participants were asked to compare the two cases. Scenario 1 was perceived as overall more harmful, as it had broader implications for general health. In contrast, scenario 2, which involved self-diagnosis and minor lifestyle changes, was considered less severe. Some even noted that the new habits could still have positive effects, despite being based on inaccurate information. This potential positive effect is supported by the fragment below.

Interview 3

"So, Gabriel still has a small chance of ending up in a positive scenario, although we tend to believe that he is more likely headed toward a negative one, because that is more probable. But in Ana's case, I could not see any chance of her moving towards a positive outcome, stuck in that loop.'
(Interview 3, participant, translated from Portuguese)

Furthermore, several respondents emphasized that anxiety, as depicted in the Ana scenario, could progressively deteriorate her mental health and potentially evolve into depression. In contrast, Gabriel's situation, a worsening of ADHD symptoms or a mistaken self-diagnosis, was not perceived as posing an immediate threat to his life. This differentiation in risk perception is supported by participant insights from Interviews 7 and 8.

Interview 7

"Because anxiety and depression share the same root.
So this anxiety can turn into depression. I find it very concerning.
It's a decision that could seriously harm her, especially her academic future."
(Interview 7, participant, translated from Portuguese)

Interview 8

"I don't know, I can't judge the difficulty involved, but in my view, depression and anxiety would lead to other, more drastic situations—potentially even death—more than just a learning difficulty."
(Interview 8, participant, translated from Portuguese)

5.2.5.2. Risk Two: Mass misinformation spread

When discussing algorithms-related risks, several participants identified echo chambers or content bubbles as a major concern. Although some acknowledged that being in a 'positive bubble' could sometimes be beneficial, the majority viewed these bubbles critically. Most of the participants emphasized the risks of being repeatedly exposed to similar content, which can distort perceptions and limit access to diverse perspectives. The following excerpts illustrate how the participants described these content bubbles.

Interview 3

"And that's where it gets complicated, because the social media algorithm traps you in a bubble, right? So if you start watching a lot of misinformation videos, you'll end up in a misinformation loop."
(Interview 3, participant, translated from Portuguese)

Interview 10

“We end up reinforcing our bubble based on the algorithm, right?
So you start watching something, and it keeps showing you more.
You start thinking that’s reality, because it filters based on what you want.”
(Interview 10, participant, translated from Portuguese)

Respondent 5 also observed that, particularly in the context of mental health, the content on their feed gradually became more distressing and emotionally intense over time. This illustrates a pattern of continuous exposure to an increasingly ‘negative’ content bubble.

Interview 5

”It was really gradual, I think, because sometimes we like things without thinking much - like, ’Oh, I liked that’ and you click quickly. It wasn’t from one day to the next, or even one week to the next. Over a long period, the posts I was seeing started to become heavier, actually over years. Sometimes I’d send something to a group of friends thinking they would relate, and they’d say: ‘Hey, don’t send that kind of stuff anymore, it’s really heavy, the vibe is too dark,’ and things like that.”
(Interview 5, participant, translated from Portuguese)

5.2.5.3. Risk Three: Lack of content moderation

The moderated content was not directly mentioned by the participants, but they did point to the absence of a clear and effective Instagram policy to address fake news, highlighting a perceived gap in the platform’s responsibility to mitigate misinformation.

Interview 1

”Social media isn’t very regulated, right? But you see movies, you see—even Instagram pages that will place something like: ‘This image contains triggers, something strong here’.”
(Interview 1, participant, translated from Portuguese)

Interview 6

”Instagram’s policy is basically nonexistent when it comes to fake news, right?”
(Interview 6, participant, translated from Portuguese)

The risk of content being created solely for marketing purposes was also raised. This concern is illustrated in Interview 9, where the participant, a doctor, reflected on how even health professionals must conform to marketing-driven strategies to remain visible on social media platforms.

Interview 9

"Alright, then you want to do the right thing. You're a serious professional. You're going to build a network to share your content, and that also touches on marketing.
And then it becomes about: is it relevant or not? Sometimes I want to say something, and the [social media advisor] company says no. Like, I wanted to mention my credentials. 'No, that's not necessary.' For me, that's the bare minimum, you know? Being able to say where I graduated. But they say it's not interesting to people, so it won't get reach.
You see? So everything ends up being driven by marketing, because it's a business. That's all the platform delivers."
(Interview 9, participant, translated from Portuguese)

In addition, the participants reflected on what they called 'Instagram logic', which are the strategic design choices that underlie the platform algorithm and the broader company decisions. Specifically, they highlighted the endless scrolling feature as addictive and potentially harmful. Participant 6 even recognized that this design was not incidental, but was intentionally developed to maximize user engagement, often at the expense of user well-being. This dynamic is further explored on the algorithm dimension in the following.

Interview 2

"I think that's kind of what social media makes you do all the time, right? Like, did you like it? Watch it. Didn't like it? Just move on, I guess."
(Interview 2, participant, translated from Portuguese)

Interview 6

"The Social Dilemma is a documentary, and it's really crazy to think that the people who were there at the beginning of Instagram and Facebook were studying mechanisms to alienate people.
They took research on how to make someone stay scrolling endlessly, up to the infinite feed.
Right? It was developed precisely to keep people there like a slot machine (...) So it was really designed to make people alienated."
(Interview 6, participant, translated from Portuguese)

5.3. Values and Risk Perception

We used the co-occurrence diagram (figure 5.5) to identify the intersections between user values and the perceived risks of mental health misinformation on social media platforms.

	Risk 1: Self-medication an... 49	Risk 2: Mass misi... 39	Risk 3: Lack of co... 17
Value Autonomy 50	3	2	1
Value Integrity 54	7	11	1
Value Knowledge 91	6	3	0
Value Safety 54	3	8	12

Figure 5.5: Code Co-occurrence Diagram

This analysis reveals that *Integrity* value is significantly associated with both Risk 1 (self-medication and self-diagnosis) and Risk 2 (mass misinformation). This indicates that participants often interpret the risks associated with self-diagnosis and unverified treatment advice as a consequence of a lack of honesty or ethical responsibility in the content being shared. In particular, the connection between *Integrity* and Risk 2 suggests that the pressure to produce content on a scale, which is often driven by the engagement of the platform and the marketing dynamics, can lead to distortion or oversimplification of mental health information. Participants expressed concern that these practices compromise the value of *Integrity*, as content creators can prioritize reach and popularity over scientific precision.

In contrast, Value *Safety* is most strongly associated with Risk 3 (lack of content moderation), reflecting the perception that insufficient moderation poses direct threats to the emotional and psychological well-being of the user. Meanwhile, Value *Autonomy* shows lower levels of co-occurrence across all risks, suggesting that while users value self-direction, they do not always frame it as directly connected to the risks in question.

Interestingly, the value of *Knowledge* did not show a strong association with any of the identified risks. This suggests that participants do not perceive the act of seeking or valuing knowledge as a direct contributor to the emergence of these risks. Instead, *Knowledge* appears to be regarded as a neutral or even protective factor, rather than a mechanism through which misinformation or harm is propagated.

6

Technical Phase

In this chapter, we build on the STPA analysis presented in Chapter 4 by incorporating insights derived from the interview data. It will present the refined version of the initial analysis by integrating user risk perceptions. Specifically, we use these perceptions to prioritize one of the identified hazards, allowing for a more focused and contextually grounded exploration of its underlying causes and implications. Furthermore, the values and corresponding norms identified in the earlier chapters inform key refinements and recommendations. Particular attention is paid to value tensions that emerged during the analysis, which are addressed to mitigate potential risks.

6.1. Refining the Purpose of Analysis

For this iteration process, and to support a more in-depth analysis, we maintain the system boundaries and conditions previously defined in Table 4.3.

6.1.1. Loss

Although validating the selected loss was not the primary objective of this thesis, the insights of the participants nonetheless provided implicit support for the identified loss (below) as a significant and feared consequence. The fragment of interview 8 below reinforces the relevance of this loss within the system analysis by reflecting genuine concerns about the emotional and psychological harm that can result from misinformation.

Key Loss L1: Decrease in mental well-being of users due to repeated exposure to misleading or distressing mental health content on Instagram.

Interview 8

”And then I don’t know—I can’t judge how difficult it is—but in my view, depression and anxiety would lead to other situations, you know?
Situations that are more drastic, even death. More than just a matter of having difficulty learning.”
(Interview 8, participant, translated from Portuguese)

6.1.2. Hazards

As presented in Chapter 5, participants identified several risks that align with the three previously defined hazards. In particular, their perspectives on Hazard H-2, which concerns the algorithm's prioritization of high-engagement content, introduced an important nuance. Rather than focusing solely on the system's preference for popular content, the participants emphasized how recommendation algorithms contribute to the formation of personalized echo chambers. The excerpt from Interview 2 below illustrates this concern, as the participant reflects on how individuals may come to distrust family members or professionals once they are deeply embedded in a belief system shaped by algorithmic exposure. Fragment from interview 10 summarizes a general belief, that the risk lies in the algorithm's tendency to reinforce content aligned with prior engagement. Over time, this recursive exposure was perceived to escalate into increasingly intense and emotionally charged content, something participants found particularly troubling in the context of mental health.

Interview 2

"Sometimes a family member or a professional will try to question it. But she is already so deeply immersed in that belief.
Because social media creates such a favorable environment for her to believe in it.
And she just goes along with it, doesn't she?"
(Interview 2, participant, translated from Portuguese)

Interview 10

"We end up reinforcing our own bubble based on the algorithm, right?
You start watching that kind of content, and it keeps showing you more.
You start thinking that's reality, because it begins filtering what you want to see."
(Interview 10, participant, translated from Portuguese)

Given the depth and consistency of these insights and the clear connections participants made between personalization, content escalation, and emotional impact, we chose to refine Hazard H-2 accordingly. This adjustment allows the analysis to better reflect the cumulative and individualized effects of algorithmic amplification as experienced by users. The revised hazard is presented below:

H-2 The system enters a state in which the recommendation algorithm reinforces personalized echo chambers, intensifying exposure to misleading content.

6.1.3. System-level constrain

Based on system-level constraints, we have refined SC-2 in light of the revised hazard and additional insights from the interviews. Some participants demonstrated awareness of the dynamics of the echo chamber on Instagram and, in response, had intuitively developed usage behaviors to mitigate this risk. These self-regulatory strategies suggest a degree of user adaptation to the platform recommendation patterns. The following excerpt from Interview 9 serves as an illustrative example and provides the basis for the refinement of this system-level constraint.

Interview 6

"So a lot of what I know about mental health—and various other topics on social media—came from there.
Like a book I found interesting, and then I went and bought it, same with podcasts.
It all started on social networks."
(Interview 6, participant, translated from Portuguese)

Interview 8

"Well, I think it is a source of knowledge, so I end up understanding and learning some things there.
However, I always try to be critical, and whenever I find information that seems odd,
I look for a second source to understand if it is really trustworthy or not."
(Interview 8, participant, translated from Portuguese)

Participant 8, who reported adopting a similar strategy of intentionally avoiding engagement with content that triggered emotional reactions or an impulse to respond, noted that such content eventually reappeared in her Instagram feed despite these efforts. For this reason, the analysis on top of a system constraint to prevent echo chambers is relevant and necessary.

SC-2[H-2] The recommendation system must periodically introduce contrasting or diverse content to disrupt echo chambers.

This safety control action alone introduces a new, not yet applied, recommendation to the system. In the following, we will explore how this can be operationalized.

6.2. Refine Unsafe Control Actions

Based on the selected hazard, H-2, we developed Unsafe Control Actions (UCAs) for all types of UCA. Table 6.1 presents the descriptions of these UCAs.

UCA Code	UCA Type	Unsafe Control Action Description
UCA-1	Not providing	The algorithm fails to introduce diverse or contrasting content when needed, allowing the echo chamber to persist.
UCA-2	Providing when unsafe	The algorithm introduces content that is too similar to previous content, reinforcing user bias.
UCA-3	Incorrect timing	The algorithm introduces diverse content too late, after the user has already been conditioned by repetitive exposure.
UCA-4	Stopped too soon	The algorithm stops introducing diverse content prematurely, failing to disrupt the echo chamber over time.

Table 6.1: Unsafe Control Actions for Hazard H2: Recommendation Algorithm Reinforces Echo Chambers

6.3. Designing recommendations based on values

In this section, we present a set of recommendations grounded in the previously established norms and values, with particular attention to the value tensions identified throughout the analysis. These recommendations were developed following the iterative process described in Figure 3.3, as introduced in Chapter 3.

6.3.1. Recommendation 1

As mentioned above, the first layer of recommendations was inspired by one of the interviews. At the end of each interview, participants were asked *In your opinion, is there anything that could be done to prevent this type of situation?*, and from this question we collected many creative solutions. The fragment that inspired this initial solution is presented below.

Interview 4

"I mean, the Instagram algorithm must be able to detect if, say, 80% of the content that a person views is about anxiety. So maybe there could be a strategy where, whenever we are dealing with something that could be a public health or mental health issue, there is an educational component. Like, for every five Reels someone watches, one would have to be from an entity that is competent to talk about the topic. "
(Interview 4, participant, translated from Portuguese)

Starting with the cascade process presented in Figure 3.3, we have derived the system-level constraint into a systemic solution, inspired by the above fragment. The first idea is presented in the box below and will be visually presented on the system further in this chapter.

Recommendation 1 – Echo Chamber Breaker: The recommendation system is supported by an "echo chamber breaker" mechanism. This dedicated algorithm operates in parallel with the primary recommendation engine and is designed to ensure that user receive specialized content is introduced within a defined time interval. Its function is to actively monitor user exposure patterns and intervene with content from verified institutions when repetitive content loops are detected, thereby disrupting echo chambers and promoting healthier informational environments.

In the following steps, we assess which of the identified values are supported by this initial recommendation. Since this recommendation was derived directly from the system-level constraint, it is not yet expected to mitigate unsafe control actions (UCA), as they were formulated based on the risks emerging from this very concept. Subsequent recommendations will build on this foundation and specifically address the unsafe control actions identified in the system analysis.

When considering the tensions of the values identified previously, this initial recommendation has the potential to alleviate the conflict between the norms of *Accessibility* and *Specificity*, both linked to the value of *knowledge*. Users are looking to have access to knowledge and information and do so through social networks. However, the design of social networks values generic and nonspecific information. By periodically introducing content from credible medical sources, the recommendation system can enhance *Specificity* without compromising *Accessibility*. Furthermore, this solution still enhances *Integrity*, as it promotes the dissemination of more authentic and trustworthy content.

6.3.2. Recommendation 2

The next phase involves expanding and refining the initial recommendation presented earlier. In this stage, we again draw on the insights from the interviews, particularly those aligned with the value of *Safety*. The following fragment presents inspirational aspects for this next recommendation phase.

Interview 1

"I remember, for example, some older movies, commercials, even soap operas... Eventually, people started to criticize the lack of care in the way certain scenes or situations were portrayed.
After those concerns emerged, more attention was paid to these issues and trigger warnings started to appear.
(Interview 1, participant, translated from Portuguese)

The idea presented by the person in the interview is that the type of media consumption on Instagram is less controlled and moderated than movies and commercials, specifically in terms of trigger warnings. However, our next idea is described below.

Recommendation 2 – Trigger warnings: The content moderation algorithm should ensure that any content identified as related to mental health includes a trigger warning, along with a message encouraging users to seek professional help if needed.

This recommendation has the potential to alleviate the tension between the norm of *Explorability*, associated with the value of *Autonomy*, and *Self-honesty*, tied to the value of *Integrity*. We argue this based on the expectation that trigger warnings when paired with a message encouraging users to seek professional help can interrupt the cycle of overidentification that several interviewees identified as problematic. By introducing a moment of reflection, this measure serves as a gentle reminder that, while self-exploration is valid, professional support remains essential to accurately assess and address one's mental health symptoms.

Additionally, Recommendation 2 may also contribute to reducing UCA-2, where the algorithm introduces content that is too similar to previous content. Even if repetitive content continues to appear, the inclusion of trigger warnings and prompts encouraging users to seek professional help ensures that this content is framed more responsibly.

6.3.3. Recommendation 3

This recommendation is based on an interview excerpt from participant 8. Her remarks suggest that, while Instagram offers verification for public figures and celebrities, it does not extend the same verification mechanism to doctors or health-related professionals.

Interview 8

"Yes, and you can't judge things on Instagram the same way you would with the verification badge for celebrities.
There's no verified account that truly indicates trustworthy health-related content."
(Interview 8, participant, translated from Portuguese)

Therefore, the idea is to initially expand this resource to only medical institutions. We recommend this exactly because, as identified by Marocolo et al. (2021), in Brazil false or misleading health information is widespread, with some cases even involving healthcare professionals as sources or distributors (Marocolo et al., 2021).

Recommendation 3 – Verified Health Institutions: Trustworthy and health-related institutions should be given verified status on Instagram. This would help users identify credible health content more easily and distinguish it from unverified or potentially misleading information.

This recommendation can serve as a complement to the initial proposal presented in Recommendation 1. Specifically, the content that is periodically introduced through the “echo chamber breaker” mechanism could be sourced from these verified health-related accounts. This strategy contributes to addressing the tension between the norms of *Immediacy* and *Accuracy*, both linked to the value of *Knowledge*. It offers users quick access to reliable evidence-based information, thereby preserving the efficiency of content delivery while enhancing its credibility. However, we did not observe a direct connection between this recommendation and the remaining unsafe control actions (UCA). As such, we proceed to the next phase of recommendation development.

6.3.4. Recommendation 4

This next recommendation is informed by interview excerpts related to the norm of *Controllability*, associated with the value of *Autonomy*. Participants expressed appreciation for having some degree of control within the platform, particularly when using features such as the ’I don’t want to see this type of content’ button. Building on this insight, we propose to expand this form of user control specifically to medical and mental health–related content, allowing users to more actively manage their exposure to sensitive or potentially distressing material.

Recommendation 4 – User-Controlled Health Content Filter: Users should have the autonomy to activate a ”Show only health-validated content” setting on Instagram. This feature would allow individuals to filter their feed to show only content verified by trusted health sources.

In this way, users gain direct control over the new algorithm proposed in Recommendation 1. By activating the filter, the algorithm is manually configured to prioritize and deliver only content from verified health sources. To balance user autonomy with platform business interests, this feature could be designed with a defined time window, remaining active for a limited duration before requiring reactivation. This approach maintains user controllability while reducing potential conflicts with Instagram’s engagement-driven content strategies.

Recommendation 4 directly addresses the issues outlined in UCA-1, UCA-3, and UCA-4. By granting users the ability to activate the system manually, it ensures that diverse and verified content is introduced when it is otherwise not provided, mistimed, or prematurely discontinued. Additionally, this recommendation helps mitigate the tension between the norms of *Immediacy* and *Accuracy*. For users seeking quick and accessible content, the feature offers a fast and straightforward way to access reliable information.

6.4. Summary of Recommendations

A summary of all recommendations and their relationship to value tensions and Unsafe Control Actions (UCAs) is presented in Table 6.2. In particular, the fourth recommendation is the most comprehensive in addressing UCAs 1, 3, and 4. This is understandable, as it introduces a new control mechanism that enables users to directly influence the behavior of the recommendation algorithm. As N. G. Leveson (2011) argues, preventing future accidents requires a shift from simply avoiding component failures to designing and implementing control structures that actively enforce system constraints.

However, this user-level control acts as an additional controller that operates the recommended algorithmic echo chamber breaker. According to N. G. Leveson (2011) systems-theoretic perspective, the introduction of multiple controllers with overlapping or poorly coordinated responsibilities may compromise performance. Therefore, while the recommendation improves autonomy and safety, we recommend that user-initiated control be given priority over automatic triggering of the algorithm.

Mitigated Element	Rec. 1 *Echo chamber breaker*	Rec. 2 *Trigger warnings*	Rec. 3 *Verified institutions*	Rec. 4 *User-controlled filter*
UCA-1: Not providing				✓
UCA-2: Providing when unsafe		✓		
UCA-3: Incorrect timing				✓
UCA-4: Stopped too soon				✓
Immediacy vs. Accuracy			✓	✓
Accessibility vs. Specificity	✓			
Explorability vs. Self-honesty		✓		

Table 6.2: Summary of Recommendations and Their Alignment with UCAs and Value Tensions

6.4.1. Assessment based on values

Our analysis was initially motivated by the identification of value tensions, with a focus on prioritizing their mitigation. To ensure that the proposed ideas and recommendations do not introduce new tensions between the identified values and norms, we evaluated how each recommendation aligns with and impacts these values. This assessment was guided by previously identified norms, which allowed us to critically reflect on whether the interventions support or potentially conflict the principles expressed by the users throughout the study. Table 6.3 summarizes how each value interacts with each recommendation.

Then it was possible to perceive that there is a specific norm and to be more affected by the recommendations and suggestions, and that *understandability* within the value *knowledge*.

6.5. Implementation of the Recommendations in the System

To help visualize how the recommendations fit into the system we previously designed, and also to show how they interact with each other, we created the diagram shown in Figure 6.1. This new version of the system includes new controllers, as well as new control actions and feedback loops. To make the diagram easier to understand, the original elements of the system are colored gray and each recommendation is represented in a different color.

Starting from the top of the diagram, Recommendation 2 introduces a new goal condition for content moderators that reinforces safety: the requirement for trigger warnings on mental health-related content. This goal is passed on to the algorithm developers, who are responsible for translating it into product specifications and implementing it within the content moderation algorithm. Once this mechanism is in place, content that contains trigger warnings is processed and distributed by the engagement algorithm as part of the regular content flow.

Moving further down the diagram, we see the introduction of a new controller: the account verification team. Although this is not a new concept on Instagram, verification is already in place for public figures and celebrities (Instagram Help Center, 2024). The same process can be extended to trusted health-related institutions. This would signal to users that the account is supported by qualified scientific or medical professionals, thus improving the credibility and visibility of evidence-based health information on the platform. This account verification team receives the user and institution requirements and feedback on the decision to be verified or not.

Value	Rec. 1	Rec. 2	Rec. 3	Rec. 4
Recommendation Description	*Echo chamber breaker*	*Trigger warnings*	*Verified institutions*	*User activates content filter*
Knowledge	Promotes *knowledge* by providing users with more *accurate* content over time. Content from scientific sources tends to be more *specific*, although this may impact *understandability*.	Trigger warnings might impact *reflexivity*, as content initially perceived as triggering may discourage engagement but could still promote reflection.	The inclusion of verified institutions enhances *accessibility* to *scientifically accurate* content. However, content based on medical language might reduce *understandability*.	By allowing users to activate a filtered view, this enhances access to *accurate* content in an *immediate* and user-controlled manner.
Autonomy	Does not directly engage with *Controllability*, as the system acts automatically.	May be seen as restrictive if users cannot opt out of trigger warnings and might impact *freedom of speech* when the user's objective is sharing content.	Might contribute to users' *explorability* when they decide to implement or test recommendations that come from verified sources.	Strongly supports *Controllability* by giving users the option to curate their experience and limit exposure to unverified content.
Safety	Indirectly contributes to *Safety* by breaking potentially harmful exposure cycles, more specifically by not exploiting users' vulnerabilities.	Enhances *transparency* by alerting users to sensitive topics.	Promotes *Safety* through verified sources that tend to be more responsible and *accountable* with the content they produce.	Strengthens *social and emotional support* by giving users access to content designed to offer care and support.
Integrity	Reinforces *Integrity* by curating authentic and evidence-based content from credible sources.	Supports *Self-honesty* by interrupting over-identification and encouraging users to seek external validation.	Strengthens *Professionalism* in health communication through visible verification.	Does not directly engage with the value of *Integrity*, as it centers more on user autonomy.

Table 6.3: Summary of Recommendations and Their Alignment with Core Values and Norms

The health-verified accounts will serve as input for the controller introduced in Recommendation 1: the echo chamber breaker algorithm. This algorithm is designed to detect repetitive content consumption patterns, indicative of an echo chamber, and periodically introduce reliable science-based content to disrupt these loops. In this context, "credible content" is defined as information originating from verified health-related accounts. Once the algorithm determines that a user's feed requires intervention, it issues a control action to the engagement algorithm, instructing it to prioritize and deliver specific verified content to that user.

If a user continues to feel that their Instagram feed is dominated by questionable or untrustworthy content, they can manually activate the echo chamber breaker algorithm. This action would trigger the system to intervene and deliver a new set of scientifically validated information, offering users more control over the content they receive. This is the last set of modifications to the system, represented in Recommendation 4.

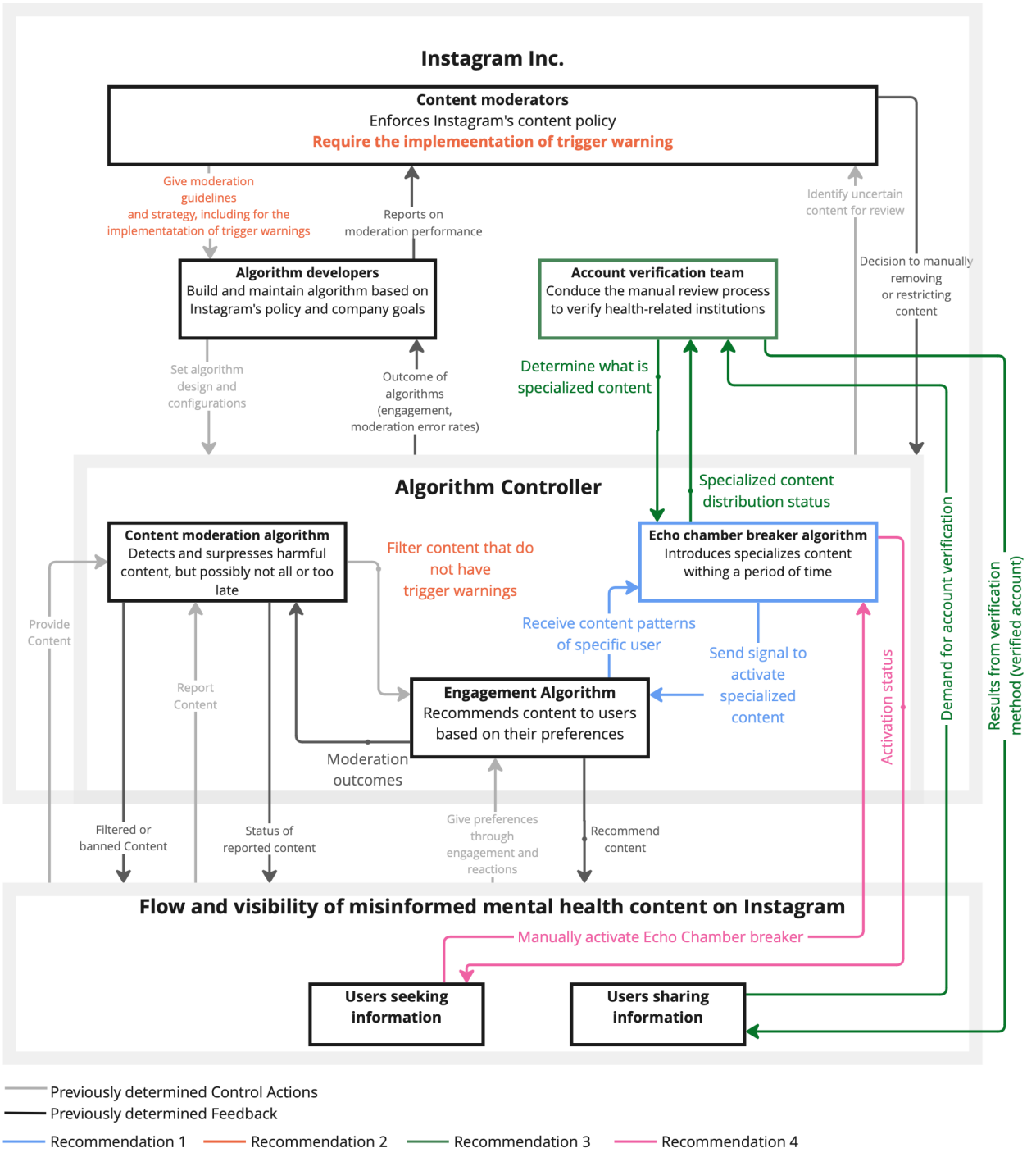

Figure 6.1: Implementation of the Recommendations in the System

6.6. Relevance and Feasibility

To assess the feasibility of the proposed solution space, we draw on existing studies that have explored or implemented similar interventions on social media platforms. These studies offer empirical insights into user behavior, algorithmic impacts, and content design strategies, providing both validation for the proposed recommendations and guidance on potential implementation challenges.

Starting from Recommendation 1 (Echo chamber breaker), a study conducted by Fernández et al. (2021) shows that popularity-biased recommendation algorithms tend to amplify misinformation over time. Therefore, an echo chamber is relevant because it aims to disrupt this cycle by introducing diversity into the content feed. In addition, the study finds that the recommendation mechanisms can be tuned to reduce the spread of misinformation by using algorithms that avoid popularity bias, such as user-based filtering (UB) instead of matrix factorization (MF).

To better assess the relevance and feasibility of Recommendation 2 (Trigger Warnings), an additional round of literature searches was carried out, as this term and its application had not emerged during the initial screenings. We then identified that recent studies on health misinformation indicate that warning labels' design, particularly in terms of color, assertiveness of language, and concreteness of message, can significantly influence user behavior, increasing the likelihood of protective actions such as information verification or reduced sharing (Varzgani et al., 2021). Drawing on fear appeal theory and construal level theory, the findings suggest that emotionally salient and cognitively engaging warnings are more effective in prompting users to act, an outcome highly aligned with the intended purpose of trigger warnings, especially in the context of sensitive topics such as anxiety and depression.

Looking at Recommendation 3 (Verified Institutions), a recent systematic review of the literature highlights that "to improve the effectiveness of misinformation correction, it is essential to systematically build public trust" (S. Zhang et al., 2024). Although trust is a broader concept that extends beyond this single recommendation, the verification institution mechanism offers a concrete opportunity to improve users' trust in specific accounts and the information they share. This reinforces the relevance and potential impact of Recommendation 3.

Furthermore, a recent TikTok field experiment found that mental health influencers who received evidence-based training toolkits produced significantly more reliable content. These results suggest that interventions supporting reliable sources are both effective and scalable (Motta et al., 2024). In this context, Recommendation 3 can gain additional relevance as a mechanism to increase the visibility and credibility of accounts that provide reliable mental health information on Instagram. Promoting such accounts can help bridge the trust gap, guide users toward evidence-based content, and reduce the harms of misinformation.

Finally, while no direct evidence was found that specifically addressed the feasibility of Recommendation 4 (User-activated content filter), the meta-analysis by Walter et al. (2021) reinforces its underlying rationale. The study shows that users who are more involved in a topic are more receptive to corrective information. By allowing users to actively opt for validated content, this recommendation leverages user autonomy: enhancing engagement, supporting informed decision making, and reducing susceptibility to misinformation.

6.7. Strategic Management Considerations

Viewing technology as a strategic corporate resource, previously identified socio-technical interventions must also respond to the goals and expectations of key stakeholders outlined in Table 4.2. This section discusses the managerial implications of these interventions and outlines key considerations for their effective implementation within Instagram's broader strategic framework.

As noted in Fernández et al. (2021), Instagram's recommendation system is closely related to engagement and monetization. Implementing interventions such as echo chamber breakers or trigger warnings (Recommendations 1 and 2) can reduce the time spent on emotionally charged or reinforcing content, potentially affecting core engagement metrics. However, Li et al. (2023) argues that integrating the principles of responsible innovation, defined as designing and deploying technology to meet societal needs while anticipating

and addressing social impacts, can improve the long-term resilience of a company. Rather than being seen as a burden of compliance, responsible innovation is increasingly viewed as a strategic asset. Companies that embed it into their operations tend to gain adaptability, public trust, and legitimacy, factors essential for success in a rapidly evolving digital landscape. In contrast, as noted by Li et al. (2023), neglecting broader concerns, such as sustainability or social responsibility, can increase exposure to long-term reputational and operational risks.

In addition, the proposed recommendations would impact a variety of Instagram stakeholders, including users, mental health professionals, advocacy groups, regulators, and advertisers. From a strategic management perspective, it is essential to anticipate potential resistance or skepticism, particularly around content moderation and algorithmic filtering. Transparent communication will be key, and involving external advisory bodies, such as the WHO or national health ministries, can help improve credibility and public trust. For example, the recommendation of ”verified institutions” could be framed as a partnership initiative, offering increased visibility to trusted health sources while demonstrating the commitment of the platform to responsible innovation in the eyes of both users and policy makers.

6.7.1. Challenges and Trade-offs

One of the main challenges is that each recommendation would need close collaboration between teams such as policy, design, engineering, legal, and data science. For example, implementing verified institutions would require a clear process for vetting accounts, legal rules for accountability, and technical systems to prioritize their content. This means that Instagram would need to invest resources, adjust workflows, and possibly create new team setups.

From a compliance point of view, the proposed interventions raise critical concerns related to censorship, data usage, liability, and fairness, especially under frameworks like the GDPR, the EU’s Digital Services Act (DSA), and Brazil’s LGPD. These regulations require that any content filtering or personalization mechanism, such as echo chamber breakers or user-activated filters, be transparent about how personal data are collected and used (G. F. do Brasil, 2018). Moreover, platforms could face increased legal liability if users rely on health-related content flagged or promoted by the system (e.g., through verified institutions) and still face negative consequences. Additionally, interventions such as trigger warnings or filtering may raise concerns about freedom of expression and must be carefully designed to respect user autonomy, avoid undue censorship, and ensure that diverse viewpoints remain accessible.

7

Discussion and Conclusion

This chapter reflects on the key findings of the thesis, offering a critical perspective on the contributions, limitations, and future directions of the research. It begins by clarifying the main theoretical and methodological contributions, such as the integration of System-Theoretic Process Analysis (STPA) with Value Sensitive Design (VSD). Then these findings are placed within the broader academic literature to highlight where this study aligns with or diverges from existing work. The chapter also addresses the generalizability of results, considering the limitations of sample diversity and stakeholder inclusion, and concludes by outlining promising directions for future research and design practice.

7.1. Discussion

This research began with the premise that misinformation about mental health that circulates on social media poses significant risks to both individuals and society. Throughout the thesis, a pragmatic and normative perspective was adopted using science as a basis for defining misinformation, while also reflecting on what should be considered valid within broader social and ethical contexts. In this section, we draw connections between the contributions of this research and the existing literature, highlighting where new insights have emerged and where earlier findings are confirmed or extended.

In reviewing the literature, Tang et al. (2024) highlights that health misinformation is multifaceted and layered, making it difficult to define or regulate in absolute terms. To clarify these layers, this study identified four key human values—*Knowledge*, *Autonomy*, *Safety*, and *Integrity*, as well as the tensions that influence how users interpret and engage with mental health content on Instagram. Our findings align with Tang et al. (2024), showing that users often define misinformation as content lacking scientific support, but frequently describe it as 'gray', unverified but emotionally supportive or personally helpful. By identifying the values that shape these perceptions and revealing tensions such as *Immediacy vs. Scientific Accuracy*, *Accessibility vs. Specificity*, and *Explorability vs. Self-Honesty*, this study further illustrates the complex conflict between scientifically verified information and content perceived as personally meaningful.

The value-based lens adopted in this study further helps to explain the ambiguity around what is considered misinformation. As the interviews revealed, content that may be harmful to some users can be perceived as helpful or benign by others. Several participants indicated that if a piece of information did not present an immediate or direct risk to them personally, they did not see it as problematic and in many cases were willing to engage with or even rely on it.

Our findings also reinforce those of Naslund et al. (2019), who highlighted the importance of social media as a space for accessing peer support among individuals with mental illness. In our study, several identified values aligned with this insight, particularly illustrating how users value community-based support and the sharing of lived experiences. Norms such as *social and emotional support*, *free speech*, and *specificity* emerged as particularly relevant, since content grounded in personal experience tends to be highly contextual and tailored to individual situations.

Although this thesis did not aim to directly investigate stigma, our findings echoed similar concerns raised by Naslund and Deng (2021). Specifically, future research should further explore the interplay between structural, social, and self-stigma, and examine how structural stigma may exacerbate the negative psychosocial effects of mental health stigma. Several interviewees described how social resistance to seeking mental health support primarily driven by fear of judgment or shame can increase individuals' vulnerability to misinformation on Instagram, as they can turn to online content instead of professional help.

Content moderation was extensively discussed in the literature, but was rarely mentioned by participants in this study. We believe that this may be due, in part, to the nuanced and often subjective nature of health information. As highlighted by Willcox (2025), moderation processes are largely opaque and involve a combination of human and automated decisions. This aligns with participant concerns that algorithms may struggle to assess whether a piece of content is truly harmful to a particular individual. Furthermore, content classified as misinformation in this thesis, such as spiritual practices or knowledge shared by indigenous communities, may not be perceived as such by users themselves. Finally, Willcox (2025) argues that platforms should go beyond a purely risk and safety-driven approach to moderation and foster environments where users and creators feel heard, valued, and empowered. This recommendation is consistent with our findings on the value of *Autonomy*, which emphasizes user agency and control within digital platforms.

Some participants were also medical professionals, offering valuable insight from the perspective of the content creator. An interviewee, a physician, shared that although she aimed to post accurate and detailed health information, her social media team advised her to simplify her content to better align with the platform engagement metrics. Even including her academic credentials in her profile was deemed irrelevant. This illustrates how even health professionals are pressured to conform to the attention economy of the platform, sometimes at the expense of scientific accuracy. These dynamics support the observations of Marocolo et al. (2021), who found that professionals themselves can become unintentional sources of misinformation.

Another significant theme that emerged was the role of the government in addressing misinformation. Many participants saw public institutions as responsible for implementing policies and regulations, yet expressed uncertainty about how they could be effectively enforced. A recurring concern was the lack of trust in government, which raises important questions about how interventions will be received by the public. Although Jia et al. (2021) suggests that governments can support the public through awareness campaigns and reliable information dissemination, our findings indicate that such efforts must also address deep-rooted skepticism. Building trust and credibility must therefore be a central component of any public-facing strategy.

The interview process revealed how users perceive the risks and potential harms associated with recommendation algorithms, particularly the formation of echo chambers and filter bubbles. Nguyen (2020) emphasizes that trust is central to epistemic health, and the interviews surfaced various mechanisms that users employ to assess trustworthiness, ranging from evaluating the relevance of the source to relying on visible signals such as account verification. However, trust remains a significant challenge in mitigating misinformation, and the effectiveness of proposed interventions in fostering epistemic trust is yet to be determined. This issue is

further discussed in the limitations of the study and presents an important avenue for future research.

Reflecting on the core findings of this thesis, the four proposed recommendations not only aim to improve the overall safety of the system, but also to strengthen the autonomy of users in navigating mental health misinformation on social media. However, as previously noted, certain trade-offs may arise, particularly regarding the norm of *understandability*. When information is presented with greater detail or scientific accuracy, it can become less accessible or harder for some users to understand, potentially limiting its effectiveness.

7.1.1. Methodological Reflections

Some of the key contributions of this research come from the selection of the methodology and the applied methods. In particular, the combination of system analysis and Value Sensitive Design (VSD) remains an emerging area, with no fully developed framework that integrates both approaches. This thesis contributes to this ongoing exploration by demonstrating how the two can be meaningfully combined to address complex socio-technical challenges, such as misinformation about mental health.

This methodological choice proved relevant when analyzing complex socio-technical systems. In the Brazilian context, this involves health institutions, content management entities, and government and public bodies. As noted in the literature gap section, STPA has not previously been applied to analyze similar problems in this context. One of the main challenges in doing so was narrowing the scope of the research while still maintaining analytical depth and respecting the complexity of the system.

An important methodological link between the analysis of the system and the semi-structured interviews was the use of fictional scenarios. Inspired by Tang et al. (2024), these scenarios turned out to be a helpful way to understand how participants perceived certain situations, without relying entirely on their ability to recall past experiences. In addition, the creation of the scenarios came directly from the STPA process, which made an interesting connection between the system-level analysis and the more qualitative parts of the investigation.

In the end, the design of the recommendations was developed iteratively on top of the STPA framework to minimize both value tensions and unsafe control actions. The integration of value-sensitive design in the solution phase focused primarily on analyzing how specific norms conflicted with each other and how these tensions could be mitigated, rather than addressing the broader values in a more abstract sense.

7.2. Summary of key contributions

The main contribution of this research is the integration of System-Theoretic Process Analysis (STPA) with Value Sensitive Design (VSD) to investigate mental health content on Instagram. Through the use of semi-structured interviews and fictional scenarios, the study shows how these two approaches can be meaningfully combined, allowing system-level safety analysis to be informed by user values and perceptions. Methodologically, it demonstrates the feasibility of embedding the VSD principles within STPA, enabling the development of actionable recommendations that are both technically grounded and ethically aligned.

This work builds on previous findings by emphasizing how community support and emotional resonance influence the perceived credibility of mental health content, strengthening earlier research such as Naslund et al. (2019). It also contributes to the still limited body of work on mental health misinformation specifically (Starvaggi et al., 2024), clarifying how certain risks are perceived by users and which values shape their interaction with this content on social media. Furthermore, the findings highlight how stigma and a lack of trust in public institutions can lead people to seek support and information through social media rather than formal health

systems, pointing to the critical role of trust building in the design of future interventions.

7.3. Findings

To bring clarity to the findings of this thesis, this section addresses the four research sub-questions that were developed to guide the investigation and ultimately support the answer to the main research question.

7.3.1. RQ1: The Emergence of Risk and Harm

Mapping the emergence of risks and harms was essential for informing recommendations aimed at mitigating their impact. This process combined peer-reviewed literature with system-level analysis to identify key elements driving the propagation of misinformation and how these elements interact.

Rather than trying to capture all possible risk scenarios, the analysis focused on three interrelated levels: individual, societal, and platform (Instagram). These levels correspond to distinct feedback loops identified in the causal diagrams and reflect the layered dynamics through which misinformation-related harms emerge.

At the individual level, risks are primarily shaped by misinformation reinforcement loops (R1 and R2). Exposure to unverified content increases the belief in misinformation, which drives behavior change, such as self-diagnosis or rejection of professional guidance, and leads to psychological deterioration. This reinforces users' vulnerability, increasing the likelihood of further exposure and harm.

At the societal level, the engagement and credibility loops (R3 and R4) play a key role. Misinformation that generates high engagement is perceived as more credible and is more widely shared, strengthening public belief. This widespread acceptance not only amplifies exposure, but also weakens institutional trust and strains public health systems, creating a self-reinforcing cycle of harm.

At the Instagram (platform) level, Algorithmic and Profit-Driven Loops (R5 and R6) sustain the circulation of harmful content. Instagram's recommendation system promotes high engagement content, regardless of accuracy, while financial incentives discourage proactive moderation. This creates a reinforcing loop in which misinformation remains visible because it aligns with platform goals, not public well-being.

Together, these three layers illustrate how the emergence of risks and harms is not the result of isolated user decisions, but the product of interacting feedback loops within a socio-technical system. This systemic perspective guided the formulation of hazards and informed the first phase of the System-Theoretic Process Analysis (STPA).

7.3.2. RQ2: User Values

Through semi-structured interviews, key user values were identified. These values were then connected to specific norms that helped clarify which dimensions of each value were most meaningful to users in practice. Although these values are explored in greater depth in Chapter 5, they can be broadly summarized as the value of *knowledge*, *autonomy*, *safety*, and *integrity*.

7.3.3. RQ3: Risk perception

User risk perceptions were inferred through semi-structured interviews, which were supported by fictional scenarios designed to explore how participants might feel and react in specific situations. In general, users tended to be flexible when assessing the risks of making medical decisions based on information found on

social media. Many expressed the belief that individuals have the right to explore, experiment, and be curious, especially when it comes to mental health. For some, consuming such content could even promote self-awareness during difficult moments.

At the same time, participants recognized the risks of entering a cycle of exposure to negative content. They highlighted how the algorithm tends to reinforce user behavior, leading to echo chambers where similar types of content are repeatedly shown, especially those that are harmful or emotionally triggering. For many, this was seen as problematic: interacting with just one harmful content could lead to being exposed to more of the same, reinforcing distressing narratives rather than helping to break them. The lack of content moderation mechanisms was not explicitly mentioned by participants, and as a result, no direct assessment could be made based on this aspect.

Another interesting aspect of risk perception, though beyond the scope of this thesis, is how differently participants view risks related to physical harm compared to those related to mental health. Situations were often perceived as more serious or risky when they involved the potential for physical consequences. Several participants suggested that this may be because physical harm is visible and tangible, whereas mental health harm is often less immediately apparent or harder to recognize. This distinction appeared to influence how users evaluated the severity of different types of health-related information. We will explore these findings further in the recommendation section of this thesis.

7.3.4. RQ4: Operationalizing systemic interventions

The operationalization of systemic interventions began by prioritizing the primary user-perceived risks to be addressed through the STPA framework. Based on this, the value tensions were mitigated through the incorporation of targeted systemic interventions designed to align with both the safety constraints and the user's values.

More specifically, user perception of risk played a critical role in the identification of system hazards that could potentially lead to harm. The ideas shared during the interviews were used both to define system-level constraints and to inspire the development of recommendations. These recommendations were developed through an iterative process: we began by proposing an initial intervention, which was then assessed in terms of its ability to mitigate unsafe control actions (UCA) and reduce value tensions. We continued this process with subsequent recommendations, ensuring that each step contributed to addressing previously unmitigated UCAs or unresolved value conflicts. The iterative design phase ended once all identified UCAs and value tensions had been addressed in a meaningful way.

7.4. Conclusion

This research aimed to identify systemic interventions capable of mitigating the risks and harms associated with misinformation about mental health on Instagram. By combining system analysis with value-sensitive design (VSD), the study explored how risks emerge within the platform and how they could be addressed through interventions grounded in both technical structures and user values. We did so by answering the following research question.

What socio-technical interventions can mitigate the risks and harms of mental health misinformation on Instagram, considering the perspectives of young adults in Brazil?

The research process led to the development of four interrelated recommendations: (i) an echo chamber breaker to diversify content exposure, (ii) trigger warnings to promote safer and more reflective engagement, (iii) the verification of trusted health institutions to increase the visibility of credible sources, and (iv) a user-controlled content filter to allow individuals to opt in to scientifically validated information. These recommendations were not designed in isolation; rather, they interact and reinforce each other within the system. For example, verified accounts help feed credible content into the echo chamber breaker, while user-controlled filters add autonomy and adaptability. Trigger warnings function as a systemic safety layer, guiding users toward more conscious engagement with sensitive content.

Crucially, these solutions emerged from the lived experiences of users and were operationalized through a system-theoretic lens, introducing new controllers, control actions, or constraints in the platform structure. The study showed that user perceptions of misinformation are shaped by subjectivity and different levels of risk awareness. It also demonstrated that effectively addressing these risks requires interventions that not only reduce harm but also align with core values such as *Knowledge*, *Autonomy*, *Safety*, and *Integrity*.

In sum, the main contribution of this thesis is its combined use of systems thinking and user-centered design to propose practical, value-based solutions. The findings highlight the need to view misinformation as a socio-technical issue and suggest that more holistic and connected design strategies are key to creating meaningful, lasting change.

7.5. Recommendation for future research

As highlighted in the discussion, while this thesis acknowledges the importance of public institutions in addressing misinformation, it does not dive into how the proposed interventions could be operationalized through policy. Future research could explore the regulatory pathways through which national health agencies or government bodies could enforce or support systemic changes on platforms such as Instagram. This includes examining the legal, infrastructure, and political feasibility of implementing the proposed interventions and how public trust in these institutions may affect adoption.

Although this thesis proposes a set of socio-technical interventions informed by the STPA framework, their real-world effectiveness has not been validated. Future work should prioritize controlled pilot studies and field experiments to assess interventions along key dimensions such as user comprehension, engagement behavior, emotional safety, and trust in information sources. Comparative testing across different demographic groups, geographic regions, and platform governance models would help refine recommendations and ensure contextual relevance.

Another opportunity for future research lies in understanding the factors that contribute to the believability of content in digital environments. As Nguyen (2020) emphasizes, simply increasing exposure to accurate or alternative information is insufficient to counteract echo chambers, where external sources are actively discredited. Therefore, future studies could investigate the values, heuristics, and contextual factors that shape the trust of users in content. These insights could then inform and refine the design and assessment of the interventions proposed in this study, ensuring that they not only reach users but are also perceived as credible and trustworthy.

In addition, expanding the scope of system-level analysis to address hazards beyond those covered here is a necessary next step. Although this study focused on algorithmically reinforced echo chambers, future work should consider the hazards related to emotional manipulation, commercial exploitation, and the dissemination of harmful recovery narratives.

Finally, this research sets the groundwork for rethinking the balance between user agency and platform control in shaping content exposure. Future studies should investigate how users engage with interventions such as echo chamber breaks, trigger warnings, verified health accounts, and user-controlled content filters, and how these features influence perceived autonomy, trust, and information consumption behavior.

These tools can be empirically evaluated through methods such as A/B testing and longitudinal surveys. For example, the echo chamber breaker could be tested by comparing engagement and emotional response between users shown personalized content and those receiving a diversified feed. Trigger warnings can be introduced before sensitive or misleading content, with follow-up surveys that assess emotional impact, click-through rates, and perceived safety. Verified health accounts can be evaluated by adding badges to accredited sources and comparing engagement and credibility perceptions with non-verified content. Similarly, a user-controlled content filter could be offered as an opt-in feature prioritizing scientifically validated information, with adoption, usage patterns, and misinformation belief measured over time. These assessments would provide crucial evidence on the effectiveness and ethical viability of such interventions.

7.6. Reflection on Generalizability

Although this study offers insight into user values, risk perceptions, and potential systemic interventions related to misinformation about mental health on Instagram, its generalizability is limited by several factors. The sample was primarily composed of young adults from southeastern Brazil, selected through a combination of convenience and snowball sampling. As such, the perspectives captured may not fully represent the broader population of Instagram users, particularly those from other regions, age groups, or cultural contexts.

Furthermore, while some participants had healthcare experience, the study did not systematically include key stakeholder groups such as platform designers, content moderators, or policymakers. The absence of these voices limits the comprehensiveness of the systemic analysis and the direct applicability of the recommendations to platform governance or regulatory frameworks.

Furthermore, the sample leaned heavily toward participants with higher educational attainment: Nine out of ten had or were pursuing a university degree, potentially skewing the findings toward more reflective or information-literate users. Although gender representation was relatively balanced (five men, five women), the study lacked ethnic diversity and did not include individuals identifying outside the gender binary, further limiting the breadth of user perspectives captured.

Finally, the study was conducted within the Brazilian cultural context, which may influence the applicability of its findings to other regions. Cultural factors such as communication styles, social norms, and attitudes toward mental health can vary significantly between societies. For example, Brazil is often characterized as a high-context culture, where communication is based primarily on implicit messages and non-verbal cues, compared to low-context cultures that favor explicit and direct communication (Burmann & Semrau, 2022). These cultural nuances can affect the way mental health information is perceived and shared on social media platforms. Therefore, caution should be exercised when generalizing these findings beyond the Brazilian context, and future research should consider cross-cultural studies to explore these dynamics in diverse settings.

7.7. Limitations

This study examined Instagram as a platform for mental health information, with a specific focus on the risks and harms of misinformation. Although the research used a systemic and value-sensitive approach to explore

both platform-level mechanisms and user experiences, several limitations must be acknowledged.

7.7.1. Methods

Although the combination of STPA and Value Sensitive Design (VSD) represents an innovative contribution, a key limitation lies in the absence of a fully established framework for integrating these two approaches. As this methodological intersection remains underexplored, the approach adopted in this thesis should be viewed as exploratory and may benefit from further refinement, validation, and testing in future research.

7.7.1.1. Value Sensitive Design

This study identified several specific value tensions; however, VSD does not offer a systematic method to prioritize or reconcile these tensions in the design of interventions. In addition, not all values and norms surfaced consistently in all interviews, reflecting the contextual and subjective nature of value expression. This variability complicates the process of generalizing findings and translating them into universally applicable design principles.

In addition, VSD is strongly actor-centered, emphasizing user experiences, values, and perceptions. Although this perspective is relevant, it may overlook systemic or institutional constraints. This limitation justifies its integration with STPA; however, the combined approach may still face challenges, as certain systemic factors might remain underexplored and not all user values may be fully captured or translated into design implications.

7.7.1.2. System Analysis

The system analysis in this thesis was developed using insights from the literature review and the causal diagram. However, the resulting model may not fully capture the complexities or nuances of the real world system. Its accuracy is limited by the inherent subjectivity involved in defining system boundaries, which can exclude external influences such as offline behaviors or broader policy environments.

Moreover, given the complexity of the sociotechnical system under investigation, it was necessary to narrow the scope of the investigation. As a result, not all relevant actors, feedback loops, or types of harm could be examined in depth. Although this scoping allowed for a more focused analysis, it also limits the completeness of the system representation.

As mentioned above on future research, other key stakeholders, such as policy makers, platform developers, content moderators, or mental health professionals in institutional settings, were also not explicitly included in the study. Although some of the participants were healthcare professionals, their input was analyzed from the perspective of individual users and not as representatives of that stakeholder group. Consequently, the study may have overlooked systemic, technical, or regulatory considerations that these perspectives could have provided, which limits the completeness of the system-level analysis.

7.7.1.3. Interviews Sample

Some things were identified in the interview process that present limitations for this study. First, the sample size and demographic scope were limited. The interviews were conducted with a relatively small group of young adults, mainly based in the southeast part of Brazil, which may not fully capture the diversity of experiences, cultural contexts, or interactions on the platform among broader populations. Although the participants came from different social, economic and educational backgrounds, the selection of the participants was still carried out through a combination of convenience sampling with snowball sampling. This condition adds limitations to the findings and therefore should not be generalized without caution.

Furthermore, it is important to acknowledge that the sample probably underrepresents people who are active on social networks and comfortable discussing mental health issues in online spaces. Similarly to the concerns raised by Naslund et al. (2019), this subgroup may have different perspectives, such as greater openness to online mental health support or more favorable views of social networks as a platform for mental health participation. Those who did not publicly disclose their mental health status on-line or who have limited engagement with social networks were not represented, potentially biasing the findings toward more digitally engaged users.

Finally, while the sampling strategy was designed to balance accessibility and diversity, it still carries potential biases. The initial use of convenience sampling may have skewed the sample toward individuals within the immediate network of the researcher, potentially limiting the variation in perspectives. Although snowball sampling helped reach broader and potentially more marginalized groups, it can also reinforce existing social circles and shared viewpoints.

7.7.1.4. Interviews Scenarios

Although the use of fictional scenarios proved valuable in encouraging reflection and minimizing reliance on memory, it also introduced certain limitations. Participants responded to hypothetical situations, which may not fully capture their real-life behaviors or reactions. Furthermore, during the interviews, we observed that some participants became overly focused on specific narrative details within the scenarios, which occasionally diverted attention away from the broader questions about their general perceptions and values.

Regarding the realism of the scenarios, all participants stated that the situations felt realistic to them. Some even mentioned that they personally knew people in similar situations or saw aspects of themselves reflected in the characters. Realistic scenarios are essential to help participants connect emotionally and cognitively with the problem, which in turn supports more genuine reflections on values and perceptions. However, it is important to acknowledge that these scenarios were still fictional and may not fully capture the complexity or emotional nuances of real-life experiences. As such, the responses of the participants can differ from how they would react in actual situations.

7.7.1.5. Interviews Bias

As with any qualitative study, there is the potential for bias both in the interview process and in the responses of the participants. Interviewer bias may have influenced how questions were asked or followed, potentially shaping participant interpretations or prompting specific types of responses. Furthermore, social desirability bias could have affected the way participants presented their views, particularly given the sensitive nature of mental health and misinformation. Some participants may have adjusted their responses to align with perceived normative expectations or to avoid judgment, especially when discussing their own involvement with online content. These biases may have subtly influenced the insights collected and should be considered when interpreting the findings.

7.7.1.6. Data Analysis

Although the coding process allowed for a structured analysis of user values and risk perceptions, several limitations should be acknowledged. First, all coding was performed by a single researcher, which may introduce interpretive bias despite efforts to ensure consistency through comments and memos. Second, although inductive coding allowed themes to emerge from the data, subsequent deductive phases may have limited interpretive flexibility. By organizing the findings around predefined categories, the analysis could have overlooked alternative or unexpected insights present in the interviews. Third, the process of grouping codes into

higher-order categories inevitably involved subjective judgment, which may have led to oversimplification or omission of important nuances. Finally, the selective focus on values and risk perception aligned with predefined research questions may have limited the exploration of emerging but less prioritized themes, such as community dynamics or platform affordances, which could offer additional insights into misinformation ecosystems.

7.7.2. Solution

Finally, specifically on the topic of echo chambers, Nguyen (2020) argues that while epistemic bubbles, where alternative views are simply absent, can often be addressed through exposure to diverse information, echo chambers are more resilient. In these environments, external voices are actively discredited and members are conditioned to distrust them. As such, rebuilding epistemic trust, rather than simply increasing exposure, is essential to facilitate exit.

However, this thesis did not explore in depth the factors that contribute to perceived credibility and trust formation in online environments. Although it examined exposure patterns and engagement dynamics, it did not systematically analyze why certain content or sources are perceived as believable within echo chambers. This limitation restricts the solution space and narrows the potential effectiveness of the proposed interventions, which could benefit from a deeper understanding of how trust operates in digitally mediated misinformation ecosystems. As outlined above, this also presents a valuable direction for future research.

References

Akhther, N., & Sopory, P. (2022). Seeking and sharing mental health information on social media during covid-19: Role of depression and anxiety, peer support, and health benefits. *Journal of Technology in Behavioral Science*, *7*(2), 211–226. https://doi.org/10.1007/s41347-021-00239-x

Anthony Boadle. (2023). *Brazil lawmakers to vote on controversial bill to clean up social media* [Accessed: 2025-02-26]. Reuters. https://www.reuters.com/world/americas/brazil-lawmakers-vote-controversial-bill-clean-up-social-media-2023-05-02/

Au, C. H., Ho, K. K., & Chiu, D. K. (2021). Stopping healthcare misinformation: The effect of financial incentives and legislation [Accessed: February 28, 2025]. *Health Policy*, *125*(5), 627–633. https://doi.org/10.1016/j.healthpol.2021.02.010

Barman, D., Koidl, K., & Conlan, O. (2024). Discerning individual preferences for identifying and flagging misinformation on social media. *Proceedings of ACM*, 110–119. https://doi.org/10.1145/3627043.3659545

Batya Friedman, P. H. K., & Borning, A. (2013). Value sensitive design and information systems. *Foundations and Trends in Human–Computer Interaction*, *6*, 1–92.

Bioni, B. R. (2021). Plantando sementes: O papel do seminário do cgi.br na construção de uma agenda de privacidade e proteção de dados pessoais no brasil (2010-2019). *PoliTICS*.

Birnbaum, M. L., Rizvi, A. F., Confino, J., Correll, C. U., & Kane, J. M. (2017). Role of social media and the internet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. *Early Intervention in Psychiatry*, *11*(4), 290–295. https://doi.org/10.1111/eip.12237

Bodaghi, A., Schmitt, K. A., Watine, P., & Fung, B. C. M. (2024). A literature review on detecting, verifying, and mitigating online misinformation. *IEEE Transactions on Computational Social Systems*, *11*(4), 5119–5145. https://doi.org/10.1109/TCSS.2023.3289031

Burmann, K., & Semrau, T. (2022). The consequences of social category faultlines in high- and low-context cultures: A comparative study of brazil and germany. *Frontiers in Psychology*, *13*, 1082870. https://doi.org/10.3389/fpsyg.2022.1082870

Center for Countering Digital Hate. (2022). Deadly by design: Tiktok's algorithm delivers harmful content to users [Accessed March 14, 2025]. https://counterhate.com/research/deadly-by-design/

Chen, J., & Wang, Y. (2021). Social media use for health purposes: Systematic review. *Journal of Medical Internet Research*, *23*(5), e17917. https://doi.org/10.2196/17917

Christman, J. (2009). Autonomy in moral and political philosophy (E. N. Zalta, Ed.) [Last revised: 2020].

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Connected Papers. (2025). Graph of: Prevalence of health misinformation on social media: Systematic review [Accessed May 23, 2025]. https://www.connectedpapers.com/main/ebafe48570fea85c2db32d5bc126011ead0f5e55/Prevalence-of-Health-Misinformation-on-Social-Media%3A-Systematic-Review/graph

Cotter, K. (2019). Playing the visibility game: How digital influencers and algorithms negotiate influence on instagram. *New Media & Society*, *21*(4), 895–913. https://doi.org/10.1177/1461444818815684

Cox, D., La Caze, M., & Levine, M. P. (2021). Integrity (E. N. Zalta, Ed.; Fall 2021). *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/integrity/

Czeskis, A., Dermendjieva, I., Yapit, H., Borning, A., Friedman, B., Gill, B., & Kohno, T. (2010). Parenting from the pocket: Value tensions and technical directions for secure and private parent-teen mobile safety. *Proceedings of the Sixth Symposium on Usable Privacy and Security (SOUPS)*.

da Saúde, M. (2020, September). Law no. 8080: 30 years since the creation of the unified health system (sus) [Retrieved February 17, 2025]. https://bvsms.saude.gov.br/lei-n-8080-30-anos-de-criacao-do-sistema-unico-de-saude-sus/

De Hert, M., Correll, C. U., Bobes, J., Cetkovich-Bakmas, M., Cohen, D., Asai, I. M., Detraux, J., Gautam, S., Möller, H.-J., Ndetei, D. M., Newcomer, J. W., Uwakwe, R., & Leucht, S. (2011). Physical illness in patients with severe mental disorders. i. prevalence, impact of medications and disparities in health care. *World Psychiatry*, *10*(1), 52–77. https://doi.org/10.1002/j.2051-5545.2011.tb00014.x

do Brasil, G. F. (2018). Lei geral de proteção de dados pessoais (lgpd) - lei nº 13.709, de 14 de agosto de 2018 [Accessed: 2025-02-26]. https://www.gov.br/governodigital/pt-br/lgpd-pagina-do-cidadao/o-que-e-a-lgpd

do Brasil, S. F. (2020). Projeto de lei nº 2630, de 2020 - lei brasileira de liberdade, responsabilidade e transparência na internet [Accessed: 2025-02-26]. https://legis.senado.leg.br/sdleg-getter/documento?disposition=inline&dm=8110634

Domenico, G., Sit, J., Ishizaka, A., & Nunan, D. (2021). Fake news, social media and marketing: A systematic review. *Journal of Business Research*, *124*, 329–341. https://doi.org/10.1016/j.jbusres.2020.11.037

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64. https://doi.org/10.1518/001872095779049543

Enserink, B., Bots, P., Van Daalen, E., Hermans, L., Koppenjan, J., Kortmann, R., Kwakkel, J., Slinger, J., Ruijgh Van Der Ploeg, T., & Thissen, W. (2022). *Policy analysis of multi-actor systems*. TU Delft OPEN Publishing. https://doi.org/10.5074/T.2022.004

Farr, R., & Forsyth, B. (2018). Gender differences in risk perception: A theoretical and methodological review. *Health, Risk Society*, *20*(1–2), 67–91. https://doi.org/10.1080/13698575.2018.1454625

Fernández, M., Bellogín, A., & Cantador, I. (2021). Analysing the effect of recommendation algorithms on the amplification of misinformation. *arXiv preprint*, *arXiv:2103.14748*. https://doi.org/10.48550/arXiv.2103.14748

Ferrer, R., & Klein, W. M. (2015). Risk perceptions and health behavior [Accessed March 14, 2025]. *Current Opinion in Psychology*, *5*, 85–89. https://doi.org/10.1016/j.copsyc.2015.03.012

Filho, L. T. P. (1999). Iniciativa privada e saúde. *Estudos Avançados*, *13*(35), 109–. https://doi.org/10.1590/S0103-40141999000100011

Friedman, B. (1997). Human values and the design of computer technology. In B. Friedman (Ed.), *Human values and the design of computer technology* (pp. 1–18). Cambridge University Press.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001

Friedman, B., Jr., P. H. K., Hagman, J., Severson, R. L., & Gill, B. (2006). The watcher and the watched: Social judgments about privacy in a public place. *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 101–110. https://doi.org/10.1145/1124772.1124791

Fuenfschilling, L., & Truffer, B. (2014). The structuration of socio-technical regimes—conceptual foundations from institutional theory. *Research Policy*, *43*(4), 772–791. https://doi.org/10.1016/j.respol.2013.10.010

Gerdes, A., & Frandsen, T. F. (2023). A systematic review of almost three decades of value sensitive design (vsd): What happened to the technical investigations? *Ethics and Information Technology*, *25*(26). https://doi.org/10.1007/s10676-023-09700-2

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gorwa, R., & Binns, R. (2020). From content moderation to visibility moderation: A case study of platform governance on youtube. *Policy & Internet*, *12*(3), 286–308.

Guardian, T. (2025). Social media influencers are fearmongering to promote health tests with limited evidence, study finds [Accessed: February 28, 2025]. https://www.theguardian.com/society/2025/feb/27/social-media-influencers-are-fearmongering-to-promote-health-tests-with-limited-evidence-study-finds

Helberger, N. (2020). On the democratic role of news recommenders. *Digital Journalism*, *8*(3), 361–370. https://doi.org/10.1080/21670811.2019.1623700

InfoMoney. (2024). Afastamentos por saúde mental batem recorde e crescem mais de 400% desde a pandemia [Accessed March 14, 2025]. https://www.infomoney.com.br/saude/afastamentos-por-saude-mental-batem-recorde-e-crescem-mais-de-400-desde-a-pandemia/

Instagram Help Center. (2024). *How do i request a verified badge on instagram?* [Accessed: 2024-05-07]. Instagram. https://help.instagram.com/854227311295302/?helpref=hc_fnav

Janz, N. K., & Becker, M. H. (1984). The health belief model: A decade later. *Health Education Quarterly*, *11*(1), 1–47. https://doi.org/10.1177/109019818401100101

Jia, X., Pang, Y., & Liu, L. S. (2021). Online health information seeking behavior: A systematic review. *Healthcare*, *9*(12), 1740. https://doi.org/10.3390/healthcare9121740

Köhler, S., Görß, D., Kowe, A., & Teipel, S. J. (2022). Matching values to technology: A value sensitive design approach to identify values and use cases of an assistive system for people with dementia in institutional care [Published online: July 12, 2022]. *Journal of Medical Systems*. https://doi.org/10.1007/s10916-022-01861-x

Lemert, A. (2022). Facebook's corporate law paradox. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4273011

Leveson, N., & Thomas, J. (2018). *Stpa handbook* [Available for non-profit and educational use]. Self-published.

Leveson, N. G. (2011). *Engineering a safer world: Systems thinking applied to safety*. MIT Press.

Li, W., Yigitcanlar, T., Nili, A., & Browne, W. (2023). Tech giants' responsible innovation and technology strategy: An international policy review. *Smart Cities*, *6*(6), 3454–3492. https://doi.org/10.3390/smartcities6060153

Marocolo, M., Meireles, A., de Souza, H. L. R., Mota, G. R., Oranchuk, D. J., Arriel, R. A., & Leite, L. H. R. (2021). Is social media spreading misinformation on exercise and health in brazil? *International Journal of Environmental Research and Public Health*, *18*(22), 11914. https://doi.org/10.3390/ijerph182211914

Martel, C., & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features [This review is part of the themed issue on The Psychology of Misinformation (2024)]. *Current Opinion in Psychology*, *54*, 101710. https://doi.org/10.1016/j.copsyc.2023.101710

Martelaro, N., Smith, C. J., & Zilovic, T. (2022). Exploring opportunities in usable hazard analysis processes for ai engineering. https://arxiv.org/abs/2203.15628

McLoughlin, K. L., & Brady, W. J. (2024). Human-algorithm interactions help explain the spread of misinformation. *Current Opinion in Psychology*, *56*, 101770. https://doi.org/10.1016/j.copsyc.2023.101770

Mena, P., Barbe, D., & Chan-Olmsted, S. (2020). Misinformation on instagram: The impact of trusted endorsements on message credibility. *Social Media + Society*, *6*(2), 2056305120935102. https://doi.org/10.1177/2056305120935102

Mishra, A., & Sinha, P. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(6), 1–20.

Morita, P. P., & Burns, C. M. (2011). Situation awareness and risk management: Understanding the notification issues. *Studies in Health Technology and Informatics*, *164*, 372–376. https://doi.org/10.3233/978-1-60750-709-3-372

Motta, M., Liu, Y., & Yarnell, A. (2024). "influencing the influencers:" a field experimental approach to promoting effective mental health communication on tiktok [Forthcoming]. *Health Communication*. https://doi.org/10.31234/osf.io/2xv3z

Murthy, V. H. (2021). *Confronting health misinformation: The u.s. surgeon general's advisory on building a healthy information environment*. U.S. Public Health Service, Office of the Surgeon General. https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf

Nan, X., Wang, Y., & Thier, K. (2022). Why do people believe health misinformation and who is at risk? a systematic review of individual differences in susceptibility to health misinformation. *Social Science Medicine*, *314*, 115398. https://doi.org/10.1016/j.socscimed.2022.115398

Naslund, J. A., Aschbrenner, K. A., Marsch, L. A., & Bartels, S. J. (2016). The future of mental health care: Peer-to-peer support and social media. *Epidemiology and Psychiatric Sciences*, *25*(2), 113–122. https://doi.org/10.1017/S2045796015001067

Naslund, J. A., Aschbrenner, K. A., McHugo, G. J., Unützer, J., Marsch, L. A., & Bartels, S. J. (2019). Exploring opportunities to support mental health care using social media: A survey of social media users with mental illness. *Early Intervention in Psychiatry*, *13*(3), 405–413. https://doi.org/10.1111/eip.12496

Naslund, J. A., & Deng, D. (2021). Addressing mental health stigma in low-income and middle-income countries: A new frontier for digital mental health. *Ethics, Medicine and Public Health*. https://doi.org/10.1016/j.jemep.2021.100719

Naslund, J. A., Bondre, A., Torous, J., & Aschbrenner, K. A. (2020). Social media and mental health: Benefits, risks, and opportunities for research and practice. *Journal of Technology in Behavioral Science*, *5*(3), 245–257. https://doi.org/10.1007/s41347-020-00134-x

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, *17*(2), 141–161. https://doi.org/10.1017/epi.2018.32

Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., & Boyle, M. P. (2020). The ineffectiveness of fact-checking labels on news memes and articles. *Mass Communication and Society*, *23*(5), 682–704. https://doi.org/10.1080/15205436.2020.1733613

of Brazil, G. (2025). Agência nacional de vigilância sanitária [Accessed: February 28, 2025]. https://www.gov.br/anvisa/pt-br/acessoainformacao/

of Brazil, G. (n.d.). Constitution of the federative republic of brazil [Accessed: February 28, 2025]. https://codices.coe.int/codices/documents/constitution/fad0640c-b8ca-453e-aa11-5e514bc898d1

Oksanen, A., Näsi, M., Minkkinen, J., Keipi, T., Kaakinen, M., & Räsänen, P. (2016). Young people who access harm-advocating online content: A four-country survey. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *10*(2). https://doi.org/10.5817/CP2016-2-6

Organization, W. H. (2022). Mental health: Strengthening our response [Accessed: 2025-04-08]. https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

PlushCare. (2025). How accurate is mental health advice on tiktok? *PlushCare Blog*. https://plushcare.com/blog/tiktok-mental-health/

Presidência da República Federativa do Brasil. (2014). Lei nº 12.965, de 23 de abril de 2014 [Acessado em: 26 de fevereiro de 2025]. https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm

Pritchard, D. (2008). The value of knowledge (E. N. Zalta, Ed.) [Last revised: 2016].

Rao, V., Joshi, A., Kang, S. M., Lin, S., Deshpande, P., & Lee, B. (2022). Designing privacy risk frameworks for evolving cyber-physical social systems: Knowledge gaps illuminated by the case of autonomous vehicles and bystander privacy. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. https://doi.org/10.1115/DETC2022-89997

Rismani, S., Shelby, R., Smart, A., Jatho, E., Kroll, J., Shilton, K., & Catanzaro, M. (2023). From plane crashes to algorithmic harm: Applicability of safety engineering frameworks for responsible ml. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, 1665–1680. https://doi.org/10.1145/3593013.3594045

Roberts, S. T. (2019). Content moderation [Forthcoming, copy on file with author]. In L. A. Schintler & C. L. McNeely (Eds.), *Encyclopaedia of big data*. Springer. https://escholarship.org/uc/item/7371c1hf

Roeser, S., Hillerbrand, R., Sandin, P., & Peterson, M. (Eds.). (2012). *Handbook of risk theory: Epistemology, decision theory, ethics, and social implications of risk*. Springer. https://doi.org/10.1007/978-94-007-1433-5

Roman, A. (2023). A closer look into brazil's healthcare system: What can we learn? [Accessed: 2025-02-26]. *Cureus*, *15*(5), e38390. https://doi.org/10.7759/cureus.38390

Rosa, E. A. (1998). Metatheoretical foundations for post□normal risk. *Journal of Risk Research*, *1*(1), 15–44.

Rothman, J. (2023, April). What if trigger warnings don't work? [Accessed: 2025-06-04]. https://www.newyorker.com/news/our-columnists/what-if-trigger-warnings-dont-work

Saha, M., & Banerjee, D. (2024). Revisiting social stigma in non-suicidal self-injury: A narrative review. *Frontiers in Psychiatry*, *15*, 11262124. https://doi.org/10.3389/fpsyt.2024.11262124

Silveira, A. C. R. D., Camelo, A. P., Katano, B. Y. S., Wagner, F. R., de Perdigão Lana, P., & Gatto, R. F. (2025). *Proposals to regulate digital platforms in brazil: Potential impacts on the internet* [Accessed: 2025-02-26]. Internet Society Brazil Chapter, CEPI FGV Direito SP. https://www.internetsociety.org/resources/doc/2025/proposals-to-regulate-digital-platforms-in-brazil-potential-impacts-in-the-internet/

Simpson, J. A., & Weiner, E. S. C. (Eds.). (1989). *Oxford english dictionary* [Entry: "Value, n." Accessed: January 16, 2025]. Clarendon Press. https://www.oed.com/dictionary/value_n?tab=factsheet#15828043

Southwell, B. G., Niederdeppe, J., Cappella, J. N., Gaysynsky, A., Kelley, D. E., Oh, A., Peterson, E. B., & Chou, W.-Y. S. (2019). Misinformation as a misunderstood challenge to public health. *American Journal of Preventive Medicine*, *57*(2), 282–285. https://doi.org/10.1016/j.amepre.2019.03.009

Souza, I. M.-d.-S. (2017). Brazil: World leader in anxiety and depression rates. *Revista Brasileira de Psiquiatria (Brazilian Journal of Psychiatry)*, *39*(4), 384. https://doi.org/10.1590/1516-4446-2017-2300

Starvaggi, I., Dierckman, C., & Lorenzo-Luaces, L. (2024). Mental health misinformation on social media: Review and future directions. *Current Opinion in Psychology*, *56*, 101738. https://doi.org/10.1016/j.copsyc.2023.101738

Statista. (2022). Prevalence of mental health conditions in brazil in 2022, by condition. *Statista*. https://www.statista.com/statistics/1337966/prevalence-mental-health-conditions-brazil/

Statista. (2024). Countries with the most instagram users as of [january, 2024]. *Statista*. https://www.statista.com/statistics/578364/countries-with-most-instagram-users/

Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). Prevalence of health misinformation on social media: Systematic review. *Journal of Medical Internet Research*, *23*(1), e17187. https://doi.org/10.2196/17187

Tang, H., Lenzini, G., Greiff, S., Rohles, B., & Sergeeva, A. (2024). "who knows? maybe it really works": Analysing users' perceptions of health misinformation on social media. *Designing Interactive Systems Conference*, 1499–1517. https://doi.org/10.1145/3643834.3661510

Tech Policy Press Staff. (2023). *Regulating online platforms beyond the marco civil in brazil: The controversial fake news bill* [Accessed: 2025-02-26]. Tech Policy Press. https://www.techpolicy.press/regulating-online-platforms-beyond-the-marco-civil-in-brazil-the-controversial-fake-news-bill/

The Guardian. (2022). How molly russell fell into a 'vortex of despair' on social media [Accessed March 14, 2025]. https://www.theguardian.com/technology/2022/sep/30/how-molly-russell-fell-into-a-vortex-of-despair-on-social-media

Thornicroft, G. (2008). Stigma and discrimination limit access to mental health care. *Epidemiologia e Psichiatria Sociale*, *17*(1), 14–19. https://doi.org/10.1017/S1121189X00002621

van der Wilt, G. J., Reuzel, R., & Grin, J. (2015). Design for values in healthcare technology. *Science and Engineering Ethics*, *21*(5), 1125–1141. https://doi.org/10.1007/s11948-014-9584-8

Vandeskog, B. (2024). Safety is the preservation of value [Open access under Creative Commons license]. *Journal of Safety Research*, *86*, 153–159. https://doi.org/10.1016/j.jsr.2024.02.004

Varzgani, H., Kordzadeh, N., & Lee, K. (2021). Toward designing effective warning labels for health misinformation on social media. *Proceedings of the 27th Americas Conference on Information Systems (AMCIS)*.

Vijaykumar, S., Rogerson, D. T., Jin, Y., & de O. Costa, M. S. (2022). Dynamics of social corrections to peers sharing covid-19 misinformation on whatsapp in brazil. *Journal of the American Medical Informatics Association*, *29*(1), 33–42. https://doi.org/10.1093/jamia/ocab219

Vincent, C., & Amalberti, R. (2012). What is preventable harm in healthcare? a systematic review of definitions. *BMC Health Services Research*, *12*, 128. https://doi.org/10.1186/1472-6963-12-128

W. Hussain, D. M., & Whittle, J. (2018). Integrating social values into software design patterns. *Proceedings of the International Workshop on Software Fairness*, 8–14. https://doi.org/10.1145/3194770.3194777

Walter, N., Brooks, J. J., Saucier, C. J., & Suresh, S. (2021). Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis [Epub 2020 Aug 6]. *Health Communication*, *36*(13), 1776–1784. https://doi.org/10.1080/10410236.2020.1794553

Wang, K., Fu, Z., Zhou, L., & Zhu, Y. (n.d.). Content moderation in social media: The characteristics, degree, and efficiency of user engagement [Accessed: February 28, 2025]. https://www.researchgate.net/publication/370735668_Content_Moderation_in_Social_Media_The_Characteristics_Degree_and_Efficiency_of_User_Engagement

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, *240*, 112552. https://doi.org/10.1016/j.socscimed.2019.112552

Wardle, C. (2017). *Information disorder: An interdisciplinary framework* (Accessed: January 16, 2025). Council of Europe. https://firstdraftnews.org:443/coe-report/

Willcox, M. (2025). Algorithmic agency and "fighting back" against discriminatory instagram content moderation: #Iwanttoseenyome. *Frontiers in Communication*, *9*. https://doi.org/10.3389/fcomm.2024.1385869

Williams, M., & Moser, T. (2019). The art of coding and thematic exploration in qualitative research. *International Management Review*, *15*(1), 45–55.

Witt, A., Suzor, N., & Huggins, A. (2019). The rule of law on instagram: An evaluation of the moderation of images depicting women's bodies. *UNSW Law Journal*, *42*(2), 557–592.

World Health Organization (WHO). (2022). Infodemics and misinformation negatively affect people's health behaviours, new who review finds [Retrieved September 1, 2022]. https://www.who.int/europe/news/item/01-09-2022-infodemics-and-misinformation-negatively-affect-people-s-health-behaviours--new-who-review-finds

Zhang, S., Zhou, H., & Zhu, Y. (2024). Why people accept mental health-related misinformation: Role of social media metrics in users' information processing. *Social Science Computer Review*, *42*(5), 987–1003. https://doi.org/10.1177/08944393241287791

Zhang, Y., Feng, F., Wang, C., He, X., Wang, M., Li, Y., & Zhang, Y. (2020). How to retrain recommender system? a sequential meta-learning method. *arXiv preprint arXiv:2005.13258*.

A

Causal Diagrams

A.1. Individual level

Table A.1: Connections of Causal Diagram on the Individual Level

Variable	Connection	Sign	Justification	Source
Exposure to mental health misinformation	Belief in misinformation	(+)	Social media platforms serve as primary sources of mental health information for many users	(Chen & Wang, 2021)
Susceptibility to misinformation	Belief in misinformation	(+)	Susceptibility to health misinformation means that some people will be more trusting of false information	(Nan et al., 2022)
Belief in misinformation	Misinformed behavior change	(+)	Misinformation about mental health can cause individuals to self-diagnose incorrectly and avoid professional treatment	(Southwell et al., 2019)
Misinformed behavior change	Short-term psychological impact	(+)	Inappropriate self-treatment or avoiding professional care can lead to increased anxiety, emotional distress, and worsening symptoms	(Suarez-Lledo & Alvarez-Galvez, 2021)
Short-term psychological impact	Long-term mental health deterioration	(+)	Untreated or mismanaged mental health conditions can escalate into chronic anxiety, depression, or suicidal ideation	(Akhther & Sopory, 2022; Oksanen et al., 2016)
Short-term psychological impact	Exposure to mental health misinformation	(+)	Users experiencing distress may seek more online information, further exposing them to misinformation (feedback loop)	Assumption
Long-term mental health deterioration	Exposure to mental health misinformation	(+)	Users with worsening conditions may turn to social media for answers, perpetuating misinformation exposure (feedback loop)	Assumption

A.2. Societal level

Table A.2: Connections of Causal Diagram on the Societal Level

Variable	Connection	Sign	Justification	Source
User-generated misinformation	Exposure to mental health misinformation	(+)	Misinformation is created by users and then shared on the platform	(Wardle, 2017)
Exposure to mental health misinformation	Belief in misinformation	(+)	When misinformation gains credibility, individuals internalize false health advice and act accordingly	(Southwell et al., 2019)
Misinformation shared by influencers and opinion leaders	Exposure to mental health misinformation	(+)	People with high numbers of followers and in a position of celebrities have more reach on the platform	(Mena et al., 2020)
Misinformation shared by influencers and opinion leaders	Belief in misinformation	(+)	People with large followings are more likely to be trusted, increasing credibility of misinformation	(Mena et al., 2020)
Belief in misinformation	Engagement with misinformation	(+)	False information tends to generate higher engagement than verified content	(Murthy, 2021)
Engagement with misinformation	Belief in misinformation	(+)	High engagement (such as likes or endorsements from influential figures) increases perceived credibility, even if the content is false	(Mena et al., 2020)
Monetary incentives for misinformation	Misinformation shared by influencers and opinion leaders	(+)	Influencers may share false information when they receive financial incentives to do so (e.g., products, non-scientific treatments)	Assumption
Engagement with misinformation	Exposure to mental health misinformation	(+)	Recommendation algorithms prioritize engaging content, amplifying misinformation spread	(Fernández et al., 2021)
Engagement with misinformation	Monetary incentives for misinformation	(+)	Financial incentives drive higher misinformation sharing and engagement	(Au et al., 2021)
Belief in misinformation	Public health deterioration	(+)	False health beliefs lead to incorrect treatments, worsening conditions, and emergency interventions	(Suarez-Lledo & Alvarez-Galvez, 2021)
Belief in misinformation	Societal distrust and stigma	(+)	Belief in false health information can lead to conspiratorial thinking and polarization	(Murthy, 2021)
Societal distrust and stigma	Public health deterioration	(+)	Social stigma affects help-seeking experiences of individuals	(Saha & Banerjee, 2024)
Public Health Consequences	Public health expenditure	(+)	Receiving incorrect treatment or requiring emergency care raises overall health costs	Insight from stakeholder analysis

A.3. Instagram Level

Table A.3: Connections of Causal Diagram on the Instagram-Level

Variable	Connection	Sign	Justification	Source
User-generated misinformation	Prevalence of misinformation on social media	(+)	Misinformation is created by users and then shared on the platform	(Wardle, 2017)
Prevalence of misinformation on social media	Exposure to mental health misinformation	(+)	False information tends to generate higher engagement than verified content	(Murthy, 2021)
Exposure to mental health misinformation	Advertising revenue from misinformation	(+)	More engagement means users spend more time on the platform, increasing ad impressions	(Domenico et al., 2021)
Exposure to mental health misinformation	Algorithmic amplification of misinformation	(+)	Recommendation algorithms prioritize engaging content, amplifying misinformation spread	(Fernández et al., 2021)
Algorithmic amplification of misinformation	Exposure to mental health misinformation	(+)	Algorithm-driven amplification increases the number of users exposed to false information	(Fernández et al., 2021)
Advertising revenue from misinformation	Platform profitability	(+)	More ad revenue contributes directly to the platform's profitability	(Domenico et al., 2021)
Platform profitability	Incentive to moderate misinformation	(-)	Since misinformation drives engagement and revenue, platforms have little economic incentive to remove or limit its spread	(Lemert, 2022)

B

Unsafe Control Actions

Control Action	Not providing causes hazard	Providing causes hazard	Too early, too late, out of order	Stopped too soon, applied too long
Give moderation guidelines and strategy	UCA-1: Developers are not given clear guidelines for identifying mental health misinformation [H-3]	UCA-2: Developers are given vague or biased guidelines that overlook subtle misinformation [H-3]	UCA-3: Guidelines are issued only after harmful trends have already gone viral [H-3]	UCA-4: Guidelines are revoked before long-term misinformation campaigns are resolved [H-3]
Set algorithm design and configurations	UCA-5: Algorithm is not designed to deprioritize engagement with potentially harmful content [H-2]	UCA-6: Algorithm is configured to prioritize content based on engagement, including misinformation [H-2]	UCA-7: Algorithm update is applied after misinformation has already gained traction [H-2]	UCA-8: Harmful configurations remain active even after being flagged by experts [H-2]
Identify uncertain content for review	UCA-9: System does not flag questionable mental health claims for human review [H-1][H-3]	UCA-10: System flags legitimate but unconventional mental health advice, suppressing helpful content [H-3]	UCA-11: Uncertain content is reviewed only after mass exposure and engagement [H-1][H-2]	UCA-12: Review ends before deeper investigation into recurring misinformation patterns [H-3]
Filter content	UCA-13: Harmful mental health misinformation is not filtered from feeds [H-1][H-2][H-3]	UCA-14: Accurate but alternative mental health perspectives are wrongly filtered [H-3]	UCA-15: Filtering occurs after users have already acted on misinformation [H-1][H-2]	UCA-16: Content remains filtered even after being verified as safe and evidence-based [H-3]
Provide content	UCA-17: Reliable mental health resources are not shown when users search for help [H-1]	UCA-18: Misinformation about mental health is pushed to users based on engagement [H-1][H-2]	UCA-19: Educational content appears after harmful content has already influenced behavior [H-1]	UCA-20: Harmful content continues to be suggested after user reports or low engagement [H-2]
Report content	UCA-21: Reporting tools are not available when users encounter mental health misinformation [H-3]	UCA-22: Valid content is reported, undermining trust in the system [H-3]	UCA-23: Report button is available only after content is fully consumed or engaged with [H-3]	UCA-24: Reporting function disappears too quickly or is disabled for certain accounts [H-3]
Give preferences through engagement and reactions	UCA-25: User preferences are not captured, making it hard to avoid triggering or harmful content [H-1]	UCA-26: Engagement signals are used to boost misleading or sensational mental health posts [H-2]	UCA-27: System weighs engagement from vulnerable users out of context, promoting harmful trends [H-2]	UCA-28: Past reactions continue to influence feed long after users' preferences have changed [H-2]

Table B.1: Unsafe Control Actions (UCAs) and related hazards for Model Control Structure

C

Interviews questions

Before the interview itself, the participant will receive a consent form, outlining the purpose of the study, the voluntary nature of participation, and the measures for data confidentiality. Any questions or concerns raised by the interviewee about the research, their rights, or the interview process will be addressed before proceeding.

C.1. Part 1: Engagement with mental health information system on Instagram

Category 1: Engagement with Mental Health Content

- How do you typically use Instagram in your daily life?
- Do you encounter mental health related content on Instagram? If not, why do you think it does not come to you?
- Do you think it is all right or not all right for Instagram to be used as a space for sharing mental health information? Why or why not?
- Do you interact with mental health content on Instagram? If yes, what usually motivates you to do so? If not, why don't you?
- How do you usually feel when you come across mental health information on Instagram?
- What makes the information about mental health on Instagram feel relevant or helpful to you personally? Why?

Category 2: Encountering Non-Credible Mental Health Information

- Have you ever encountered mental health information on Instagram that seemed controversial? How did you respond, if at all? Why?
- What influences your decision to trust or question mental health information on social media? Why?
- Would it be okay for a person to change their opinion about mental health guidance and treatment based on Instagram?

C.2. Part 2: Reflection on the actions of characters from fictional scenarios

Individual Scenario Questions (asked after each scenario)

- What are your initial thoughts or feelings about our persona's experience? Can you tell me more about that?
- Do you think it is all right or not all right that this is happening to our persona? Why or why not?
- How would you feel if this persona were a close friend? Why do you think that is?
- If this person asked for your advice, what would you suggest they do next?

C.3. Part 3: Comparison and Solutions

- In your view, which scenario felt more realistic? Why?
- Did any of the scenarios feel more risky or harmful than the other? Why?
- In your opinion, is there anything that could be done to prevent this type of situations? (If thinks scenarios are "not all right")
- Is there anything else you would like to add about your experience with mental health content on Instagram?

D

Interviews fragments - Values

D.1. Autonomy

Interview 1

Paragraph 14: I think in the sense of sharing the experience, it can be reasonable, like, for the person to express themselves, to connect with others who have had similar experiences.

Paragraph 44: So, I think she said something, something about her mental health experience, but it was more like, she brought it from her own experience.

Paragraph 59: Having doubts, or being more curious, I wanted to go deeper or discuss it, to the point of questioning things, discussing them – I think that's acceptable.

Paragraph 71: And sometimes we try to solve it in our own world, like, by searching on Google, or even using ChatGPT.

Paragraph 140 (1): Like, there might be situations where, oh, to reduce anxiety levels, a certain practice can be used and it might actually have an effect. Likewise, like, if someone can't focus, they might realize that, I don't know, using the Pomodoro technique is something that works well for them and it can have a positive effect.

Paragraph 140 (2): Now, if someone thinks they have issues with focus or anxiety and they see other people's experiences, and want to test if those techniques work for them, I think that's valid. Like, changing routines, doing something different and all that – I think it's valid to give it a try.

Interviews 2

Paragraph 45: I think it's nice that it might have raised the possibility for him to name a feeling he has and then investigate and understand it further.

Interviews 3

Paragraph 12: There's a girl I follow – I met her at an anime event – and she's autistic. She makes videos about anime, games, that kind of stuff, and in some of them, she talks about her autism. She shares how her first experiences were until she got the diagnosis. She says things like: "These are the things I used to feel, and later I understood they were related to autism. So, if you feel something like that, it's worth seeking a professional."

Paragraph 20: I used to dislike it, but after I understood that a large part of the current generation uses social media more than they consume other types of content, my view changed.

Paragraph 46: Trying to change your habits? Okay, you can try changing habits, but doing that without any support or foundation is problematic.

Paragraph 49: Even if it doesn't harm them, sometimes people adopt habits that end up improving their lives. Maybe they get lucky. That can happen. But it's still inadequate because you're changing your habits based on nothing. You know?

Paragraph 51: Then they said something like: "I don't want to take medication, I don't want to go down that road, so I'll try to take care of myself first, and if it gets really bad, then I'll seek help."

Paragraph 64: Gabriel might be lucky and adopt a habit that unintentionally ends up improving his life. Like, he might come across fake videos, sure, but he could also stumble upon something useful. Let's say he struggles with attention – he might accidentally find a video about the Pomodoro technique and realize it works for him.

Interviews 4

Paragraph 16: The things that appeared to me in that sense, I would look at them, and if they referred to a study I might use, I would usually like, save, or share them with the co-authors I was working with in the research.

Paragraph 20: I think it's nice – even if it doesn't change much for me, it might appear for someone else and be very relevant for raising awareness, sharing strategies, or suggesting professionals who could help if the person is facing some mental health challenges.

Paragraph 26: When it's something like that, I usually hit dislike and, if possible, mark that I don't want to see that again. But honestly, I don't recall anything specific about mental health that looked like misinformation or something controversial.

Paragraph 33: When she sees a post giving advice she hadn't heard before, it might change her perspective about seeking help or support. But the opposite could happen too, right?

Paragraph 43: I wouldn't immediately ask her to stop following that content, because that could create a barrier between us when it comes to discussing this topic. Instead, I'd recommend looking for a second opinion or talking to a psychologist or psychiatrist – in short, a health professional.

Paragraph 46: I wouldn't oppose him seeing something online and trying to apply it to his daily life, as long as it made sense for organizing his routine.

Interview 5

Paragraph 28: I absorb information a lot. And sometimes it takes a while for me to form my own critical opinion about it. I have so much empathy and willingness to consider what's said to me that I end up taking it all as truth. Most things I hear—even when I'm a bit skeptical—I still think, "But what if it's true?" So, I end up becoming a bit of a hostage to it.

Paragraph 34: Because of pride and this feeling that I had to go through it alone, I didn't feel like I could count on him. So, I ended up not seeking therapy, thinking like, "I need to save up money and get financially organized to afford it myself."

Paragraph 36: I also had a lot of those feelings, like, "No, I don't need this. I don't want this. I don't want to be a burden." I'd tell myself that talking to friends or watching a video about mental health for an hour was enough.

Interview 6

Paragraph 8: That's a voluntary thing, right? I'm the one who chooses who I follow and all.

Paragraph 9: Even though I access and consume mental health content because I chose to do so—and I think that's a smarter way to use Instagram than following rich influencers who have nothing to do with me—the algorithm and the app as a whole weren't designed for that. They're designed to create dependence, because it's a free service where you're actually the product.

Paragraph 23 (part 1): I already had some knowledge. But after I started dating him, getting to know him, my awareness of LGBT+ topics grew a lot, and that was largely through social media. I started following some influencers, of course I also read and studied—my Kindle is full of LGBT-themed books—but a lot of what I learned about how to deal with myself came from following pages like "Chapadinha de Endorfina," "Bom Dia Óbvios," and "Epistemia," which is also really good. So yes, some social media pages really help.

Paragraph 23 (part 2): So, much of my knowledge about mental health and other topics came from social media. I'd find a cool book or podcast there first and then go buy the book later.

Paragraph 33: Unlike something like fasting. So, as long as it's not a tip that could harm someone's physical or mental health, I think it's okay to try. But that still doesn't replace seeing a health professional.

Paragraph 34: So, as long as it's not a tip that harms Gabriel, I think it's fine. If it's doing him good, great. Now, in Ana's case—saying fasting cures anxiety—I found that pretty absurd.

Paragraph 36: For me, it depends, because it's not clear what kind of tips or techniques Gabriel is applying. If it's not harmful to his physical or mental health, I think it's okay to try.

Paragraph 37: I think it's okay to try some tips we see online, as long as they aren't explicitly or implicitly harmful to us as individuals.

Paragraph 48: These are massive platforms that hold this data. If they aren't regulated by the government, they'll do whatever they want. Like Elon Musk talking about freedom of speech on Twitter, when really he just wants to allow the far right to make racist comments.

Interview 7

Paragraph 31: There's the "Not Interested" option on Instagram. Back then, I'd just select that and scroll past. I never felt extreme anger to the point of reporting something, but I always used the "Not Interested" feature so I wouldn't get recommended that kind of content.

Paragraph 66: Now, in Gabriel's case, he didn't say exactly what methods he's using to manage his ADHD. So, I can't say anything for sure. It could benefit him—or it could harm him.

Interview 8

Paragraph 12 (1): Well, I think it's a path to knowledge. I end up understanding some things and learning there. However, I always try to be critical, and whenever I find something strange, I search for a second source to understand whether it's actually reliable or not.

Paragraph 12 (2): But I always say Wikipedia is never reliable. I think Instagram is the same. It's not a reliable source, but it can be a first point of contact with a topic that will lead you to keep researching to understand if it's real and socially accepted.

Paragraph 28: It keeps showing up, it appears more and more. But I think that, depending on how it affects me—positively or negatively—if it's negative, I end up avoiding that content, and eventually, it stops showing up. I think that's kind of how Instagram works. But yes, it stays visible for a while.

Interview 9

Paragraph 22: Sometimes I avoid things through denial. Like, if I see something that doesn't fit with my reality, I won't engage. I don't want to keep seeing things I can't apply to my life. But that's my own denial—something I need to work through.

Paragraph 38: It depends. It's not like I've always believed one thing. If I see something new, I think it's totally valid for people to change their minds. That's the whole point of gaining knowledge. Once you have contact and mastery of a topic, you can choose how you relate to it. But being influenced—that's another matter.

Paragraph 41: That's the main thing. Once again, there are good practices—not even necessarily alternative ones. I see them more as complementary. I'm not just talking about medication or not; there are many treatment paths, and the person who can define what's best for you is the health professional who's treating you. That's why getting a diagnosis is important—to understand what stage you're in.

Paragraphs 68–69: That's how I see it: there should be filters on what can be said. But then it becomes a freedom of expression issue. I don't want to be someone telling others what they can or can't say. I've been seeing this a lot lately, and it's wearing me out—especially through music. It's really consuming me.

Interview 10

Paragraph 9 (1): In my opinion, I think it's quite appropriate, because in a way, even though there's no filter or quality control, I've picked up a lot of helpful things I actually use in my daily life.

Paragraph 9 (2): It combines something useful with something enjoyable in a lighter way. These days, I spend more time on social media than sitting down to watch a movie. Sometimes you only have, like, 30 minutes—so instead of a full film, you scroll and watch clips. That's how it is.

Paragraph 42: For example, I found out about a medication for nausea on TikTok, and it worked great. I saw it there, went out and bought it, and it was amazing.

Paragraph 45: Sometimes we just evaluate based on results. Like, maybe it worked for her. Maybe it's actually helping. I don't know.

D.2. Integrity

Interview 1

Paragraph 32: So, I think it's really important not to generalize and always either have some scientific basis or at least provide some guidance, like encouraging people to seek help or consult a doctor, or to talk to someone.

Paragraph 47: It does have an influence, but in this case, she wasn't trying to generate engagement. She was just sharing something personal, which might actually help a lot of people.

Paragraph 77: They are chasing engagement and numbers, and not really looking at Anna's specific case.

Paragraph 89: She needs to carry a lot of responsibility, because this is a very serious topic.

Paragraph 44 (duplicate reference): That makes me question credibility—because I think it's more aligned with what I said earlier. I trust institutions and organizations that have some foundation or basis. It's not just someone speaking off the top of their head… like an influencer. I don't think influencers have that kind of credibility. Honestly, I don't even follow many of them, but sometimes I come across things they say, especially when their statements go viral.

Interview 2

Paragraph 14 (1): I think it's possible to use social media to talk about this, but it requires a lot of responsibility and care in how things are said and what's being discussed.

Paragraph 14 (2): And I doubt that everyone creating content has that level of responsibility. Creating content has almost become mandatory—you have to post, you have to be relevant on social media. That leads to mass content creation that seems superficial and lacks deep reflection.

Paragraph 27 (1): She's speaking from personal experience, which is a very different place, and that's something you can't really question.

Paragraph 27 (2): I find it really hard to engage with mental health content created by someone who doesn't have at least some kind of education or practical experience in the subject.

Paragraph 45 (1): I understand that having a diagnosis is super important—it frees many people once they realize they have potential. But I don't think social media is the place to get that diagnosis.

Paragraph 47: Yes, diagnosis is liberating and important, but there's also a treatment process that needs to follow.

Paragraph 53 (1): So how can institutions and the government—which I believe have the main responsibility—help by providing access to professionals and tools? It's about making access easier, offering things at low cost or even for free.

Paragraph 53 (2): And social media reinforces what you already started believing. That turns into deep-rooted beliefs, and then it becomes really difficult to change them.

Paragraph 14 (3): These days, it's become a meme or a joke—like saying "I have ADHD" or "I'm bipolar." People are using mental health terms as part of their everyday vocabulary. It's crazy. We talk so casually about serious conditions, in totally unrelated contexts. I think the danger lies there.

Paragraph 45 (2): That puts people affected by these issues at risk. And I also think it stops us from really addressing the root of the problem. Everything gets individualized, but many of these are social and structural problems being placed on individuals to solve.

Interview 3

Paragraph 28: He associated it like, "Oh, I'm distracted, I forget things, I'm stressed in daily life. So I must have ADHD." But he didn't really understand it. Everything started to look like a symptom.

Paragraph 30 (1): I know they're professionals, and there's that thing where people show they're certified—you go to their page, and they list their qualifications, or show proof that what they're saying is legitimate.

Paragraph 30 (2): Because even if it's misinformation, or said with the wrong intention, the fact that they encourage you to go to a professional still leads you toward the right path. Even if they say something wrong, at least they're pointing you toward getting help.

Paragraph 36: The information came from someone's imagination, but it still gets amplified by people who say it worked for them.

Paragraph 47 (1): He said, "No, I don't have ADHD," and he got angry. He wanted a diagnosis, because he wanted to feel like he belonged with that group of people who were posting about it. It was a special case because he was upset that he *didn't* have it. I was surprised—like, shouldn't he be happy? But he wanted that to be the explanation for who he is. He didn't want to work on it; he wanted the validation of saying, "I act this way because I have ADHD."

Paragraph 47 (2): Then he found out his problem was actually stress. The psychologist told him he probably had a lot of trauma and was repressing things—and that it had nothing to do with ADHD. He was furious—he

came to us really upset.

Interview 4

Paragraph 28: When I see someone forcing something I know doesn't make sense—or when it's fake news—I find it really problematic.

Paragraph 41 (1): I think it's inappropriate. Even though I'm not from the field, I can recognize that what's being said has no scientific support. It's just a blogger talking.

Paragraph 41 (2): They just want more followers. They know this topic generates engagement.

Paragraph 52: Sometimes you stop being yourself, relying too heavily on strategies or tools you saw online. Instead of making small, meaningful improvements, you end up mimicking an ADHD routine that may not even apply to you.

Paragraph 65 (1): Now people say, "I have ADHD, and that's why I can't study." Like that settles it and explains everything.

Paragraph 65 (2): Maybe we need some kind of filter to prevent cases where people are being influenced by accounts more interested in engagement than in raising awareness.

Interview 5

Paragraph 7: We would still fall into that guilty pleasure type of situation—like, "I like this, I relate to it, but it also makes me feel bad."

Paragraph 13: There's more. I have a friend who studied psychology and created an account about it, which I followed and liked. My psychiatrist also has a clinic with a social media page and a YouTube channel, so she shared a lot through those. I had those direct references in my own life.

Paragraph 20: I actually feel addicted to it. Like I said—it's a guilty pleasure, what people call a "guilty pleasure."

Paragraph 28: I absorb information very easily. It takes me a while to develop my own critical view of it. I have so much empathy and willingness to consider what people say that I end up treating most of it as truth—even when I have some doubt, I still think, "But what if it's true?" So, I end up kind of trapped by it.

Paragraph 42: I don't identify as having ADHD. But if you're exhausted, extremely stressed, not sleeping well, not eating right—of course your memory won't function properly.

Paragraph 55: When you start consuming that kind of content every day, influencers become like gods—the main reference in someone's life. And maybe that wouldn't be a problem if it wasn't about such delicate, sensitive topics. But people seem to be looking for someone to idolize, to put on a pedestal.

Interview 6

Paragraph 8: Yes, but I did that intentionally. I surrounded myself with a bubble of people I know and influencers I admire—people who communicate seriously and respectfully. So I deliberately chose to follow people who align with what I believe in, especially in mental health.

Paragraph 19: I decided not to follow that type of content, but when I saw the number of followers, I was shocked. Like—two million people really believe in what this girl is saying?

Paragraph 30: And it's not an easy process. We all have ups and downs. So, all these promises that "if you do this, you'll be 100% better"—that's a lie. No one feels 100% all the time.

Paragraph 44: I've seen posts from bloggers saying things like, "I had to lose X kilos to fit into my bridesmaid dress, so I stopped eating for days." And I'm like, wow—you're saying this to tons of people, and you don't know what kind of health conditions those people have.

Interview 7

Paragraph 19: Even in the fitness space, I sometimes have to stop watching certain content because it starts turning into jokes or comparisons I find inappropriate.

Interview 8

Paragraph 20: I tend to follow more things related to studying. For example, I told you my cousin is doing a postdoc in this field. So, anything she posts gets my attention, because I know it's a reliable source—I know her background.

Paragraph 34 (1): I believe people should share more reliable information. So whether I share something or not is also a personal choice.

Paragraph 34 (2): If you don't trust a source and you're skeptical, sharing it can lead to spread of misinformation. That can influence people and even change opinions.

Interview 9

Paragraph 31 (1): Sometimes I get so outraged that I want to send things to others, but I try not to do that anymore because it boosts engagement for that type of post.

Paragraph 31 (2): Like, I'd share something saying, "This is absurd," but then that just generates more engagement for it. So now I just let it fade away on its own.

Paragraph 36 (1): Sure, everyone can have access to knowledge. But I really value formal education and investment in learning. So if I see just anyone sharing something...

Paragraph 36 (2): ...even if the content is good, if the person doesn't really master the topic—even if they speak well or seem convincing—I filter a lot based on who is talking.

Paragraphs 51–52: People often confuse receiving a diagnosis with simply recognizing they're exhausted.

It's different. We live in such a fast-paced "rat race" that we don't stop to see this. So with ADHD, we need to ask—how much of this is about routine?

Paragraph 54: Again, it's important to have a general understanding, but it's not Gabriel's job to become a PhD in ADHD. He doesn't need to know everything—it's up to a qualified professional to assess and diagnose.

Paragraph 67: One thing I strongly believe in is filtering who is spreading knowledge. That's essential to me. Consumers also need to be educated to know how to choose credible sources.

Paragraph 68 (1): I think it's one thing for a psychologist to talk about mental health therapies—or a therapist, someone trained in the area.

Paragraph 68 (2): But the platforms should also guide people so they consume content from those with the proper training and knowledge. Many people can't make that distinction on their own.

Paragraph 69 (1): So, if you're a serious professional, you try to do the right thing, build a network to share your content. But then you face marketing barriers. You want to share your credentials—say when you graduated—but the company says it's not relevant. People don't care about that, they say. So, important messages don't get visibility.

Paragraph 69 (2): So everything ends up being shaped by marketing—because it's a business, and that's what the platform prioritizes.

Paragraph 70: We need to reeducate the public. I recently spoke with a mentor about this. It's infuriating that people spreading misinformation aren't trying to educate—they're focused on attention and reach. And while qualified professionals stay silent, these other voices take over. Unless professionals step up and claim their space, we'll keep seeing unqualified people dominating.

Interview 10

Paragraph 30 (1): Because it's a social media platform, it becomes very easy for people to create content without any real oversight. And sometimes the things that go viral...

Paragraph 30 (2): ...are exactly those posts that use entertainment tricks more than actual helpful content. What I mean is, sometimes a blogger will reach way more people than a doctor ever could.

D.3. Safety

Interview 1

Paragraph 8: I wasn't trying to monitor anyone.

Paragraph 14: But when you see something and realize you're not going through it alone—that someone else is experiencing something similar—I think that can be good for self-expression, or even for finding someone with a similar issue, or to have an exchange.

Paragraph 26: Maybe it's a form of escape—like, to distract myself, when I don't want to think, when I want to avoid something.

Paragraph 71 (1): It's hard to verbalize that we have a problem, you know?

Paragraph 71 (2): It's a very fragile support network. It's not a real support system. I don't know what it is—it's a virtual support network of people who don't even know Ana.

Paragraph 89 (1): Social media isn't very regulated. But you see movies, or even Instagram pages, that use warnings like "this contains triggering content," or similar disclaimers.

Paragraph 89 (2): They warn that it may be triggering and suggest, "If you're experiencing something like this, seek a doctor."

Paragraph 89 (3): But when it comes to influencers, how do you assess accountability? They're not forcing anyone to do something, but still—it's a form of influence.

Paragraph 101 (1): Social media is very dynamic. I think the first step is to think about how to regulate this type of content, even considering possible penalties.

Paragraph 101 (2): At the very least, if someone is talking about mental health, there should be a warning—like in movies—saying "seek help" or something similar. That should be the minimum.

Paragraph 110: I don't know to what extent platform connections matter—like, I can share something on WhatsApp, and it's all connected through Meta. It's all the same system.

Paragraph 32 (revisited): I think everyone goes through more fragile, vulnerable moments when it comes to mental health—even if you're not going through depression or another disorder, you still may not be doing well.

Paragraph 101 (3): I think this is all evolving with time and maturity. I remember some older movies, ads, or soap operas that eventually started getting criticized for showing certain scenes without proper care. Over time, public concern grew, and the industry began to act. I don't know the specifics, but I imagine there are even regulations or penalties now for that sort of thing.

Interview 2

Paragraph 14: How is this influencing others? Some people—especially those still forming their opinions—might have less critical thinking. They could identify with it and feel deeply impacted.

Paragraph 35: When it comes to mental health, people are more vulnerable. For me at least, it's a more sensitive area than other topics.

Paragraph 53: I don't think the platform is exempt from responsibility. The platform needs to take accountability for this kind of thing. What actions are they taking when it comes to moderation?

Paragraph 6: For me, there's also value in knowing what's happening with people close to my network.

Interview 3

Paragraph 8: If you look at it without proper guidance, it all fits—you check every box.

Paragraph 10: So, like, I think if you have proper training and you're pointing people in the direction of seeking professional help, then that's fine.

Paragraph 20: Information that leads you in the right direction—like encouraging you to seek professional help, or people sharing their experiences and saying to go get checked—I think that's great. I used to dislike it, but now I think it's important.

Paragraph 30: A lot of people start their videos by saying "seek professional help." They tell their story, but somewhere in the video, at the beginning or written in the caption, they include that message.

Paragraph 42: It angers me—because it could have been me. If I had been younger, with a different mindset, or if I hadn't had the education or friendships or awareness I have, it could have been me.

Paragraph 49: So, he'll have to look for a new reason—even if he never wants to actually solve his issue. I'm sure that as new things keep showing up in his life, since he was someone who sought help once, he'll keep looking. Even if someone tells him it's autism or something else, if videos keep popping up...

Interview 4

Paragraph 24: She was suffering a lot. She needed channels to talk and vent, without having to pay for a psychologist or psychiatrist—which many people can't afford, depending on their social or economic situation. So, I think it's interesting that there's an outlet for people who need this kind of support.

Interview 5

Paragraph 7: I was really into it—liking and watching a lot of posts about mental health, personality traits, relationships, and breakups. Interestingly, Instagram thought I was a woman and started pushing content focused on breakups from a female perspective. I found that curious, but over time the content became heavier, a little sadder—with videos edited in a melancholic way. And eventually, it started making me feel worse.

Paragraph 13: There were other things too. I have a friend who graduated in psychology and made an account about it—I followed him and liked his content. My psychiatrist also had a clinic with a page and a YouTube channel, and she would share a lot there, which I followed. So, I had those direct references in my life too.

Paragraph 15: That search bar—anything can come up. I feel more vulnerable there. Like something might appear that won't do me any good. It just pops up and overwhelms me.

Paragraph 20: It makes me feel good, not because it's treating something I think about all the time, but because I feel guided, supported. I don't always have access to my psychiatrist or people around me with the information I need. So, I feel like it gives me light—a point of reference.

Paragraph 28 (1): I think it depends on the moment I'm in. If I'm feeling more vulnerable, I start to identify with anything—and believe it's true. I feel like I know less than I do.

Paragraph 28 (2): If you're not careful, you start defending things that drag you down. But to answer your

question...

Paragraphs 30–31: It's different when you visit a profile where the picture is some anime drawing or has a weird name. Or worse, those Instagram accounts with names like "Depressed World" or something like that—there's a lot of that out there.

Paragraph 34: I'm not going to stop eating, but I can understand how people in a bad moment, under stress, might not think straight. The number of likes, comments, shares, followers—it all gives a false sense of credibility.

Paragraph 36: I know it's irrational, but it feels appropriate at the time. Rationally, I know it's inadequate. Anyone from the outside would say that. But when you're in that place, you don't see another option—not until I went to therapy.

Paragraph 41: It makes sense—because that's when you're most free, but also most tired. And you think that scrolling endlessly on your phone will help. Hours pass, and suddenly it's the middle of the night, and you've ruined your sleep schedule.

Paragraph 55: I think the most important thing is to talk to the people close to you about how you're feeling.

Interview 6

Paragraph 9 (1): This isn't mental health policy. Have you seen the documentary *The Social Dilemma*? It's wild to think that the people who were there at the beginning of Instagram and Facebook were studying ways to alienate users. They took research on how to keep people scrolling endlessly—like infinite scrolling, which is literally based on a slot machine mechanism. Your brain gets hooked, waiting for a reward. These platforms were designed to keep people consumed and addicted.

Paragraph 9 (2): So yes, even though I choose to consume mental health content, and I think that's a smarter way to use Instagram than following rich influencers who have nothing to do with me, the platform still wasn't made for this. It was made to create dependence. It's free, but that's because *you* are the product.

Paragraph 23 (1): Just to add something—it's really crazy how we can use social media in a very abusive way toward ourselves. Like only following people and pages that show a life completely different from ours. And then you feel totally lost, unsure how to deal with that gap.

Paragraph 23 (2): There's a difference between media and reality. But you can also use social media to connect with spaces that talk about self-care, about how to support yourself.

Paragraph 27: Instagram doesn't have any clear policy for fake news. There's no fact-checking system. There's no tagging that says, "This video talks about mental or physical health."

Paragraph 33: Even among the people I follow—people I feel are careful when producing content—there's often still no message saying "go seek a health professional."

Paragraph 48 (1): The government should also play a role on these platforms. Not just because of mental health—although that's important—but also because of data. These are big platforms with huge amounts of

data.

Paragraph 48 (2): If governments don't regulate these platforms, they'll do whatever they want. Elon Musk talks about free speech on Twitter, but really he just wants to let the far right make racist comments.

Paragraph 48 (3): If governments don't step in to regulate social media and big tech companies that handle our data so effectively, then that's a huge problem. That's what really gets to me.

Interview 7

Paragraph 40 (1): I don't think it's very trustworthy. There's been a recent change in fact-checking policies across platforms—on X, on all of Meta. So I think it's even harder to trust now.

Paragraph 40 (2): You could've typed the same thing 10 seconds earlier into ChatGPT, and it would just generate anything, based on whatever source it pulls.

Interview 8

Paragraph 40: I hear a lot of my students saying things like, "Why is that person perfect and I'm not?" That may be related to social media, because all we see there is the good stuff. Rarely do people post about crying or going through hard times.

Paragraph 48: You can't treat Instagram like it has a verification badge for trustworthy health information. There's no checkmark that guarantees the content is reliable.

Paragraph 61: What I'd expect from Instagram, for example, is something like I said before—some way to verify whether something is reliable or not. After a post, maybe there should be a validation label or system. Of course, I know it's complicated, but it's a necessary step.

Paragraph 63: All of this connects to the mental health discussion we're having. Sometimes people just aren't prepared to handle difficult situations, and that can make them more fragile. I think this will have serious consequences in the future.

Interview 9

Paragraph 28: I said, "This is genius," because people take information and then apply it in a totally absurd way—turning it into misinformation. That absolutely relates to mental health.

Paragraph 68: That's why platforms should be educational. People need to consume content from professionals who are truly qualified. But users don't always know how to filter for that.

Interview 10

Paragraph 7: Even though I didn't explicitly say on Instagram that I went through a breakup, it knows. It somehow picks up on that. Even some of the thoughts I've evolved over time—it picks up on those too.

Paragraph 48: The algorithm ends up reinforcing your bubble. You start seeing something, and then it shows you more of the same. Eventually, you think that's reality—because the system is filtering everything based

on what you engage with.

D.4. Knowledge

Interview 1

Paragraph 14: I think in the sense of sharing experiences, it can be reasonable—for someone to express themselves and connect with others who've gone through similar things.

Paragraph 32: You can't treat these things as absolute truths. Sometimes it's just a catchy or impactful phrase, but it doesn't apply universally. Mental health is deeply personal.

Paragraph 53: So she brought up her own experience and gave it context. I didn't find that prejudiced or offensive.

Paragraph 20: Sometimes I see posts that feel overly positive, not serious or reflective. My algorithm often leads me to content about habits—and if you can't keep up with it all, it can lead to burnout.

Paragraph 44: That's where I feel a sense of credibility—institutions and organizations that are backed by something. It's not just someone saying whatever comes to mind… like an influencer. Honestly, I don't follow influencers much, but sometimes I come across their comments or hear about the buzz they generate.

Interview 2

Paragraph 6: At some level, I do use Instagram to stay informed about what's happening in the world. I follow some news accounts that I trust, so it also serves as a way to stay up to date.

Paragraph 10: It also helps to democratize information and create identification. Like, "I feel this too," and people relate within that space.

Paragraph 12 (1): Social media isn't something I deeply connect with. I often find the approach too superficial—based on personal experiences, which are valid, but still individual.

Paragraph 12 (2): Yes, the good side is that we're talking about it more, which allows change. But it also creates a shadow—a risk of trivialization. I feel that social media often trivializes the topic, even when professionals are involved.

Paragraph 12 (3): It's not a topic I believe can be addressed properly in a 30-second reel. It just feels way too shallow.

Paragraph 14: I think there's a big risk in spreading this kind of content superficially—it contributes to trivialization, which is what I'm concerned about.

Paragraph 18 (1): It was a long video—about 8 minutes. That's pretty long for Instagram.

Paragraph 18 (2): I think it's important to bring depth, other perspectives, and other kinds of knowledge into this type of analysis.

Paragraph 23: What catches my attention most is when I see things that feel very shallow—that's what makes it easier to question.

Paragraph 25 (1): We're making or changing diagnoses—or taking major decisions about our mental health—based on Instagram content made to be general.

Paragraph 25 (2): It can make you think, "Wow, I relate to this, maybe I have it too." But then you need to take the next step—seek a professional or medical help.

Paragraph 27 (1): She's speaking from a personal experience, which is a totally different context—one you can't easily question.

Paragraph 27 (2): I find it really difficult to engage with mental health content from someone who doesn't have at least some training or practical experience.

Paragraph 33: Social media can be useful. It can bring up something that leads someone to suspect they may be experiencing anxiety, or something similar.

Paragraph 45: I think it's great when it raises the possibility for someone to name a feeling they have—and then go investigate it and try to understand.

Paragraph 47 (1): Treatment for this kind of issue is so dependent on your specific needs.

Paragraph 47 (2): Yes, the diagnosis is important—it helps you understand yourself—but there's also a treatment process that must follow.

Paragraph 49 (1): Maybe the first scenario is more realistic, because I feel like we're living in a time when people want simple solutions to complex problems.

Paragraph 49 (2): Social media reinforces that idea—"This will be quick, it will work fast." Especially for younger generations, that first version feels more real.

Paragraph 6 (revisited): For me, it's also about knowing what's happening with people in my close network.

Interview 3

Paragraph 10: There are a lot of people who only use social media and don't access news websites—they don't follow anything happening in the world unless it's through social media. So yeah… I think it's an interesting space.

Paragraph 12: There's a girl I follow—I met her at an anime event—and she's autistic. She makes videos about anime, games, and she also talks about her autism—how her early experiences were before she got diagnosed. She says, "these things I used to feel, later I understood it was autism. So if you feel something like that, it's worth seeking a professional."

Paragraph 13: I also think it's cool when someone shares their experience. They went through that moment of not understanding themselves.

Paragraph 20: If you manage to access pages that are well-informed, you also end up in a loop of good content. The problem is, I'm not sure we really have access to the Instagram algorithm. That part, for me, is the tricky one.

Paragraph 22: I like personal experiences because they're wrapped in something that feels close to many people's reality.

Paragraph 24: Those generalist accounts really bother me. I saw some videos this week… I also have ADHD. I forget everything. I don't even remember where my car keys are… and those generalist posts reach a lot of people.

Paragraph 30: When you open someone's profile and see their credentials—their degree, their professional info—it gives me a sense of trust. It's different from just a random person talking.

Paragraph 34: I see a lot of personal experiences. Sometimes someone shares a story that really makes you reflect. You think, "I've had a similar experience… maybe this affects me too?" If I had never seen that on Instagram…

Paragraph 51: Some people say, "I don't want to take medication, I'll try to take care of myself first. If I get worse, I'll look for help." Others just say, "I have this and that's it," and they don't want to seek help.

Interview 4

Paragraph 14: Nowadays, I think we need to share scientific information, of course in a more accessible, everyday language for the general public.

Paragraph 14: And one way to reach people assertively and consistently is through social media—especially Instagram.

Paragraph 20: It's showing up for me, and maybe it won't change much for me, but for others it could be really relevant. This kind of information can raise awareness and help people learn some strategies and professionals they can turn to if their mental health is a bit fragile. So I think it's interesting.

Paragraph 22: When we're talking about high levels of anxiety or actions we can take to improve mental health conditions, this type of public information helps. It includes useful tips and statistics on growing mental health issues. I think it raises awareness and helps people understand this isn't just being dramatic.

Paragraph 24: Sometimes people might feel like they're alone—but they see others going through it too, and realize they're not.

Paragraph 31: I see that some people base their content on research, even if they're not from academia, and they follow trustworthy profiles. That gives me some confidence—like, this person doesn't follow pages spreading disinformation or fake news.

Paragraph 30–31: I mean, sometimes there's just a comic strip with some text and an illustration. But what's the source of that?

Paragraph 33: When someone sees a post with guidance they hadn't heard before, it can shift their perspective and lead them to seek support. But the opposite can happen too.

Paragraph 39: The person may have studied scientific research methods, but because of their mental state—stress, crises—they forget what they know and follow advice that seems reliable, without seeking professional help.

Paragraph 46: Whether it's due to attention deficit or impulsivity, the content reaches them in accessible language or through humor, which captures their attention.

Paragraph 46: They end up identifying with it and trying to solve the issue they recognized in themselves. I think it's a strategy to get better organized and more focused.

Paragraph 46: Even if the content was made for people with ADHD, often it's just practical advice for daily organization. I understand their situation.

Paragraph 63: I think health and education institutions—like the WHO, Ministry of Health, Ministry of Education—should produce content to offer this type of guidance.

Paragraph 65: A lot has improved in terms of awareness due to increased access to information.

Interview 5

Paragraph 18: When I liked something, it was because I identified with it, right? It really resonated with my own experiences, but often it made me feel bad, because sometimes I had to let go of those situations, and I was just left remembering them, which triggered me.

Paragraph 24: I prefer personal stories that feel more grounded in reality. I like hearing stories in general—it's easier to identify with them and to see the people sharing as human.

Paragraph 30: I don't think it helps, especially when there's no depth—when you consume things in a shallow, superficial way.

Paragraph 42: For a long time, I consumed a lot of this kind of content. The Instagram algorithm kept feeding me more about ADHD. I saw myself in it and thought: I have ADHD, no doubt.

Interview 6

Paragraph 8: I surrounded myself with a bubble of people and influencers I respect. I made a conscious choice to follow people who align with what I believe in—especially around mental health.

Paragraph 14: When something resonates with my own experiences—or when I want to send it to someone I care about—that motivates me to engage.

Paragraph 16: Silver-bullet messages bother me. Like "if you don't wake up early, you'll be unhappy." I prefer content that shows multiple ways to reach happiness and well-being.

Paragraph 23: A lot of what I've learned about myself and about mental health came from social media—pages like *Chapadinha de Endorfina*, *Epistemia*, and others.

Interview 7

Paragraph 15: I think it's tricky to use Instagram for this kind of content. Maybe there are people doing a good job, but is the information true? I think it's hard to verify that. Personally, I wouldn't use it.

Interview 8

Paragraph 12: Well, I think it's a path to knowledge—I end up understanding and learning some things there. But I always try to be critical, and when I find something suspicious, I look it up elsewhere to see if it's actually trustworthy.

Paragraph 12: Like I always say—Wikipedia is not reliable, and I think Instagram is the same. It's not a reliable source, but it can be an initial contact point with a topic, and from there you dig deeper to see what's real and widely accepted.

Paragraph 18: That's based on experience, right? Something I've been through grabs my attention—that's when I watch the whole video and try to understand.

Paragraph 24: Things from my family experience sometimes upset me—remind me of hard times. I've had depression, so it brings back those memories. It does affect me, but not in a way that brings me down for a long time.

Paragraph 36: I see this a lot—students want to fix things right away. The internet feeds that sense of urgency. It gives the impression you can solve everything on your own, without any support.

Paragraph 38: I think it comes from how the internet generates anxiety—from seeing perfect people and thinking, "I'm not like that." It's tied to that immediate mindset, not going through a real process.

Paragraph 63: I've thought about quitting teaching, leaving academia, and not continuing with a PhD—because it's hard dealing with young people who are more impatient, who stay in their bubble, and don't want to deal with effort or struggle—they want everything fast.

Interview 9

Paragraph 16: Well, because it reaches a large number of people. It's a topic that really needs to be talked about more, and like it or not, this is a tool that's going to spread that information to a very large audience.

Paragraph 28: I said, "man, that's genius," because that's what happens—people take a piece of information like that, and the way they apply it becomes absurd misinformation. I think that has a lot to do with mental health too.

Paragraph 41: You have to know—it's not just about relating to something that you give yourself a diagnosis. Like, you watch 10 reels and say, "Oh wow, I think I have anxiety." And then what? "Okay, I'll treat myself," or "I'll seek help," you know?

Paragraph 45: Another issue is that although I believe professional help is essential, it's not accessible. Not everyone can go to a psychologist or psychiatrist to take care of their mental health. It's just not financially accessible for most people.

Paragraph 45: Very few people can actually access this kind of help, and there's still a huge stigma. People sometimes prefer to deal with it on their own because they don't want to expose themselves or admit it's a problem. So it's easier to try and find alternatives on your own.

Paragraph 48: People may have all the knowledge—they know what the right tools are—and still ignore them.

Paragraph 52: Then they start looking for miracle solutions. Like thinking a single medication for ADHD will fix everything in someone's life.

Paragraph 52: That's the case with Gabriel and many others—they relate to content that reflects things from daily life, even without a diagnosis. These posts make you identify with the symptoms, but only a professional can really tell the difference.

Paragraph 54: Gabriel doesn't see the bigger context—and it's not his fault. He doesn't have the technical knowledge to know those are different issues. He just connects the dots, without seeing what's around them. I think that's really difficult.

Paragraph 54: And sometimes even professionals try to simplify their content so non-experts can understand—but only the first part of the message ends up being shown.

Interview 10

Paragraph 5: I use Instagram to consume information, so I treat it as both an informative and entertaining platform.

Paragraph 9: In my opinion, it's pretty useful—although there's no filter or quality control, I've found things I actually apply in my daily life.

Paragraph 11: The short, quick format makes it easier—I don't approach it like I'm sitting through a class or anything.

Paragraph 27: Now that I see more in-depth content appearing, I value it more—when it's presented in a more structured way, it captures my interest.

Paragraph 30: The content that gets more reach is often the one that uses entertainment techniques, not necessarily what's best in terms of quality. Like, sometimes a blogger will reach more people than a doctor.

Paragraph 31: When it comes to health, the audience may not have enough criteria to judge if something is accurate or not. Like with economics or other fields—sometimes I just think, "that's crazy," and scroll past. I mostly treat it like someone sharing an opinion or a different perspective.

Paragraph 40: Social media simplifies things too much. Like, “she fixed her issue in 5 minutes.” She shared an experience that solved her problem—and then you wonder if it might solve yours too.

Paragraph 40: It’s like this: the cost of trying something yourself is low. You don’t need to go through all the steps—find a doctor, get a referral, book an appointment. You can just try something now and see if it works.

Paragraph 42: That simplification makes it more accessible. The cost of trying a potential solution is low compared to the possible benefit—you might get lucky and it works.

Paragraph 45: I usually avoid getting into arguments about things I don’t know much about. I just move on.

E

Interviews fragments - Risk Perception

E.1. Risks related to Hazard 1

This section presents excerpts from interviews related to Hazard 1: Individuals act on unverified mental health information without professional validation, leading to inappropriate self-diagnosis or treatment.

Interview 2

Paragraph 25: It's about us, you know, making a diagnosis, changing a diagnosis, making any decision about something as fundamental as mental health—based on Instagram content that was made to be generalist.

Paragraph 42: These fasting trends—to me, they seem very… I don't know, distantly related, and with real risk of impacting her overall health.

Interview 3

Paragraph 24: I saw some videos on Instagram this week. I also have ADHD. I forget everything. You should see me—I don't even remember where my car keys are. That's it. These generalist accounts reach a lot of people.

Paragraph 28: And then he linked it like, "oh, I'm distracted, I forget things, I'm stressed out every day—so I must have ADHD." He didn't understand that… everything starts becoming a symptom.

Paragraph 38: It's going to make her mental state worse. And she'll never seek professional help. She'll just keep getting bombarded with more and more videos. I imagine that if she keeps believing it all works like that, things will get worse.

Paragraph 36: She's going to stay stressed about college. She'll keep having anxiety problems. And it'll get worse, because she's starting a fasting routine that no nutritionist recommended.

Paragraph 47: He said, "I don't have ADHD," but he really wanted the diagnosis. Because he felt upset not having the same validation as all those people posting online. This case was unique—he was angry not to have it. And I was like, "wait, shouldn't you be happy?" But he wanted that to be the explanation for why

he was the way he was. He didn't want to fix it—he wanted the excuse to say, "I'm like this because I have ADHD."

Paragraph 47: Then he found out that his real issue was stress—and that the psychologist said he probably has unresolved trauma. It had nothing to do with ADHD. And he got really angry—he came to us furious.

Paragraph 47: When someone criticized him at work, he'd say, "but I'm like this because I have ADHD"—just to see if people would let him off the hook for it, you know?

Paragraph 47: It's also about belief. The issue is the way people use a label to justify things—like, "I'm like this because…" and they don't plan to change. That's what bothers me. It's like saying, "I'm like this because the stars were aligned a certain way when I was born."

Paragraph 49: So now he'll look for another reason—even if he never wants to fix his problems, I'm sure that as more things go wrong in his life, because he once sought help, he'll keep looking. Even if people say, "you have autism" or "you have this or that," he'll keep watching those videos.

Paragraph 49: And what if that makes his mental health worse in another way? What if it damages his relationships with people close to him? That's where it gets more complex.

Paragraph 64: So, Gabriel still has a small chance of ending up in a good situation—although we tend to believe the likelihood is that he'll go down a negative path. But in Ana's case, I really didn't see any chance of a positive outcome. She seemed stuck in that loop.

Interview 4

Paragraph 39: Because of the condition she's in—due to stress, to crises—she forgets what she knows and ends up following what seems like the quickest and easiest solution.

Paragraph 48: So I think, in his case, self-diagnosing is inappropriate. While using one strategy or another might be fine, the way he embraced everything from social media all at once wasn't.

Paragraph 52: Many times, you stop being yourself. Using these strategies you saw on social media as a crutch—when you could simply improve a thing or two—doesn't mean you need to follow every routine of someone with ADHD.

Paragraph 58: In my view, Ana's case is more serious. The text makes it clear she's following influencers and changed her eating habits—which directly affects her health. Doing weekly fasting incorrectly can harm her. The case itself shows her condition worsened afterward.

Paragraph 65: Now I have ADHD—so that explains why I can't study anymore.

Interview 5

Paragraph 34: My parents had to push me a lot. My ex too—we weren't even dating yet. There was all this pressure, like, "what kind of impression am I going to make if I quit?"

Paragraph 41: In recent years, since I started working and having more to do, more info to deal with, I've started forgetting things more. Feeling more tired. I now see it's from stress, poor sleep, too much information, bad diet…

Paragraph 42: I don't identify as having ADHD. I just know I'm exhausted, under serious stress, not sleeping well, not eating well—my memory is affected.

Paragraph 52: Ana's case is riskier—because it involves treatment, especially dietary. Looking up info about disorders is one thing. But fasting as a treatment? That's dangerous. At least Gabriel was just searching for coping strategies.

Paragraph 52: Fasting reduces her food intake. I know what it's like to go without eating—it impacts stress, memory. Ana's case is riskier than Gabriel's.

Interview 6

Paragraph 6: It's not great—when I spend too much time on Instagram, I start comparing myself to others. I compare myself to the best version of everyone, because that's all we show online—our best version.

Paragraph 23: I'll just add this—because it's wild how we can use social media in a really abusive way toward ourselves. Like only following people and pages that show lives completely different from ours. Then we're left disoriented, not knowing how to deal with that gap.

Paragraph 33: Unlike fasting—if it's not a tip that harms physical or mental health, I think it's fine to try. But it still doesn't replace seeing a health professional.

Paragraph 34: If the tip isn't harmful, and it's helping Gabriel, then okay. But Ana doing fasting to "cure anxiety"? That's outrageous.

Paragraph 44: You don't know what might happen. So I find it really serious—these statements like "you must fast," "only eat protein, not carbs." These rules—what you can or can't do—they're all very relative.

Interview 7

Paragraph 43: Fasting to solve this issue? I don't think that's a solution—it'll likely make things worse later.

Paragraph 47: Because anxiety and depression are closely linked. That anxiety could turn into depression. I think it's very worrying—it could harm her future academic life.

Paragraph 64: Ana's case—because fasting might harm her daily health. Depending on how it's done, it can be harmful.

Interview 8

Paragraph 24: It reminds me of things I've been through with my family. I've had depression—it was a very complicated time. These memories come back. It does affect me, but not in a way that brings me down for

too long.

Paragraph 38: Maybe this comes from what the internet creates—the anxiety from seeing others as perfect and not yourself.

Paragraph 40: It's really about that. I see many of my students say, "why is that person perfect and I'm not?" And I think that could be from social media. Online, we only see the beautiful stuff. It's rare for someone to post themselves crying or in a rough phase.

Paragraph 46: These "trendy disorders"—I think it's because things show up a lot on Instagram, and then everyone gets constantly exposed to the same info.

Paragraph 57: I can't judge the difficulty, but depression and anxiety, in my view, can lead to much more drastic situations—like death. More than just struggling to learn.

Interview 9

Paragraph 41: You need to know—it's not just because you relate to something that you give yourself a diagnosis. Watching 10 reels and going, "oh wow, I have anxiety"—then what? Are you going to treat it or seek help?

Paragraph 41: You can't just self-diagnose and self-medicate. These are serious things. People need help—and professional support.

Paragraph 43: It's inappropriate—she's delaying her thinking, stopping herself from finding the right tools.

Paragraph 51–52: People often confuse giving a diagnosis with recognizing they're in an exhausting routine. It's not the same. We live in such a rat race that we don't even stop to realize this. So when people say ADHD, sometimes it's just the routine.

Paragraph 64: ADHD is being overexposed. People think the medication will make them super productive. But there are other ways to achieve that. As a doctor, I find it very worrying. This medication has many side effects and creates dependency.

Paragraph 65: It's not that it won't help—you take it and suddenly you're super productive. But then you become dependent on it. That's what I find dangerous.

Interview 10

Paragraph 48: Gabriel… His story is actually very common. For example, I often make personal or even third-party diagnoses. I'll say, "that person probably has autism," just based on how they interact socially.

Paragraph 48: Today I'm able to make diagnoses, and sometimes we even get it right. It's kind of like astrology—you relate something and go with it.

E.2. Risks related to Hazard 2

This section presents excerpts from interviews related to Hazard 2: The recommendation algorithm prioritizes high-engagement content regardless of accuracy, increasing exposure to harmful misinformation.

Interview 1

Paragraph 20: I think maybe my algorithm doesn't really lead me to that kind of video.

Paragraph 101: Sometimes, the algorithm itself might be the one doing that, right?

Interview 2

Paragraph 14: And I question whether everyone creating content has that level of responsibility. Creating content has practically become a requirement now. You have to be relevant on social media, you have to produce… and that leads to a mass of content that feels like it's just being made for the sake of it, without much thought.

Paragraph 23: That's kind of what social media does to you, right? If you like it, you watch it. If you don't, you scroll on. I think that's the norm.

Paragraph 53: But I don't think the platform is exempt from responsibility. I believe the platform needs to be held accountable for this kind of thing. What are they doing in terms of moderation?

Paragraph 53: Social media does that—it confirms what you've already started to believe. And then those beliefs become very deep-rooted, which makes things really difficult.

Paragraph 53: Sometimes a family member or a professional tries to question it, but the person is already so deep into that belief. Social media creates such a favorable environment for believing in it that they just won't question it.

Interview 3

Paragraph 20: It's tricky because the social media algorithm traps you in a bubble. So if you start watching a lot of misinformation videos, you end up in a misinformation loop.

Paragraph 20: On the other hand, if you manage to access well-informed pages, you can get stuck in a loop of good content too. The problem is, I don't know if we actually have access to Instagram's algorithm. That's the complicated part for me.

Paragraph 27: Because of the algorithm, I end up seeing a lot of ADHD videos. That's what shows up the most. I don't know if it's the algorithm—but it feels very targeted.

Paragraph 36: She's getting information that someone made up, and then she keeps getting bombarded by more content and people saying it works for her.

Interview 4

Paragraph 10: Usually, when something like that shows up, it's because some profiles I follow post about research.

Paragraph 60: She was already having episodes, and ended up falling into a vicious cycle. So she kept doing it more and more, and I imagine the videos she's watching now are reinforcing that.

Paragraph 60: The ”drug” here is watching those videos of people doing well with fasting. She thinks she'll feel better too. She thinks she's escaping, but actually she's digging herself deeper. That's what I find most harmful.

Interview 5

Paragraph 7: We start liking certain things, and Instagram starts shaping the feed—pulling us more in one direction.

Paragraph 7: I was liking and watching a lot of posts about mental health, personality behaviors, relationships, breakups. Curiously, Instagram assumed I was a woman and started showing me lots of breakup advice, as if I were the feminine side of the relationship. I found that really interesting, but over time, it became heavier—sadder. The videos started getting more emotional, even animated ones—and eventually, it started affecting me.

Paragraph 15: That search bar… anything can come up. I feel more vulnerable. Like something might pop up that won't do me any good—or will just suck me in.

Paragraph 42: For a long time, I started consuming more of that kind of content. The Instagram algorithm just kept throwing more at me. ADHD stuff—so many posts about symptoms. I thought, “that's me. They're describing me. I must have ADHD—no doubt.”

Interview 6

Paragraph 8: So I started following people that relate to me—like, their content aligns with what I believe in. And I ended up surrounding myself in a bubble, right?

Paragraph 11: On Instagram, in that explore section, it's really rare for anything about mental health to show up. The algorithm shows me really random stuff—like people popping pimples. Or people who lost weight, because I'm also trying to lose weight.

Paragraph 13: It's not a mental health policy. I don't know if you've seen that documentary *The Social Dilemma*, it's crazy to think that the people who were there at the beginning of Instagram and Facebook were studying ways to alienate people. They used research on how to keep people endlessly scrolling—like an infinite feed. It was developed exactly for that. It's like a casino slot machine—you pull and wait to see what comes up. That's what happens in our brain. It was really designed to keep people alienated.

Paragraph 34: There's also this side of products and supplements. Sometimes the ads really go overboard.

Paragraph 41: Instagram sometimes… we get caught in the flow. We don't leave, and certain things just stick in our heads.

Paragraph 42: Instagram doesn't have a clear policy for fake news. There's no fact-checking, no clear direction like "this video is about mental or physical health."

Paragraph 46: It's not an easy process—we have ups and downs. So all these promises like "do this and you'll be 100%"—that's a lie. No one is at 100% all the time.

Paragraph 72: *The Social Dilemma* is a documentary, and it's wild to think that people at the beginning of Instagram and Facebook were studying how to keep people endlessly scrolling. They used research to develop the infinite feed like a slot machine. That's how our brains get hooked. It was truly designed for alienation.

Interview 7

Paragraph 2: If you don't watch much of a video, Instagram doesn't recommend it.

Paragraph 9: Even in the fitness space, sometimes I have to stop watching. It starts going into this territory of comparison jokes that I don't find funny.

Paragraph 10: Yeah, like if you start following fitness stuff, it quickly branches off. The algorithm assumes what you want and pushes similar things—I don't really enjoy that.

Interview 8

Paragraph 4: I have a cousin who works with mental health, and she posts content—especially during important dates like Yellow September. I think because I click on it, it pulls in other related posts, so I end up seeing more mental health stuff.

Paragraph 13: It starts appearing more often. And as it starts affecting me—positively or negatively—I might stop accessing it. Then with time, it fades. I think that's how Instagram works. But yeah, it shows up for longer.

Paragraph 41: I've been giving up on being a teacher, staying in academia, or continuing my PhD because young people today are more immediate, stuck in their bubbles. They don't want the hard stuff—they want the quick fix.

Interview 9

Paragraph 2: Sometimes you see one post, and if you show interest, Instagram pulls in a bunch more. Sometimes nothing appears for a while, but usually there's a lot of medical content—about fatigue, stress, burnout.

Paragraph 49: We need a platform that's different—so people can access content from those who are truly qualified and trained. But people won't always know how to filter it.

Paragraph 51: Sure, you want to do things right, be a serious professional. You build a network to share your content. But then marketing comes in. Sometimes I want to say something based on my training, but companies say no. I think it's basic—I should be able to say where I studied. But it's "not interesting" for the audience. So it doesn't get reach. People don't care. I want to talk about something? "That won't work." Everything is too driven by marketing, because it's a business. That's all the platform delivers.

Paragraph 52: Everything is too driven by marketing, because it's a business. That's all the platform delivers.

Interview 10

Paragraph 2: A lot, even though I didn't explicitly say on Instagram that I broke up. But it knows. It can pick that up. Even some thoughts I've had as I processed things—it could follow.

Paragraph 13: Later, I think about whether that content makes sense or applies to me. It creates curiosity. Sometimes I mention it to others—like, the algorithm is so good. And I see that in my bubble, a lot of people are getting similar content.

Paragraph 17: But since it's a social media platform, people can very easily create content without much oversight. And sometimes what goes viral…

Paragraph 21: In general, I think the algorithm itself starts differentiating. For instance, if you look at an evangelical's or Bolsonaro supporter's Instagram, you'll see totally different content being delivered.

E.3. Risks related to Hazard 3

This section presents excerpts from interviews related to Hazard 3: Financial incentives undermine content moderation, allowing harmful misinformation to remain online.

Interview 1

Paragraph 89: Social media isn't that regulated, right? But you see movies, or even Instagram pages, that put warnings like "this image contains triggers," something strong ahead.

Paragraph 89: Now, influencers—how do you even evaluate if they should be punished? They're not forcing anyone to do anything, but it's still an exclamation point.

Paragraph 101: Social media is very dynamic. I think the first step is figuring out how to regulate this kind of content, including possibly introducing penalties.

Paragraph 101: I also think everything evolves with time and maturity. I remember how, in the past, movies, ads, and soap operas showed scenes without much concern. Later, people started criticizing that lack of care. After that, there was more attention to sensitive content—I'm sure there's even regulation or penalties for that now.

Interview 2

Paragraph 53: But I don't think the platform is exempt from responsibility. I believe it needs to be held accountable for moderation and what it does in that space.

Interview 3

Paragraph 66: I think the platform should filter fake news overall. But it's really hard to filter every kind of content. I honestly don't know how that could even be done.

Interview 6

Paragraph 27: Instagram doesn't have an established policy for fake news. There's no fact-checking, and no content labeling like "this video is about mental health" or "this is about physical health."

Paragraph 48: There's also the question of the government's role on these platforms. Not just because of mental health concerns—I think that's one angle—but also because of data issues.

Paragraph 48: These are huge platforms holding our data. If governments don't regulate them, they'll do whatever they want—like Elon Musk talking about free speech on Twitter when he really just wants to let the far-right make racist comments.

Paragraph 48: If governments don't regulate social networks or big techs that manage so much of our data, then yeah, that really worries me.

Paragraph 48: These platforms have the technology to understand what content is about mental health, what's about medicine, or law, etc.

Paragraph 48: These people in charge need to understand that they have to work to make the world a better place. And I don't see how that happens without taking mental health seriously and regulating social networks even just a little.

Interview 7

Paragraph 40: I don't find it very trustworthy. And now with the recent policy changes—fact-checking across platforms like X and Meta—I think it's gotten even more complicated.

Paragraph 68: It's a very open-ended question. Platforms have expanded things—like introducing community notes. But that's controversial too, because it's based on majority votes, and the majority doesn't always know the truth.

Interview 8

Paragraph 48: You can't treat Instagram like it has a "verified" tag for trustworthy health info like it does for celebrities. There's no verified badge for reliable health content.

Paragraph 61: From the platform itself—like Instagram—I would expect something like I mentioned before about verification. Knowing whether something is reliable or not is crucial. Maybe after a post, there could be a way to validate if it's real. I know it's hard, but it's something that could help.

Interview 9

Paragraph 67: Something I strongly believe in is filtering who spreads knowledge. That's something I've mentioned before—it's essential. The content I consume matters. I think most people are just consumers, but they should be more equipped to handle that. You know?