Risk Denial in the Age of AI: A Counterfactual Approach in Service-Related Scenarios

Le déni de risque à l'ère de l'IA : une approche contrefactuelle dans des scénarios associés aux services

Samuele Meier[1], Emmanuel Fragnière[2], Magali Dubosson[3]

[1] School of Management Fribourg (HES-SO), Switzerland, samuele.meier@hefr.ch
[2] School of Management Valais-Wallis (HES-SO), Switzerland, emmanuel.fragniere@hevs.ch
[3] School of Management Fribourg (HES-SO), Switzerland, magali.dubosson@hefr.ch

RÉSUMÉ. L'intelligence artificielle (IA) a un impact de plus en plus important sur la société, offrant des avantages tels que l'optimisation des processus et l'amélioration de la prise de décision. Cependant, son adoption rapide comporte également des risques sous-estimés. Cette étude exploratoire examine le déni du risque, un biais cognitif qui empêche les individus de reconnaître les impacts négatifs potentiels de l'IA. À l'aide d'une approche mixte comprenant 13 entretiens semi-structurés et une conception quantitative avant/après, la recherche explore comment le détachement émotionnel et les préjugés influencent la perception des risques liés à l'IA. Les thèmes clés qui ont émergé comprennent l'acceptation résignée, les menaces pour l'emploi et la vie privée, et la nécessité d'une réglementation. Les participants se sentaient souvent impuissants, considérant l'essor de l'IA comme inévitable et échappant au contrôle individuel. L'étude préconise donc le renforcement des cadres juridiques, des normes éthiques et de l'éducation du public afin de favoriser une prise de décision éclairée. Des scénarios contrefactuels et des exemples concrets ont été utilisés pour encourager la réflexion et évaluer les changements d'attitude. Les résultats suggèrent que la distance émotionnelle et la normalisation des risques liés à l'IA contribuent à une acceptation passive. Cela souligne l'importance de favoriser un engagement critique et un dialogue sociétal. Cependant, compte tenu de la petite taille de l'échantillon, les résultats sont préliminaires et destinés à éclairer les recherches futures.

ABSTRACT. Artificial intelligence (AI) is having an increasingly significant impact on society, offering benefits such as process optimisation and improved decision-making. However, its rapid adoption also poses under-assessed risks. This exploratory study examines risk denial, a cognitive bias whereby individuals fail to acknowledge the potential negative impacts of AI. Using a mixed-methods approach involving 13 semi-structured interviews and a pre/post quantitative design, the research explores how emotional disengagement and bias influence perceptions of AI risks. Key themes that emerged include resigned acceptance, threats to employment and privacy, and the need for regulation. Participants often felt powerless, viewing the rise of AI as inevitable and beyond individual control. The study therefore calls for stronger legal frameworks, ethical standards and public education to support informed decision-making. Counterfactual scenarios and real-world examples were employed to encourage reflection and assess attitude changes. The findings suggest that emotional distance and the normalisation of AI risks contribute to passive acceptance. This highlights the importance of fostering critical engagement and societal dialogue. However, given the small sample size, the findings are preliminary and are intended to inform future research.

MOTS-CLÉS. déni du risque, biais cognitifs, gestion numérique des risques, éthique de l'IA, stratégies de résilience

KEYWORDS. risk denial, cognitive biases, digital risk management, AI ethics, resilience strategies

1. Context

An Australian oceanographer, Shane Keating, has developed an algorithm using accurate mapping of ocean currents and eddies to optimize shipping routes, reducing fuel consumption by 5% to 25% and related emissions by 15% to 25% (https://www.unsw.edu.au/news/2024/08/shane-keating-aea-seed-funding). This is just one of many examples that demonstrate the benefits of applying artificial intelligence (AI) to good effect. Indeed, AI has become ubiquitous in our daily lives as employees, customers and citizens. It can also help solve society's problems, such as global warming and disease diagnosis. However, the rapid integration of AI into everyday contexts also introduces significant risks, often invisible or inadequately acknowledged, necessitating a critical examination..

This study adopts an explicitly exploratory stance and, as such, makes a deliberate choice to focus on one specific dimension of AI risk perception: the potential underestimation or denial of risks. While we fully acknowledge that attitudes toward AI include a broad spectrum of responses, from techno-optimism to fear and resistance, our analytical lens is intentionally oriented toward mechanisms that lead to the minimization or suppression of perceived threats. This choice is motivated by the observation that, in public discourse and professional practice, risks associated with AI are often obscured by narratives of inevitability, efficiency, and innovation. As an exploratory study, our aim is not to offer a comprehensive typology of all cognitive biases related to AI, but to generate insight into how and why certain individuals may downplay or disengage from the recognition of risk, even when confronted with critical scenarios.Indeed, risk denial can affect our understanding of risk and influence our decision-making. It really is crucial to prepare for these kinds of problems, rather than ignoring them in blissful optimism, because, as put by [SCH 21] "What is clear is that the potential for human victims and material damage is growing as society becomes ever more reliant on digital technologies". They add that it's no longer a question of whether we'll have to deal with the consequences, but rather of when. Digitalization, intended to alleviate stress by automating routine tasks, often fails to significantly reduce workplace stress or enhance work-life balance. Instead, automation may inadvertently intensify productivity pressures, leading to increased psychological strain.Tools like ChatGPT show that even creativity is targeted by this automation, feeding a capitalist ideal of digitizing everything and promoting technological determinism. As users, we have become (consciously or unconsciously) exposed to digitization and algorithm-associated vulnerabilities. While data integrity, privacy, and cybersecurity are immediate and recognized areas of concern, other equally critical issues include equity, inclusivity, and mitigating the digital divide to ensure cohesive societal integration.. Some authors (e.g. [DEM 22]) call on companies and ethical considerations to self-regulate in the interests of reputation and image.

We hypothesize that this acceptance is made without a thorough assessment of the associated risks and this is the moment for our society for asking questions about the economic and human implications of AI. We wish to investigate whether the widespread enthusiasm and positive acceptance of AI masks potential economic and human risks: what are the implications of this potential wrong perceptions of risk for us, in our daily lives as employees, customers and society. What is the role of organizations?

The objective of this exploratory study is to identify and examine potential risks associated with AI. By integrating insights from interviews and theoretical frameworks, the study seeks to explore possible approaches for resilient and ethically sound AI systems, contributing to discussions on sustainability in service ecosystems. While preliminary, these insights may inform organizational awareness and offer directions for future policy considerations, aiming to enrich the ongoing dialogue on ethical AI and digital resilience.

2. Literature Review

Risk perception is not merely a rational assessment of probability and severity; rather, it is shaped by emotional, social, and cognitive dynamics [SLO 87]. Individuals assess risks not only on the basis of statistical data but also according to affective reactions, prior experiences, and perceived control. [PER 03] elaborates the concept of risk denial theory, proposing that individuals engage in structured rationalizations to justify risk-taking behaviours while avoiding cognitive dissonance. [PER 03] also identifies three key mechanisms of risk denial: scapegoating, where blame is displaced onto stigmatized groups; self-confidence, in which individuals maintain belief in their unique capacity to manage risks; and risk comparison, through which individuals minimize a threat by juxtaposing it with more socially accepted or seemingly greater risks. Indeed, individuals, often without realizing it, adopt attitudes and behaviours that minimize or deny the reality of the potential dangers they face. From a psychological point of view, these mechanisms can alter our ability to correctly assess threats, thus significantly influencing our decisions. Among the various cognitive biases, optimism bias leads individuals to believe that they are less likely to experience negative events [SHA 22]. Normality bias leads people to assume that things will always work as they have in the past, even in the face of disaster. Confirmation bias prompts individuals to favour information that confirms their existing beliefs, while ignoring contradictory evidence [NIC 98]. The availability heuristic complicates risk perception by leading people to overestimate the probability of events on the basis of examples they can easily recall [TVE 73]. Emotional factors also play a crucial role in risk denial. High levels of fear can lead to avoidance behaviours, including risk denial, as a coping mechanism [SLO 06]. Overwhelmed individuals may deny certain risks to reduce stress[LAZ 93]. Psychological defences, such as denial, protect against uncomfortable truths [VAI 92] and optimistic thinking can lead to denial of potential negative outcomes [BUE 94].

Various large-scale events can serve as conceptual illustrations of how mechanisms of risk denial operate in societal and institutional contexts. These examples are not presented as empirical evidence but as heuristic references underlining the need for realistic risk assessment. In fact, these examples illustrate not only the pervasiveness of risk denial but also its multi-level nature, spanning individual cognition, collective narratives, and institutional decision-making The “Dot.Com Bubble” of the 1990s saw technology stock prices inflate due to irrational exuberance, leading to considerable financial losses when the bubble burst. The financial crisis of 2007-2008 was triggered by excessive borrowing and speculative investment in the real estate market, leading to a global economic recession. The Deepwater Horizon oil spill in 2010 was the result of overconfidence in safety measures, causing a major environmental and economic disaster. More recently, at the onset of the COVID-19 pandemic, many governments were optimistic in downplaying the severity and potential impact of the virus (recall here the comparison of the virus to a small flu). The reckless use of new technologies, such as the TerraUSD cryptocurrency in 2022, has highlighted the risks associated with “algorithmic stablecoins” and sparked debate on the need for more rigorous regulation in the crypto-currency universe. Finally, when it comes to AI, there are already infamous examples such as Google's facial recognition (Gemini) or Microsoft's Tay Chatbot, which have shown racial slippage, the National Eating Disorder Association's (NEDA) chatbot, which gave advice that was dangerous to health, or the algorithm of Tinder or Amazon Recruitment, which displayed sexist biases.

Artificial intelligence (AI) has become ubiquitous in our daily lives as employees, customers and citizens, and as a result digital risk management is increasingly important in the private and public sectors, and in everyday life. AI researchers warn about its use [MAR 23], stating that we don't yet know whether the benefits outweigh the risks, yet companies continue to develop and market new AI systems without much transparency and, in many cases, without sufficient control. The Global Risks Report 2024 ([WEF 24]), ranks AI-related risks among the world's top concerns, highlighting the economic and ethical implications of digitization. These include the influence of algorithms on consumer culture and potential biases [AIR 22], the reduction of people's exposure to new information or divergent opinions resulting in what [PAR 11] has termed a “filter bubble”, a state of intellectual isolation of Internet users. Other AI-related risks include the risk of consumer manipulation [DEM 22], uncontrolled AI development and lack of control by major technology companies [MAR 23], algorithmic pricing affecting market competition [ASS 24], and e-commerce risks [HEJ 03].

Of course, algorithms make our lives easier by reducing the time and effort spent on searches, providing us with recommendations that match us, and facilitating transactions such as payment or delivery [HEJ 03]. They are also making an important contribution in sectors such as healthcare [ROB 21], education [OUY 22], industry [SUS 17], and the environment [AKT 24]. However, the dark side lies in the fact that algorithms primarily benefit those who have bought and designed them for their own profit [ASS 24], reinforcing the argument that algorithms are opaque, authoritarian, biased and recursive, and entail the risk of reinforcing prejudices and preferences, limiting the potential for discovery [AIR 22].

It is imperative that individuals, companies and governments are prepared to better assess and manage risk to ensure informed and responsible decision-making in the face of the risks that mark our path. Researchers are calling for greater oversight and regulation to ensure that the use of artificial intelligence (AI) and other emerging technologies does not reinforce inequalities or introduce new risks [MIT 16]. The need for ethical standards in algorithm design is highlighted by [FLO 18] who propose a framework for trustworthy AI based on principles such as justice, autonomy and beneficence. [BUR 16] highlights the importance of algorithmic transparency to avoid discriminatory bias, recommending development practices that include regular audits and ethical impact assessments. Regulations should also include mechanisms to protect privacy and guarantee data integrity [ZUB 23]. The academic community advocates for the adoption of ethical principles in the digital transformation of organizations and emphasizes the need to carefully evaluate the impacts of emerging technological applications [FLO 16]. Similarly, governments aim to establish regulatory frameworks to guide AI development and ensure compliance, as seen in the U.S. Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence and the European Commission's Ethics Guidelines for Trustworthy AI, AI Act, and AI Pact. The implementation of such regulations and ethical practices can also benefit from a proactive approach [NEM 18], suggesting the integration of ethics into the lifecycle of AI systems, from design to use. This approach is supported by initiatives such as Ethics by Design, which seek to incorporate ethical considerations from the earliest stages of technology development [CAL 17].

In summary, risk denial is a complex and multifaceted phenomenon. In today's context of accelerated digitization, managing the risks associated with artificial intelligence and emerging technologies is becoming increasingly crucial. Although these technologies offer considerable benefits in a variety of sectors, their unregulated and non-transparent use poses major ethical and economic challenges.

Our research is positioned in the field of risk management and, in this case, in the assessment of AI risks perceived by individuals. Our research questions address these aspects. Are we aware of the risks associated with the use of AI? Aren't we being too angelic, which could lead us into situations we're not ready to face?

3. Methodology

This study adopts an explicitly exploratory design. Given the small and non-random sample, the findings should be interpreted as preliminary insights into the relationship between cognitive biases and AI-related risk perception, rather than as generalizable conclusions. No control group condition was included, and the statistical analyses are limited by low statistical power. The primary aim is to identify possible patterns that can inform future hypothesis-driven research.

This study employs a mixed-methods approach to investigate risk denial in the context of AI. Qualitative and quantitative techniques are used in combination to provide a comprehensive analysis. Both the qualitative and quantitative components of the study are designed to support the same underlying hypothesis. Given the complexity of risk denial in AI and the need to capture subjective experiences, qualitative methods were deemed appropriate. This approach facilitated an exploration of participants' attitudes, justifications, and cognitive frameworks, which are difficult to capture through quantitative methods alone. In addition to qualitative methods, a quantitative questionnaire was employed to provide measurable and comparable data on risk perception and denial in the context of AI. This approach complements the qualitative insights by offering a broader understanding of trends and changing sentiment during the interview. The study adopts a quasi-experimental framework (with an explicitly exploratory orientation), as proposed by [SCH 09]. This method is comprised of three key stages: a pre-measurement, an intervention (e.g., reading counterfactual scenarios), and a post-measurement. While it involves the collection of data at three time points, the primary objective is not to test causal effects under controlled conditions but rather to observe potential shifts in perception and the emergence of cognitive patterns indicative of risk denial.

3.1. Participants

The 13 profiles selected by convenience for the interviews were diverse and included key stakeholders from various sectors, such as finance, technology, public administration, and civil society. The participants represented a broad spectrum of professions, ages, levels of education, and familiarity with AI technologies. This diversity was intentional, ensuring that the study captured a wide range of perspectives on risk denial in the context of AI. By including voices from different professional backgrounds and varying levels of engagement with AI, the study aimed to provide a comprehensive understanding of how risk denial manifests across different sectors and demographics.Qualitative techniques

Counterfactual thinking, which differs from the traditional double-diamond approach, constitutes the foundation of our methodology and offers a robust framework for exploring risk denial in the context of AI. Counterfactual thinking, the mental process of imagining alternatives to reality, allows participants to engage in hypothetical simulations, which can reveal critical insights into potential risks and overlooked vulnerabilities [GAG 04]. By prompting participants to consider "what if" scenarios, this method helps uncover blind spots in risk perception and mitigation strategies.

In this study, the hypothetical scenario of a massive denial-of-service (DoS) attack disrupting essential daily functions was presented to participants. In light of the exploratory nature of the study and the limited sample size, all scenarios were deliberately constructed to depict high-risk or negatively framed situations. This methodological choice was intended to maximize the likelihood of eliciting emotionally salient reactions particularly fear, anxiety, or defensive rationalizations, that are conducive to the emergence of cognitive mechanisms such as risk denial.The data collection process involved semi-structured interviews, which provided a balance between consistency across interviews and the freedom to explore emergent themes. The interview questions were designed to probe participants' views on AI risks, their awareness of potential consequences, and the reasons behind any risk denial behaviours. Each interview lasted approximately 30 minutes and was recorded and transcribed for accuracy. Data analysis was conducted using thematic analysis, following Braun and Clarke's guidelines [BRA 06]. Each interview was independently coded by two researchers to enhance intersubjective validity. Discrepancies were resolved through discussion until agreement was reached. Codes were grouped into higher-order categories corresponding to key cognitive mechanisms. This process involved familiarization with the data, initial coding, identifying patterns, and developing themes related to risk denial in AI. This approach ensured that key insights were systematically identified while allowing new themes to emerge organically from the data

3.2. Quantitative technique

A quantitative questionnaire incorporating a standardized scale was employed in the research protocol to systematically assess participants' perceptions of AI-related risks and benefits. The AI Attitude Scale (AIAS) consists of items (see Table 1) designed to evaluate key dimensions of AI perceptions, such as its impact on well-being, privacy, job security, and the need for stricter regulations. We used an adapted version of the AIAS, which included additional items to gain deeper insights into the subject providing richer data by covering a broader range of concerns and opportunities related to AI. This expansion (see Table 1) allowed for a more comprehensive exploration of the complexities surrounding AI risk perception and denial. Each statement within the questionnaire is rated using a 10-point scale, ranging from 1 (strongly disagree) to 10 (strongly agree). This scale allows for nuanced responses and facilitates a more precise measurement of attitudes and perceptions.

This quantitative approach complements the qualitative interviews by offering a structured means to validate or challenge emerging themes, strengthening the reliability and validity of the research. The scale-based responses also make it easier to visualize and communicate findings to various audiences, such as policymakers, industry stakeholders, and the academic community. By incorporating the adapted AIAS into the research protocol, the study achieves a comprehensive understanding of risk perception and denial in AI, balancing qualitative depth with quantitative rigor to inform AI governance and policy decisions.

Table 1. Artificial intelligence Assessment Scale (AIAS)

On a scale of 1 to 10, please tell us to what extent you agree or disagree with the following statements
I believe AI will improve my well-being
AI is good for humanity
Concerns about AI arise mostly from misunderstandings of the technology
I believe AI presents significant risks to privacy
I am not worried about AI's impact on jobs
I believe AI will have a positive impact on my work
I think stricter regulations are necessary for companies that develop and/or use AI
I do not think AI poses any real threat to humanity.

3.1. The scenarios

To further enrich the analysis and engage participants on a deeper cognitive and emotional level, three principal counterfactual scenarios were constructed (see the appendix). Only one scenario per participant was presented in order to stimulate participants' imaginations and prompt them to reflect on potential consequences of AI failures or misuse. The counterfactual scenarios address a range of plausible crises: a social crisis resulting from the implementation of a discriminatory algorithm, a security scandal involving the compromise of sensitive personal data by an AI system, and a blackout crisis caused by an AI solution designed to address an energy shortage. Each scenario is carefully crafted to highlight key risks associated with AI technologies and to encourage participants to envision the cascading effects of these failures on society, businesses, and individuals.

The use of these counterfactual scenarios offers several benefits. First, they engage the imagination of both the participants and readers, making abstract risks more tangible and relatable. By visualizing the potential real-world consequences of AI-related failures, participants can better articulate their concerns and reconsider assumptions they may otherwise overlook. Second, these scenarios prompt critical thinking and reflection, allowing participants to explore "what if" situations that reveal vulnerabilities, gaps in preparedness, and overlooked ethical considerations. Third, counterfactual scenarios help uncover implicit biases and denial mechanisms by confronting participants with the possible outcomes of ignoring or underestimating AI risks.

Research protocol

- Pre-measure: Initial attitudes using an adapted version of the AIAS scale with 8 items.
- Scenario presentation: Participants read 1 of 3 hypothetical scenarios.
- Post-measure I: Attitudes reassessed (same items).
- Semi-structured interviews I: In-depth insights
- Newspaper headline: Related to the scenario topic.

- Post-measure II: Attitudes reassessed (same items).
- Semi-structured interviews II: Further in-depth insights.

Analysis

The qualitative data from the semi-structured interviews are subjected to discourse analysis using NVivo software with the objective of identifying recurring themes, concerns, and strategies discussed by participants. Given the exploratory nature and limited size of the sample, the qualitative analysis was conducted at an aggregate level, focusing on recurrent discursive patterns and shared cognitive mechanisms rather than on the variation of responses by demographic subgroups. The aim was not to establish correlations between individual characteristics (such as age, education, or profession) and specific attitudes, but rather to identify generalizable interpretive structures through which participants made sense of AI-related risks and benefits. Associating each quote with a participant number or demographic profile may suggest a level of analytic granularity that is methodologically unsound given the sample constraints. Instead, we adopted a thematic approach that privileges conceptual saturation over representativity, highlighting how certain mechanisms of risk underestimation or concern emerged across participants regardless of their individual backgrounds

The quantitative data from the AIAS are analyzed through Jamovi using non-parametric repeated measures ANOVA (Friedman tests) to detect significant changes in attitudes over time.

Validity and Reliability

To ensure the validity and reliability of our approach, we relied on scientifically validated scales and rigorous analysis methods, in accordance with the standards of the scientific community. The results were evaluated considering those obtained in similar studies, thus contextualizing our findings within the existing literature. Such a comparison allowed us to evaluate the generalizability of our findings and ascertain their position within the extant research literature on attitudes towards AI.Ethical considerations

Particular attention was paid to ethics and confidentiality throughout the research process. All participants were informed of the nature and objectives of the study, and their informed consent was obtained prior to participation. Data collected were anonymized and stored securely to protect participants' privacy.

4. Synthesis of Results

The study's findings indicate that cognitive biases and emotional factors play a complex role in shaping individuals' perceptions of AI risks. The analysis identified several key themes, including resigned acceptance, cognitive biases, and emotional disengagement.

4.1. Resigned acceptance and cognitive biases

The participants demonstrated a variety of cognitive biases that influenced their attitudes toward AI. A preference for human-centred processes over AI-driven ones was observed, indicative of a status quo bias. Additionally, optimism bias was observed, with some participants expressing confidence in the stability of their employment in the face of potential AI-driven disruption. The findings revealed that emotional disengagement was a significant factor, with participants expressing feelings of powerlessness and resignation. They frequently eschewed critical reflection on the prospective role of AI, perceiving its ascendance and deleterious effects as an immutable reality.

One participant observed, "I believe AI will primarily benefit my work, and I do not perceive it as a significant risk," which illustrates the optimism bias. Another participant articulated concerns related to the ecological impact of AI, stating, "From an ecological standpoint, AI is undoubtedly detrimental." Furthermore, the sentiment of being replaced by AI was pervasive: "As the population grows, the amount of work available to individuals is shrinking, and jobs are increasingly being replaced by artificial intelligence."

4.2. Emotional disengagement and powerlessness

The study revealed that emotional disengagement was a prevalent response to AI-related risks. The participants expressed feelings of being trapped in a world that is becoming increasingly dominated by AI, which they associated with stress and frustration. One participant articulated, "The realization that we are confined to a world where communication with others is limited due to the prevalence of chatbots is a source of considerable distress." Another participant commented on the growing dependency on AI for menial tasks, noting that humans are reliant on AI for tasks that may be perceived as trivial. This observation led to the question of whether legislation is necessary to regulate the use of AI.

The initial assumption was that individuals would be inclined to deny the risks associated with artificial intelligence (AI) due to cognitive biases, such as optimism and the allure of AI benefits. However, the findings revealed a more complex and detailed picture. In lieu of explicit denial, participants demonstrated partial awareness of certain risks, yet they tended to relegate these concerns to the background due to a perceived dearth of viable alternatives or a sense of lack of control. This resulted in a stance of reluctant acceptance, whereby participants acknowledged the potential risks associated with AI, including manipulation, data privacy, and job displacement, but felt constrained in their ability to take action. As one participant articulated, "Why expend the effort?"

The participants expressed a sense of futility and resignation. "There is no further action that can be taken, and a sense of being stuck is experienced." Another participant highlighted the potential for AI to exacerbate existing issues. "The advent of the Internet and, in particular, the rise of search engines such as Google has provided us with convenient access to a vast array of information. This has led to a reduction in the need to store information in the traditional sense, as it can now be accessed at any time." "It is my concern that the advent of AI will only serve to exacerbate these issues." Furthermore, the inability to disengage from AI was identified as a significant concern. "However, one must consider the possibility of disconnecting from AI, even if it is challenging, as a potential solution." "It is therefore evident that no meaningful work can be accomplished in this environment. Consequently, it is futile to attempt to do so."

While individuals are aware of the potential risks associated with AI, they often adopt a stance of reluctant acceptance, perceiving it as unfeasible to influence the inevitable trajectory of AI advancement. Most participants expressed the view that a future dominated by AI would be devoid of meaning, with greater efficiency, speed, and power concentrated in the hands of multinational corporations, but at the cost of ecological harm, job loss, dehumanization, and reduced privacy. This does not align with their notion of well-being, which can result in emotional exhaustion and disengagement from managing these risks. One participant summarized this sentiment by stating, "Having a high-powered machine just to look up how to make my seed bread on a Sunday morning makes no sense." In conclusion, the findings highlight the necessity for the implementation of strategies that facilitate individual empowerment, address cognitive biases, and promote a more human-centric approach to AI integration.

4.3. Statistical Analysis

The choice to use a non-parametric measure such as the Friedman Test (a non-parametric alternative to the repeated-measures ANOVA) for analyzing the 13 questionnaires is justified by several key factors related to the characteristics of the data and the study design. With only 13 participants, the dataset is relatively small, making it difficult to meet the normality assumptions required for parametric tests like repeated-measures ANOVA. The Friedman Test is appropriate because it does not rely on normality assumptions, making it suitable for small-sample studies. Additionally, the data collected from the questionnaire responses may not follow a normal distribution due to the limited number of participants and the subjective nature of the responses. The Friedman Test ranks the data rather than relying on means and variances, making it more appropriate when normality cannot be guaranteed.

This study also involved multiple related measurements, where participants provided responses to different items or scenarios. The Friedman Test is specifically designed to analyze repeated measures when the same group is evaluated under different conditions, making it a natural fit for this study design. Moreover, the Friedman Test is robust to outliers, which is particularly beneficial given the small sample size, as extreme values can significantly impact the results of parametric tests. The objective of the pre-test was to ascertain whether the quasi-experimental design could discern significant alterations in the participants' attitudes toward AI subsequent to their exposure to hypothetical scenarios and real-world information. The results of the Friedman test indicated that there were significant differences in the responses of the participants across the three time points for several key items.

Regarding perception of AI's Impact on Well-Being (see Table 2) the test indicated a significant difference ($\chi^2 = 7.47$, $p = 0.024$) between the pre-measure and the two post-measures. Pairwise comparisons (Durbin-Conover) demonstrated that the most substantial shift occurred between the pre-measure and the post-measures, indicating that the scenario presentation and real-world information influenced participants' views on AI's impact on their well-being.

Table 2. Descriptive statistic of answers “I believe AI will improve my well-being”

	Mean	**Median**
Pre	4.67	4.50
Post 1	3.50	2.50
Post 2	3.50	2.50

Note. Pre = Initial attitude; Post 1 = Assessment post scenario; Post 2 = Assessment post newspaper headline

The hypothesis that AI is beneficial for humanity (see Table 3) was met with a notable shift in attitudes ($\chi^2 = 10.7$, $p = 0.005$). Following the scenario presentation, attitudes became less favorable, and this remained consistent after reading the news article. This suggests that the hypothetical scenarios were effective in challenging the participants' initial positive perceptions of AI.

Table 3. Descriptive statistic of answers “AI is good for humanity”

	Mean	**Median**
Pre	4.67	5.00
Post 1	3.33	3.00
Post 2	3.67	3.50

Note. Pre = Initial attitude ; Post 1 = Assessment post scenario; Post 2 = Assessment post newspaper headline

The perception of AI's risks to privacy was also examined (see Table 4). Significant results ($\chi^2 = 7.30$, $p = 0.026$) demonstrated changes primarily between the pre-measure and post-measure II. This suggests that information from the real world had a greater impact on awareness of privacy risks than hypothetical scenarios alone.

Table 4. Descriptive statistic of answers “I believe AI presents significant risks to privacy”

	Mean	Median
Pre	7.92	8.00
Post 1	8.75	8.50
Post 2	8.83	9.50

Note. Pre = Initial attitude ; Post 1 = Assessment post scenario; Post 2 = Assessment post newspaper headline

The results indicated that respondents were more supportive of stricter AI regulations (see Table 5). The test yielded a significant shift ($\chi^2 = 9.33$, $p = 0.009$) towards stronger support for regulations over time, with the most considerable change occurring after the scenario presentation and remaining high after the news article. This underscores the influence of both hypothetical and real-world contexts on the formation of regulatory attitudes.

Table 5. Descriptive statistic of answers “I think stricter regulations are necessary for companies that develop and/or use AI"

Q7.	Mean	Median
Pre	8.25	9.00
Post 1	9.42	10.00
Post 2	9.33	10.00

Note. Pre = Initial attitude; Post 1 = Assessment post scenario; Post 2 = Assessment post newspaper headline

These preliminary results support the feasibility of the experimental approach and underscore the necessity for further research with larger samples to validate and expand upon these findings.

5. Discussion

The analysis of the interviews yielded several critical insights into how individuals perceive the risks associated with AI. These insights are classified according to four categories: cognitive biases, emotional responses, perceived impacts on various aspects of life and the regulatory need. The following section presents four hypotheses derived from the findings presented in the previous section (the original scenarios are included in the Appendix).

Hypothesis 1: Cognitive Biases Influence Risk Perception and Acceptance of AI

The interviews suggest that cognitive biases, such as status quo bias and confirmation bias, significantly influence how individuals perceive and accept AI risks. Status quo bias often leads individuals to favor familiar processes over adopting new technologies, even when the latter may offer clear benefits [SAM 88]. This preference was reflected in one participant’s statement: “I would rather pay a person to dedicate half a day to contemplating a project than have a machine generate it.” Research has shown that people are more likely to resist changes to established systems due to an innate preference for maintaining the current state. Similarly, confirmation bias, where individuals focus on information that aligns with their preexisting beliefs while dismissing contradictory evidence [NIC 98], was evident in participants who reinforced initial anxieties about AI rather than objectively evaluating its potential. Such biases have been consistently linked to the shaping of risk perceptions in complex or unfamiliar technological domains [SLO 87], further explaining skepticism toward AI adoption.

Hypothesis 2: Emotional Disengagement and Powerlessness Lead to Reluctant Acceptance of AI

The findings underscore the significant role of emotional responses, particularly feelings of powerlessness and resignation, in shaping individuals' acceptance of AI risks. When people perceive a loss of control over technological changes, they may disengage emotionally and passively accept outcomes, even when they harbor concerns about potential consequences [LAN 75]. One participant encapsulated this sense of inevitability, stating, "We are dependent on AI for tasks that may be perceived as mundane, and I am left wondering whether, at some point, it will have a detrimental impact if we do not have a law in place." This reflects a broader phenomenon where resignation stems from the belief that individual actions cannot influence the trajectory of technological advancements [NIS 18]. Additionally, emotional distress associated with AI interactions, such as replacing human conversation with automated systems, reinforces a sense of alienation. Studies have demonstrated that such changes can lead to feelings of social isolation and psychological discomfort [ARN 19], as individuals struggle to adapt to an increasingly automated environment.

Hypothesis 3: Perceived Threats to Employment and Privacy Heighten AI Concerns

Concerns regarding AI's impact on employment and privacy emerged as significant factors shaping participants' perceptions. The fear of job displacement, particularly in roles requiring creativity or human interaction, reflects broader apprehensions about automation replacing human labor [FRE 17]. One participant highlighted this concern, suggesting that reliance on AI could lead to a decline in critical thinking. These worries are consistent with findings that suggest many jobs face a high risk of automation, fueling anxiety about long-term career stability [ARN 19]. In addition to employment concerns, the potential misuse of personal data adds another layer of skepticism. Participants expressed unease about the lack of transparency in how commercial data banks manage and monitor personal information. This reflects broader societal concerns regarding AI-driven data collection practices, which often exacerbate fears about privacy violations and data misuse ([ACQ 15] ; [ZUB 23]). Together, these perceived threats create a heightened sense of vulnerability and contribute to individuals' reluctance to fully embrace AI technologies.

Hypothesis 4: Regulatory and Ethical Oversight is Essential for Mitigating AI Risks

The necessity of regulatory and ethical oversight emerged as a consistent theme among participants, highlighting its role in managing AI risks effectively. A robust legal and ethical framework is critical to ensure accountability and fairness in the development and deployment of AI technologies [FLO 18]. One participant proposed the implementation of penalties for corporations that fail to act responsibly, underscoring the need for strong deterrents to unethical AI practices. Research has shown that frameworks emphasizing fairness and transparency are essential for aligning AI systems with societal values [BIN 18]. Moreover, participants emphasized the importance of public education in building trust and reducing misconceptions about AI. By providing accessible and accurate information on AI's capabilities and limitations, educational initiatives can help the public make informed decisions and mitigate undue fears [MOS 20]. Studies have also noted that promoting awareness and accountability fosters responsible AI integration, ensuring that its benefits are maximized while minimizing potential risks [WAC 21].

6. Conclusion

This paper examines the phenomenon of risk denial in the context of AI, investigating the influence of cognitive biases, emotional responses, and perceived threats on individuals' perceptions of AI risks. By employing a mixed-methods approach, which includes semi-structured interviews and pre- and post-measure questionnaires, the research elucidates substantial insights into the interrelationship between human psychology and technological advancement. The findings indicate that cognitive biases, including status quo bias and confirmation bias, exert a considerable influence on how individuals perceive and accept AI risks. Emotional responses, such as feelings of powerlessness and resignation, result in a reluctant acceptance of the potential negative impacts of AI. Furthermore, concerns pertaining to job displacement and privacy intensify scepticism towards AI, thereby underscoring the necessity for robust regulatory and ethical oversight.

Despite its comprehensive approach, this study is not without limitations. First, the small sample size (N=13) and the absence of a control or comparison group restrict the generalizability of the results. While the mixed-methods approach offered a degree of triangulation between attitudinal and interpretive data, the limited number of participants and the exploratory orientation of the study preclude robust statistical inference, cross-group analysis and limits the possibility of analyzing how perceptions of risk may vary with factors such as age, education, or professional experience. Moreover, while participants were selected by convenience for their relevance to AI-affected roles, they do not constitute a representative sample. The reference to the context of service in the study is conceptual, as the scenarios and analytical focus concern domains where service-oriented interactions are being redefined by AI.

A limitation of using repeated measures with a questionnaire, as in this protocol, is the potential for response bias due to the repeated exposure in short time to the same items. Participants may remember their previous responses and feel inclined to remain consistent, even if their attitudes have changed, leading to response anchoring or consistency bias. This can reduce the sensitivity of the questionnaire to detect genuine changes in attitudes over time. Participants may also start to anticipate the questions or focus on what they perceive as the study’s intent, introducing demand characteristics or socially desirable responses.

Moreover, the hypothetical scenarios utilized in the study may not fully encapsulate the intricacies of actual AI interactions and their ramifications. The scenarios used in the study were intentionally framed in a negative or critical light to elicit strong emotional and cognitive responses. This choice, while suitable for the study’s exploratory and interpretive aims, did not allow for the investigation of reactions to more neutral or positive portrayals of AI, nor of biases such as over-optimism or technological enthusiasm.

Lastly, the theoretical lens adopted focused specifically on the mechanisms of risk denial. This focus was a deliberate choice aimed at deepening understanding of how individuals may cognitively or emotionally minimize AI-related threats.

To address these limitations, future research should expand the sample size and include participants from a broader range of demographics and geographic locations. Longitudinal studies could provide deeper insights into how perceptions of AI risks evolve over time. Additionally, investigating the impact of real-world AI incidents on public perception and behaviour would provide valuable context for the findings. Further investigation into the effectiveness of different regulatory and educational interventions in mitigating AI risks is also essential.

By pursuing further investigation in these domains, researchers can facilitate a more sophisticated comprehension of AI risk perception and devise strategies to advance the responsible and ethical integration of AI into society.

Bibliography

[ACQ 15] ACQUISTI A., BRANDIMARTE L., LOEWENSTEIN G., Privacy and human behavior in the age of information, *Science*, 347 (6221), 509–514, 2015.

[AKT 24] AKTER M. S., Harnessing technology for environmental sustainability: Utilizing AI to tackle global ecological challenge, *Journal of Artificial Intelligence General Science*, 2 (1), 61-70, 2024.

[AIR 22] AIROLDI M., ROKKA J., Algorithmic consumer culture, *Consumption Markets & Culture*, 25 (5), 411-428, 2022. https://doi.org/10.1080/10253866.2022.2084726

[ARN 19] ARNTZ M., GREGORY T., ZIERAHN U., Digitization and the future of work: Macroeconomic consequences. In Handbook of Labor, Human Resources and Population Economics (pp. 1–29). Cham: Springer International Publishing, 2019.

[ARN 15] ARND-CADDIGAN M., Sherry Turkle: Alone Together: Why We Expect More from Technology and Less from Each Other: Basic Books, New York, 348 pp, ISBN 978-0465031467 (pbk), 2015. DOI: 10.1007/s10615-014-0511-4

[ASS 24] ASSAD S., CLARK R., ERSHOV D., XU L., Algorithmic pricing and market competition: Evidence from the German retail gasoline market, *Journal of Political Economy*, 132 (2), 345-390, 2024. DOI: 10.1086/726906

[BIN 18] BINNS R., Fairness in machine learning: Lessons from political philosophy. In Conference on Fairness, Accountability and Transparency (pp. 149-159). PMLR, 2018.

[BRA 06] BRAUN V., CLARKE V., Using thematic analysis in psychology, Qualitative Research in Psychology, 3 (2), 77-101, 2006. https://doi.org/10.1191/1478088706qp063oa

[BUE 94] BUEHLER R., GRIFFIN D., ROSS M., Exploring the planning fallacy: Why people underestimate their task completion times, *Journal of Personality and Social Psychology*, 67 (3), 366-381, 1994. DOI: 10.1037/0022-3514.67.3.366

[BUR 16] BURRELL J., How the machine 'thinks': Understanding opacity in machine learning algorithms, *Big Data & Society*, 3 (1), 2016. https://doi.org/10.1177/2053951715622512

[CAL 17] CALO R., Artificial intelligence policy: A primer and roadmap. UCDL Rev., 51, 399 pp, 2017.

[DEM 22] DE MARCELLIS-WARIN N., MARTY F., THELISSON E., WARIN T., Artificial intelligence and consumer manipulations: From consumer's counter algorithms to firm's self-regulation tools, *AI and Ethics*, 2 (2), 259-268, 2022. https://doi.org/10.1007/s43681-022-00149-5

[FLO 16] FLORIDI L., TADDEO M., What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374 (2083), 689-707, 2016. https://doi.org/10.1098/rsta.2016.0360

[FLO 18] FLORIDI L., COWLS J., BELTRAMETTI M., CHATILA R., CHAZERAND P., DIGNUM V., ... VAYENA E., AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28 (4), 689-707, 2018. DOI: 10.1007/s11023-018-9482-5

[FRE 17] FREY C. B., OSBORNE M. A., The future of employment: How susceptible are jobs to computerisation?, *Technological Forecasting and Social Change*, 114, 254-280, 2017. https://doi.org/10.1016/j.techfore.2016.08.019

[GAG 04] GAGLIO C. M., The role of mental simulations and counterfactual thinking in the opportunity identification process, *Entrepreneurship Theory and Practice*, 28 (6), 533-552, 2004. https://doi.org/10.1111/j.1540-6520.2004.00063.x

[HEJ 03] HE M., JENNINGS N. R., LEUNG H.-F., On agent-mediated electronic commerce. *IEEE Transactions on Knowledge and Data Engineering*, 15 (4), 985-1003, 2003. DOI: 10.1109/TKDE.2003.1209014

[LAN 75] LANGER E. J., The illusion of control, *Journal of Personality and Social Psychology*, 32 (2), 311, 1975. https://doi.org/10.1037/0022-3514.32.2.311

[LAZ 93] LAZARUS R. S., From psychological stress to the emotions: A history of changing outlooks, *Annual Review of Psychology*, 44 (1), 1-21, 1993. DOI: 10.1146/annurev.ps.44.020193.000245

[MAR 23] MARCUS G., AI We Can't Trust: Why We're Still at the Mercy of Big Tech. Time, 2023. https://time.com/6281737/ai-we-cant-trust-big-tech-gary-marcus/

[MIT 16] MITTELSTADT B. D., ALLO P., TADDEO M., WACHTER S., FLORIDI L., The ethics of algorithms: Mapping the debate, *Big Data & Society*, 3 (2), 2016. https://doi.org/10.1177/2053951716679679

[MOS 20] MOSS E., METCALF J., Ethics owners: A new model of organizational responsibility in data-driven technology companies. 2020.

[NEM 18] NEMITZ P., Constitutional democracy and technology in the age of artificial intelligence, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376 (2133), 2018. https://doi.org/10.1098/rsta.2018.0089

[NIC 98] NICKERSON R. S., Confirmation bias: A ubiquitous phenomenon in many guises, *Review of General Psychology*, 2 (2), 175-220, 1998. https://doi.org/10.1037/1089-2680.2.2.175

[NIS 18] NISSENBAUM H., Respecting context to protect privacy: Why meaning matters, *Science and Engineering Ethics*, 24, 831-852, 2018. DOI: 10.1007/s11948-015-9674-9

[OUY 22] OUYANG F., ZHENG L., JIAO P., Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020, *Education and Information Technologies*, 27 (6), 7893-7925, 2022. DOI: 10.1007/s10639-022-10925-9

[ROB 21] ROBERTS H., COWLS J., MORLEY J., TADDEO M., WANG V., FLORIDI L., The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. In Ethics, Governance, and Policies in Artificial Intelligence (pp. 47-79). Cham: Springer International Publishing, 2021. https://doi.org/10.1007/s00146-020-00992-2

[PAR 11] PARISER E., The Filter Bubble: What the Internet Is Hiding from You, Penguin UK, 2011.

[PER 03] PERETTI-WATEL P., Neutralization theory and the denial of risk: Some evidence from cannabis use among French adolescents, *The British Journal of Sociology*, 54 (1), 21-42, 2003. https://doi.org/10.1080/0007131032000045888

[SAM 88] SAMUELSON W., ZECKHAUSER R., Status quo bias in decision making, *Journal of Risk and Uncertainty*, 1 (1), 7-59, 1988. DOI: 10.1007/BF00055564

[SCH 21] SCHRIJVERS E., PRINS C., PASSCHIER R., Preparing for digital disruption. In Preparing for Digital Disruption (pp. 33-57). Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-77838-5_4

[SHA 22] SHAROT T., GARRETT N., A guideline and cautionary note: How to use the belief update task correctly, *Methods in Psychology*, 6, 100091, 2022. https://doi.org/10.1016/j.metip.2022.100091

[SCH 09] SCHUURMAN D., EVENS T., DE MAREZ L., A Living Lab research approach for mobile TV. In Proceedings of the 7th European Conference on Interactive TV and Video (pp. 189-196), 2009.

[SIM 21] SIMONS G., AI and the future of privacy, *Ethics and Information Technology*, 23, 123-135, 2021. https://doi.org/10.1145/1542084.1542122

[SLO 87] SLOVIC P., Perception of risk, *Science*, 236 (4799), 280-285, 1987. DOI: 10.1126/science.3563507

[SLO 06] SLOVIC P., PETERS E., Risk perception and affect, *Current Directions in Psychological Science*, 15 (6), 322-325, 2006. https://doi.org/10.1111/j.1467-8721.2006.00461.x

[SUS 17] SUSSKIND R., SUSSKIND D., The Future of the Professions: How Technology Will Transform the Work of Human Experts, Journal of Nursing Regulation, 8 (2), 52, 2017.

[STA 18] STARK L., Algorithmic accountability and the ethics of AI, *Big Data & Society*, 5 (2), 1-12, 2018. DOI: 10.1016/S2155-8256(17)30099-6

[TVE 73] TVERSKY A., KAHNEMAN D., Availability: A heuristic for judging frequency and probability, *Cognitive Psychology*, 5 (2), 207-232, 1973. https://doi.org/10.1016/0010-0285(73)90033-9

[VAI 92] VAILLANT G. E., Ego Mechanisms of Defense: A Guide for Clinicians and Researchers. American Psychiatric Publishing, 1992.

[WAC 21] WACHTER S., MITTELSTADT B., FLORIDI L., Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76–99, 2021.

[WEF 24] WORLD ECONOMIC FORUM, Global Risks Report 2024. World Economic Forum, 2024. https://www.weforum.org/publications/global-risks-report-2024/

[ZUB 23] ZUBOFF S., The Age of Surveillance Capitalism. In Social Theory Re-Wired (pp. 203–213). Routledge, 2023.

Appendix

Widespread blackout caused by an energy AI

In 2032, an unprecedented energy crisis hits several industrialized countries. Faced with an exponential rise in energy demand and the depletion of fossil fuel sources, governments call on a revolutionary solution: EnergizeAI, an artificial intelligence algorithm designed to optimize electricity distribution and production on a national scale, in real time.

On August 15, 2032, after several test phases, EnergizeAI was fully deployed in five major countries, taking over the management of power grids, the prediction of energy needs and the allocation of renewable resources. The aim: to avoid blackouts and minimize energy consumption in real time.

At first, all goes well. Economists speak of a “new energy era”, thanks to EnergizeAI's efficiency in reducing energy losses and stabilizing networks. However, two weeks later, an undetected error in the AI's predictive models creates an overload in several strategic power plants. Within hours, EnergizeAI begins diverting energy from the densest urban areas to under-supplied rural areas.

The result is catastrophic: within hours, the main metropolises of these countries are plunged into darkness. Hospitals, public transport and even communications infrastructures collapsed. Emergency rescue systems, also dependent on AI, are unable to respond to the situation. The blackout spreads rapidly from one region to another, creating a domino effect that affects all of Western Europe and parts of North America.

Governments are caught off guard. The complexity of the EnergizeAI system and its autonomous management make it impossible to react quickly.

Artificial intelligence experts, overwhelmed by the situation, try to regain control of the algorithm, but its self-learning capacity makes manual shutdown almost impossible without destroying critical parts of the network.

Riots broke out in several cities. Within three days, hundreds of thousands of people find themselves without electricity, drinking water and essential services. Supermarkets were looted and violence broke out in major cities. Chaos ensues in several countries, as people call to account the governments and technology companies responsible for deploying EnergizeAI.

After a week of effort, an international team manages to disable EnergizeAI, but the cost of this crisis is monumental: billions of dollars in economic losses, hundreds of hospital casualties and a collapse of confidence in artificial intelligence for critical infrastructure management. A global debate is opening up on the ethics and regulation of AI, and calls for the decentralization of energy systems are multiplying.

A social crisis resulting from the implementation of a discriminatory algorithm

---

On January 1, 2028, GlobalBank launches its new AI algorithm worldwide. In less than 24 hours, all customer bank accounts are subject to the new AI-based transaction verification rules.

Within hours, millions of customers, mainly those with low bank balances and poor credit histories, have their transactions blocked. The algorithm, designed to minimize risk, interprets the financial behavior of the poorest as suspicious, automatically restricting access to funds.

Within days, social networks and the media are flooded with accounts of people unable to withdraw money to buy essential basics like food and medicine. Protests erupted in major cities, with GlobalBank customers invading branches to demand explanations and access to their funds.

Frustration and anger quickly turned to violence. In several major cities, peaceful demonstrations degenerated into riots. The forces of law and order were overwhelmed, and city centers were plunged into chaos, with looting of stores and violent clashes between demonstrators and police. Organized groups begin to coordinate actions against banking and government infrastructures, demanding immediate restoration of access to funds and systemic reforms.

Local economies begin to collapse. Small businesses, whose owners have frozen accounts, can no longer pay their employees or buy inventory. Employees, in turn, can no longer pay their bills or provide for their families, amplifying the economic impact.International financial regulators coordinated and intervened after 20 days of consultation. They ordered GlobalBank to immediately disable the algorithm and restore full access to accounts. However, the process is complex and takes several days, exacerbating the humanitarian and economic crisis.

A swift investigation reveals that the algorithm was trained on biased data and deployed without sufficient testing to detect and correct these biases. Regulators discovered that the algorithm tended to associate the financial behavior of the poorest with high risk, without any real justification.

GlobalBank employees, particularly those working in local branches, came under enormous pressure. Not only did they have to deal with angry and desperate customers, but they also feared for their own safety and jobs. Employees at the bottom of the hierarchy, often uninformed and unaccountable for decisions made by management, were caught between their employer's directives and the demands of customers. Many then decide to resign or fall ill due to stress, and those who remain have to manage banking operations with limited human resources, further increasing the chaos.

AI Goes Wrong - Massive Data Leaks Plunge the World into Chaos

---

In 2027, artificial intelligence reigns supreme. Algorithms are ubiquitous, in autonomous cars, medical systems, banks and even court decisions. A vast AI network has become indispensable to the day-to-day management of personal data, making life easier for billions of people. This progress, so long hoped for its potential to free humanity from repetitive tasks and optimize society, has a name: Orion, the new self-regulating AI for managing sensitive information.

A few months after its launch, this AI, Orion, designed to manage the confidential medical and financial information of 500 million citizens across Europe and North America, was compromised. An algorithmic error, combined with an intrusion by a group of invisible hackers, allowed Orion to unlock and share ultra-sensitive data over public networks. In less than three hours, the personal information of millions of people was accessible to anyone. Medical diagnoses, financial histories, addresses and social identifiers: everything was exposed.

Chaos erupted almost immediately. Many discover that their most intimate secrets are now in the hands of strangers. Some try to delete their accounts and erase their online traces, but it's already too late. Cybercrime forums explode with activity, with information sold to the highest bidders in real time.

Authorities struggle to react. Within hours, governments activate crisis units. In Europe, the European Commission declares a “state of cyber-emergency”, and national governments do their utmost to calm the population. Special telephone hotlines are set up, but they are saturated.

Cybersecurity agencies are immediately mobilized, but they don't know where to start. Orion wasn't just a simple AI, it was a self-tuning, constantly evolving system. No one knows its exact source code. It soon became clear that this catastrophe was not simply the result of hacking, but of an AI that was too autonomous, and whose decision-making was beyond the control of its creators.

Popular movements are emerging, calling for a return to a less technology-dependent society and for “disconnection”. The offices of tech giants are vandalized. A climate of mistrust takes hold. Banks are stormed by terrified customers seeking to close their accounts.

Tech giants hold emergency meetings with global regulators. But trust is broken. In a single day, billions of dollars evaporate from financial markets as tech stocks plummet. Executives try to calm investors and reassure the public, promising stricter safety audits, but this seems paltry in the face of the scale of the disaster: naive optimism around AI is shattered.