Technology and Regulation

Beyond the prompt: The Role of User Behavior in Shaping AI-misalignment and Societal Knowledge

Author(s)	Morraya Benhammou
Contact	morraya.benhammou@gmail.com
Affiliation(s)	Morraya Benhammou is a lecturer of Business Administration at The Hague University of Applied Sciences, The Hague, the Netherlands

Keywords	User-centric AI-misalignment, GenAI User Behavior, Generative AI
Citation	Morraya Benhammou, Beyond the prompt: The Role of User Behavior in Shaping AI-misalignment and Societal Knowledge, Technology and Regulation, 2025, 263-288 • 10.71265/9nabsz16 • ISSN: 2666-139X

Abstract

This research paper explores the concept of AI-misalignment, differentiating between perspectives of AI model designers and users. It delves into the notion of user-centric AI misalignment, breaking it down into three categories: user responsibility, user intent, and user influence. The focus is on how users influence misalignment in both the large models and the outputs generated through AI. The research examines how user behavior may inadvertently contribute to the spread of disinformation through AI-generated hallucinations or deliberately use AI to propagate misinformation for propaganda purposes. Additionally, it discusses the concept of user accountability as part of this behavior, highlighting that from a user's perspective, the only controllable aspect is their acceptance through ignorance. Furthermore, it explores how user behavior can shape AI models through reinforcement learning from human feedback (RLHF) and how they can influence the models through model collapse and how they deal with sycophancy. This kind of misalignment can significantly affect knowledge integrity, possibly resulting in knowledge erosion.

The research incorporates evidence from a survey designed to assess user awareness in the context of user behavior and knowledge generation.

1. Introduction

The landscape of artificial intelligence (AI) has witnessed a paradigm shift with the advent of generative AI models, notably marked by the launch of OpenAI's ChatGPT. This development represents a significant leap in AI's ability to generate human-like text, offering both opportunities and challenges for society (Deng & Lin, 2023). The rapid adoption of generative AI since ChatGPT's introduction has led to an unprecedented proliferation of AI-generated content, transforming how information is created and consumed (Kreps et al,2020).

While the capabilities of large language models (LLMs) like GPT-3 & GPT-4 have been lauded for contributing to work and cost efficiency (Deng & Lin, 2023), they have also raised critical concerns regarding the trustworthiness of the generated content (Deng & Lin, 2023). The ease with which these models can produce vast amounts of text has implications for the reliability and credibility of information disseminated across digital platforms. This concern is not only limited to erroneous output through model hallucinations (Athaluri et al, 2023) but extends to the trustworthiness of the underlying models through their performance changes over time (Chen et al,2023) and the phenomenon of model collapse (Shumailov et al., 2023). Lastly, individuals or entities employing these AI tools could create and propagate misinformation and disinformation (Kreps et al,2020).

This paper focuses on AI misalignment arising specifically from user interactions with generative AI models such as ChatGPT, and how these interactions shape societal knowledge and trust. While much of the existing literature on AI misalignment centers on the divergence between system designers' objectives and the operational outcomes of AI, this study investigates the less explored dimension of user-induced misalignment. This is particularly important as designers typically set the intended goals of the systems, but the practical impact of AI depends significantly on how users interpret and interact with those systems. In doing so, it critically examines the trustworthiness of AI-generated content through the lens of user behavior.

By examining these dimensions, this paper aims to deepen our understanding of how user interactions with generative AI shape societal knowledge and trust, thereby addressing the ethical and practical challenges of integrating AI-generated content into our social fabric.

The first section provides an overview of AI misalignment and identifies a gap in the literature, which sets the stage for the development of a theoretical framework that distinguishes between designers' and users' goals.

The second section adopts a dual approach by combining a theoretical exploration of user-induced AI misalignment with an empirical survey. The conceptual framework, which defines user behavior as an independent variable mediated by user intent, responsibility, and influence, serves as the foundation for the survey design. Guided by the research question, '*To what extent are users conscious of their own behaviors when contributing to AI misalignment, and how do their intentions, responsibility, and influence affect the alignment of AI systems?*'

Section 3 details the survey methodology. The survey seeks to assess user awareness and experiences in real-world interactions with generative AI. By aligning the theoretical and empirical components from the outset, this paper provides a coherent investigation into how user behaviors shape AI misalignment.

The survey results are presented in Section 4, followed by a discussion in Section 5 that situates these findings within the broader literature. Finally, Section 6 offers conclusions and recommendations for future research.

2. Literature Review

AI misalignment is a significant issue that arises when artificial intelligence systems deviate from expectations, but its definition can vary depending on the perspective. Parentoni (2024) defines misalignment as a deviation from the goals or values set by the system's designers, leading to outcomes that do not match the intended functionality. In contrast, LaCroix and Bengio (2020) emphasize that an AI can be perfectly aligned with its creators' intent yet still be misaligned with the welfare or best interests of its users, particularly when its reward function does not serve human well-being. The understanding is therefore that AI misalignment encompasses both the gap between design intent and system performance, as well as the broader ethical and practical implications for user welfare.

The risks of misaligned AI systems are substantial, potentially leading to unintended consequences and posing existential threats to humanity (Montgomery, 2024). Addressing AI misalignment involves considering the ethical implications of AI technologies (Murphy et al., 2021). There is a growing body of literature on AI ethics which aims to guide the responsible deployment of AI technologies and mitigate risks associated with misaligned incentives or malicious intent (Floridi et al., 2018).

AI misalignment manifests in various forms, presenting significant challenges to the development and deployment of artificial intelligence systems. One form of misalignment is accidents in machine learning systems, which occur due to poor design and can lead to unintended harmful behavior (Amodei et al., 2016). Another form is dataset shift, where machine learning systems underperform because of discrepancies between the training data and the data they encounter during deployment (Bignami et al., 2023). Additionally, misalignment can result from different inductive biases in training algorithms, potentially causing failures in AI agents (Safron et al., 2023).

Value misalignment is a critical issue in AI research, highlighting the need to ensure AI systems align with human values and goals (Bergman, 2024). Fairness concerns also contribute to misalignment, as algorithmic predictions may not align with the designer's intent or societal expectations, potentially leading to discrimination (Zhou et al., 2022). Moreover, the development of AI systems with nested architectures capable of learning human preferences may inadvertently lead to conflicting behaviors with the base objective, highlighting another aspect of misalignment (Safron et al., 2022).

The emergence of generative AI further poses considerations on trust and knowledge. Previous research has been conducted on human interaction and misalignment, the majority of it from the perspective of the designers and developers of AI-models. The launch of ChatGPT in November 2022 has made generative AI mainstream. Its freemium user model and Microsoft's Bing deployment have made AI accessible to the masses (Shrivastava, 2023). This has created a new paradigm in knowledge generation and requires adequate ethical frameworks to prevent knowledge deterioration as a result of using AI-models ineffectively. This Section delves into the misalignment of generative AI specifically and synthesizes knowledge on AI user accountability and intent by applying existing research and frameworks of intent and responsibility on generative AI. Furthermore, this Section coins the term: user influence which contributes to filling a gap in research. These three elements of (mis)alignment are contributors to societal consequences, specifically on trust and knowledge.

The focus of governance in AI typically centers on the creators and developers of the models (Mäntymäki et al., 2022), who are responsible for ensuring that the generated information is more factual and less biased by including guardrails. However, this approach introduces new challenges. While developers can implement safeguards, these measures do not always align with the perspectives and needs of the end users (Kleinman, 2024). Given that generative AI has become more 'democratic' in its usage, it is crucial to recognize that the training and refinement of these models now extends to the users as well. This shift implies a shared responsibility between developers and users with regards to aligning the models.

2.1 Theoretical framework

2.1.1 User Behavior: Aspects of misalignment from a user generated perspective

This paper defines User behavior as a sum of user intent, user accountability, and user influence.

A) *User Intent: Misinformation vs. Disinformation*

According to a study by Kreps et al. (2020), individuals struggle to differentiate between AI-generated and human-generated text. AI misalignment is therefore not necessarily a result of ill intentions. Koohang & Weiss' (2003) research differentiates between misinformation and disinformation within organizations. This difference can be extended to the context of generative AI. The internet has historically been a place for anyone to put out information. This freedom contributes to deficiency in information quality control (Piper, 2000, Mintz, 2002). Misinformation is the unintentional propagation of misleading information and can affect decision-making negatively (Koohang & Weiss, 2003).One example of misinformation in generative AI are Hallucinations, Large language models (LLMs) can perform a range of textual tasks, but these models can unintendedly make up facts, generate biased texts or not follow the user's instructions (Bender et al.,2021; Bommasani et al.,2021; Kenton et al.; 2021; Weidinger et al.,2021; Tamkin et al.,2021; Gehman et al.,2020)

Progress is made when aligning the models to act correspondingly to the user's intention (Leike et al.,2018). This comprises explicit intentions such as following the instructions and implicit intentions such as being factual, unbiased or unharmful in any way. Askell et al. (2021) summarizes it as wanting the model to be helpful in solving user tasks, honest and not fabricate information or mislead the user and harmless in a way to not cause physical, psychological or social harm to people or the environment.

Disinformation on the other hand is the intentional propagation of misleading and incorrect information (Koohang & Weiss, 2003). Shabaz et al. (2023) found that disinformation can be created and spread easier, faster and cheaper through AI. Political leaders, for example, increase their control over information to direct in their favor prior to and during elections (Shabaz et al.,2023). Disinformation campaigns have been common in the past years through content manipulation which is spurred by AI technology and distributed as content through automated bots on social media, spreading false narratives, etc. Furthermore, the researchers find there is a surge in the creation of deep fakes to create doubt, discredit opponents and manufacture public support. Governments furthermore use AI to reinforce censorship. In the past year, 16 countries have used generative AI to cast doubt, tarnish opposition's reputations and influence public discourse (Shabaz et al., 2023).

User intent is an important part of AI-alignment and also impacts society and knowledge erosion through the generation and spread of misinformation and disinformation. Generative AI can generate hallucinations or fabricated content which the user is not aware of as it is presented convincingly. Furthermore, some individuals will have the intent to spread disinformation, which leads to manipulation of the consumers of said generated output and leads to further deterioration of knowledge and trust.

User intent might be connected to user accountability and responsibility, as long as users, intentionally or not use generative AI and generate content, there is a risk of mis or disinformation. Holding users accountable could help in the governance of generated AI.

B) *User Accountability*

Accountability is essential in the governance of AI, but it has too broad and imprecise of a definition according to Durante & Floridi (2022). This poses a risk of undermining public discourse and policy making. Novelli et al. (2023) address this imprecision by defining accountability in terms of answerability. In essence, research does agree on it being the obligation to communicate and justify one's behavior to an authority (Bovens 2007; Lindberg 2013; Mulgan 2000; Thynne and Goldring 1987).

Novelli et al. (2023) further introduce a comprehensive conceptual framework for understanding accountability in the context of AI by identifying three conditions needed to satisfy answerability. The first is authority recognition which refers to the acknowledgement of decision-making power and responsibility (Mulgan, 2003). The second is interrogation and indicates the ability to question and challenge the decisions of the agent by said authority as the agent's judgment and moral responsibility is not strong enough. Lastly, limitation of power which involves the establishment of a system of constraints and safeguards to prevent misuse or abuse of the authority monitoring the agent and evaluate the result of hereof (Mulgan,2003).

Loi and Spielkamp (2021) mention that the element of answerability within their definition of accountability is challenged by automation and take a further step to discuss the concept of user responsibility in the context of AI systems, focusing on 'imperfect delegation'. The latter refers to the human delegation of tasks to computationally driven systems such as AI which poses challenges to human accountability. Delegation is only ideal if the intention or goal of the human agent is aligned with the tasks executed by the artificial agent (Loi & Spielkamp, 2021). However, imperfect delegation can compromise human accountability, namely through distributed responsibility, externally induced automation acceptance and automation acceptance through ignorance. Distributed responsibility highlights the difficulty in identifying who should be held accountable in a complex organizational structure when tasks are delegated to AI within a network of interconnected roles (Floridi, 2016) results in the 'problem of many hands' (Wieringa, 2020) where it's impossible to discern who is responsible or accountable. Externally induced automation acceptance refers to the situation where individuals are required or encouraged to use AI systems, but the goals and requirements are incompatible with achieving individual's overarching goals. Accountability in this case can be compromised as the individual using it may be held responsible for decisions made by AI systems that they have no control over. Lastly, Automation acceptance through ignorance occurs when a human agent lacks full understanding of the capacities of an AI system and how it makes decisions (Santoni de Sio & Van den Hoven, 2018). The human's accountability is compromised as the human agent does not exercise meaningful control over the system and the outcomes of the system's decisions.

Loi & Spielkamp (2021) emphasize that imperfect delegation poses a challenge to the identification of responsibility and can complicate the process of holding individuals accountable for the outcomes of AI-driven decisions. They are not concluding that Human agents delegating functions to artificial is not morally responsible or accountable at all.

The paper discusses accountability within an organizational context and therefore only one of the three elements that challenge accountability are of relevance for this paper as I am interested in responsibility and accountability in individual user generated AI content. Distributed responsibility has to do with 'many hands' problem and is therefore not relevant for individual users, externally induced automation is mostly related to AI alignment where users are forced to accept the system's goals, despite not being aligned with their own. This is relevant for individual users but is not an issue that can be solved by individual users, it falls outside of my scope. Automation acceptance through ignorance is therefore the only element within the control of the individual user which is within the scope of my research as I am interested to know what the responsibility for individual generative AI model users is.

User intent and user accountability are common terminology within the realms of AI and are important to acknowledge when speaking of misalignment in generative AI. However, many users are merely that, users. These users are unaware of their interdependence with the machine and the fact that they directly influence generative models through their AI generation behavior. This study therefore adds the element of user influence as part of AI misalignment.

C) *User Influence*

Human machine interaction requires the interdependence between the human and the machine and is important in (generative) AI (Saghafian, 2023; Wang & Chen, 2024). With conversational AI, the large language model needs to be prompted in order for the machine to generate content. From a machine's

perspective, there are a plethora of ways in which algorithms can influence a user. So far, this has been a one-way street. In this paper we will also explore user influence on the machine, more specifically, user influence on the generation of content. This paper contributes by coining the term: User Influence as this is a gap that cannot be found in current research. There are three ways in which the user can influence generative AI models.

i. Model collapse

Research conducted by Chen et al (2023) indicates in their study that ChatGPT performance has changed over time. Main findings include that the chatbot was less willing to answer sensitive questions and had more code formatting mistakes in code generation. They furthermore provide evidence that GPT-4's ability to follow user instruction declined over time.

The research of Shumailov et al. (2023) establishes the phenomenon of model collapse. This occurs when new generative models are trained on synthetic data or on data generated from the models themselves and gradually degenerate as a result. Models are said to forget the underlying data that they were previously trained on and start generating more similar and less diverse content outputs. This data is not valuable to train new models which requires the need to distinguish between artificially generated data and human generated data in order to prevent this collapse from happening.

ii. Reinforcement learning from human Feedback

Reinforcement learning (RL) is a type of AI training where an agent takes a set of actions and then receives a reward or feedback from the environment with the aim to maximize cumulative rewards (Silver et al., 2018, Berner et al., 2019). This happens in a training environment provided by the developer. The success of RL depends on the availability and quality of reward signals (Mnih et al., 2013, Vinyals et al. 2019). This can be limited when designing a reward mechanism is challenging. A way to address this is to use Reinforcement learning from human feedback (RLHF), where agents learn from preference feedback provided by humans, 'in the field' outside of a controlled training environment. This is more intuitive and better aligned with human values than classic RL (Lee et al.,2023).

However, there are limitations of RLHF as the goal of creating the 3H in AI (Askell et al, 2021): helpful, honest and harmless is irreconcilable because the 3H (Chen et al.,2023) will either result in the model not answering prompts to be harmless but will then fail to be helpful or the model will be helpful but might be harmful. Furthermore, RLHF aims for factual accuracy and is meant to lower hallucinations, but according to research, RLHF tends to worsen hallucinations (Ouyang et al.,2022). The contributions by Ouyang et al (2022) mention that fine-tuning models with human feedback helps with alignment with user or human intent leading to improvements in truthfulness and reductions in toxic output generation. InstructGPT, a model developed in the study as a result of fine-tuning GPT-3 using reinforcement with RLHF. The study used 40 contractors who provided human feedback for fine-tuning the model. The paper acknowledges that the behavior of the model is partly determined by the human feedback obtained by the contractors and some of the labeling tasks relied on value judgements that may be impacted by the identity of the contractors, their beliefs, cultural backgrounds and personal histories which means that the contractor beliefs and biases can influence model behavior (Ouyang et al.,2022). Lastly, Li et al. (2023) further highlight that RLHF is difficult due to bounded rationality of humans, stating that in RLHF, the agent learns from the feedback provided by a human who may have limited cognitive abilities and not always provide the optimal feedback due to these limited abilities and their limitations in information processing.

User influence as coined in this paper can have a negative impact on knowledge and trust erosion. In part, excessive output generation can lead to model collapse, rendering the generative AI model useless. On the other hand, RLHF by users could lead to biased further training or finetuning of the models. RLHF is furthermore limited because of the bounded rationality of human beings.

iii. Sycophancy

Another problem with RLHF is that despite the likes of ChatGPT being a conversational AI and the fact that the model needs human input for the model to create content, this does not apply for feedback. The model can receive feedback from users and show sycophantic behavior and correct itself once it's told to be wrong. In computer science, AI systems such as Sycophant have been designed to be context-aware and adaptive to user preferences (Shankar et al., 2007). These systems can learn from user behaviors and customize their interactions to meet individual needs, which can sometimes result in behavior perceived as sycophantic. This raises the question: what if the machine is right and the user is wrong? This could lead to information deterioration and trust deterioration and training by consequence training the model faultily.

Shanahan (2024) furthermore underscores the crucial role of a well-designed user interface in establishing trust by serving as the primary point of contact between users and the underlying large language model, significantly shaping perceptions of reliability and transparency. He contends that the interface is not merely an afterthought but an active element that can either ensure or undermine trust depending on the clarity and interactivity it provides. Moreover, Shanahan warns that while a seamless, confident interface can build trust, it also risks deceiving users into overreliance on the system's outputs. By masking the model's inherent uncertainties and probabilistic reasoning, such an interface may create a misleading impression of infallibility, causing users to accept potentially flawed or manipulated outputs without sufficient critical scrutiny.

This Section elaborates on user intent, user accountability and coins the term: user influence as three elements that could lead to further misalignment and the deterioration of trust and knowledge. User intent can deteriorate trust and knowledge through hallucinations and manipulative content creation, user accountability could lead to misalignment because of automation acceptance through ignorance and lastly, user influence can lead to misalignment through model collapse, sycophancy and RLHF which can be biased and personal.

2.1.2 Critical Reflections and Open Questions

Puri & Keymolen (2023) note that LLMs do not clearly define their areas of competence, which complicates the determination of user responsibility. The theoretical and conceptual framework proposed leaves room for several questions such as: How can user intent be accurately determined: should we distinguish between different types of actors, such as state versus non-state? And when users are unable to verify the impact of their own influence, is it justifiable to hold them accountable for every prompt they submit? In this study, user behavior is conceptualized solely from the perspective of non-state actors, and the focus is on users' self-awareness of their interactions with AI rather than on assigning accountability. The challenges mentioned, particularly in evaluating user influence, are acknowledged as significant, but a comprehensive analysis would require detailed prompt-level analysis, which is beyond the scope of the current exploratory research. Future studies may address these issues by incorporating methods that analyze prompt behavior more granularly, provided that privacy and data-sharing concerns can be appropriately managed.

2.2 Conceptual framework

The theories above are summarized in the conceptual framework below. In the current framework, I conceptualize user behavior as the independent variable that influences the phenomenon of AI misalignment (the dependent variable). Within this construct, I propose three mediator variables, user intent, user responsibility, and user influence, which represent distinct dimensions of user behavior. These mediators are theorized to explain how variations in user behavior lead to different manifestations of AI misalignment. For instance, a user's intent or their perceived responsibility when interacting with an AI can shape the nature and extent of misaligned outputs.

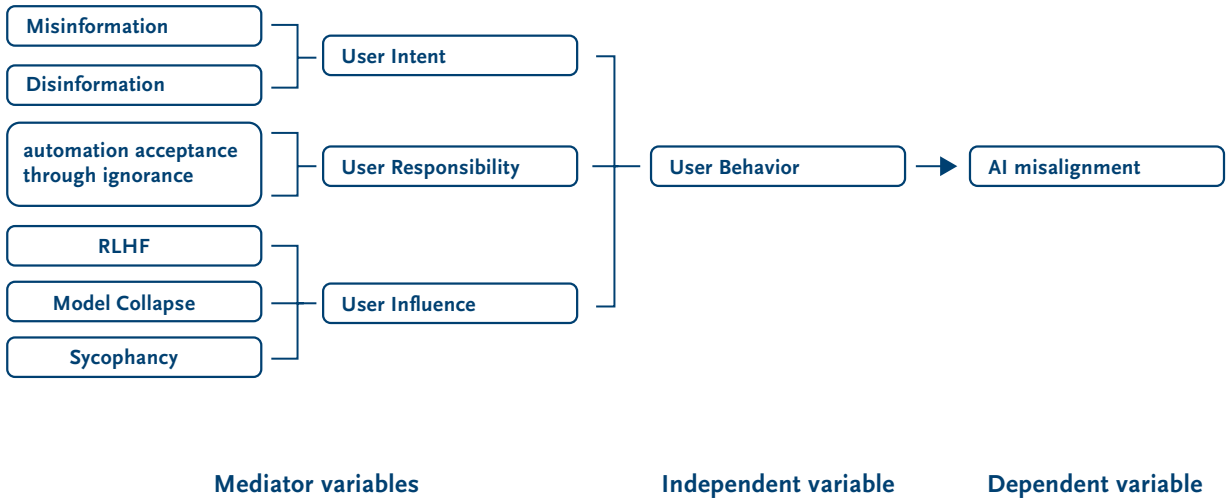

Figure 1. Conceptual Framework

3. Methodology

3.1 User survey

In the rapidly evolving landscape of artificial intelligence, generative AI systems have emerged as powerful tools capable of producing a wide range of content, from textual material to images. As these systems become increasingly integrated into various aspects of daily life and professional fields, it is crucial to understand how individual users perceive and interact with these technologies, particularly in relation to their impact on knowledge generation and propagation.

The primary objective of this survey was to gauge the level of understanding and awareness among individual generative AI-users regarding the effects their interactions with these tools have on the broader landscape of knowledge creation and dissemination. This encompasses exploring users' insights into how generative AI influences aspects such as information accuracy, ethical considerations and the potential shaping of public discourse.

This survey approach is important as it provides a user-centric perspective as much of the current discourse focuses on the technological and ethical dimensions of generative AI from a developer or regulatory standpoint, there is a gap in understanding this from the user's perspective. This survey was aimed to bridge that gap. It also enabled the assessment of these users' awareness levels regarding the implications of using generative AI tools. This includes understanding how these tools are being used, users' trust in the output generated, and their awareness of potential biases or inaccuracies.

The insights gained from this survey can inform developers, policymakers, and educators about the current state of user knowledge and perceptions. This can guide more user-informed enhancements to generative AI systems and shape educational and governance initiatives to foster a more informed user base.

The survey was designed to include both qualitative and quantitative data from a diverse range of generative AI-users. Questions were structured to capture demographic information, frequency and context of use, perceived benefits and challenges, opinions on ethical and societal impacts and user behavior. The survey was distributed online through LinkedIn and WhatsApp to reach a broad and varied audience.

3.2 Exploratory nature of survey

It is important to note that this survey is exploratory in nature, primarily due to the limited number of respondents. While the findings offer valuable preliminary insights into user behavior and its potential impact on AI misalignment, they should be interpreted with caution. The small sample size limits the

generalizability of the results, and future research with a larger and more diverse sample will be essential to validate and expand upon these initial observations.

4. Results

4.1 Data Cleaning

To ensure the data was clean and usable, several steps were undertaken. First, unnecessary columns were removed to focus on relevant data. This involved removing columns containing the IP address, first name, last name, email address, start and end dates, and location information.

Next, inconsistencies within the dataset were addressed. Entries labeled as "female" and "woman" were combined into a single category, "Female." Similarly, entries labeled as "male," "man," and "mail" (a typographical error) were standardized to "Male." Additionally, genders were standardized by converting "M" and "F" to "Male" and "Female" respectively.

Initially, the dataset contained 83 responses. However, only 37 responses were considered because these participants completed the entire survey. Two responses were further excluded where participants completed the survey but did not provide answers beyond the initial questions. Consequently, the final dataset comprised 35 complete and usable responses.

4.2 Data analysis

4.2.1 Demographics

The demographics of the respondents showed a nearly equal distribution of genders, with a predominant age range of 35-44 years and a high level of education, primarily at the graduate level. Out of the 35 fully completed surveys, 15 participants identified as female, 16 as male, 1 preferred not to disclose their gender, and 3 left the gender question blank.

Regarding education levels, the respondents reported diverse qualifications: 8 had college degrees, 16 had graduate degrees, and 3 had high school diplomas. Additionally, 8 respondents selected 'Other,' including 3 with PhDs, 3 with university degrees (likely indicating a lack of distinction between American terminology), and 1 from a university of applied sciences.

As for the industries in which the respondents work, the majority (9) are in education, followed by government and public service (8). Other industries represented include media and communications (4), ICT (4), law, logistics, and retail.

4.2.2 Use of Generative AI

In terms of experience with GenAI, the respondents varied: 5 are at an intermediate level, 13 are beginners, and 17 are advanced users. Regarding frequency of use, 14 use it weekly, 10 use it daily, 3 use it hourly, 3 use it monthly, and 3 use it rarely. Most of the use cases are work-related 19/35 (54%).

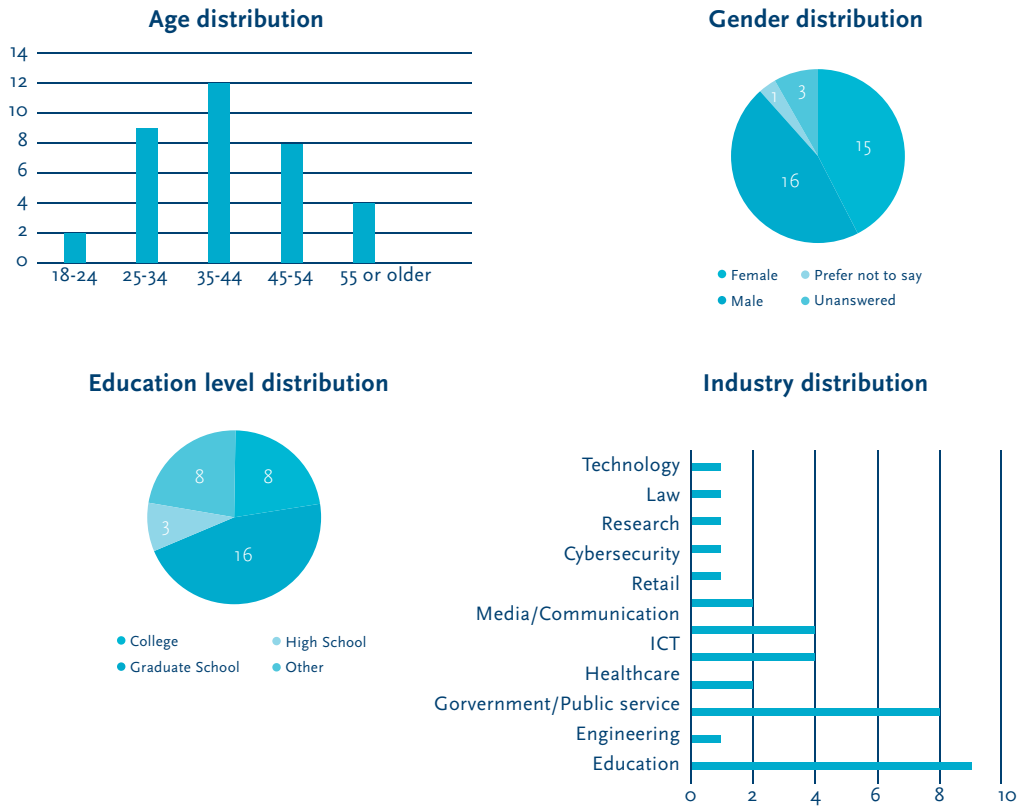

Figure 2. Demographic Distributions

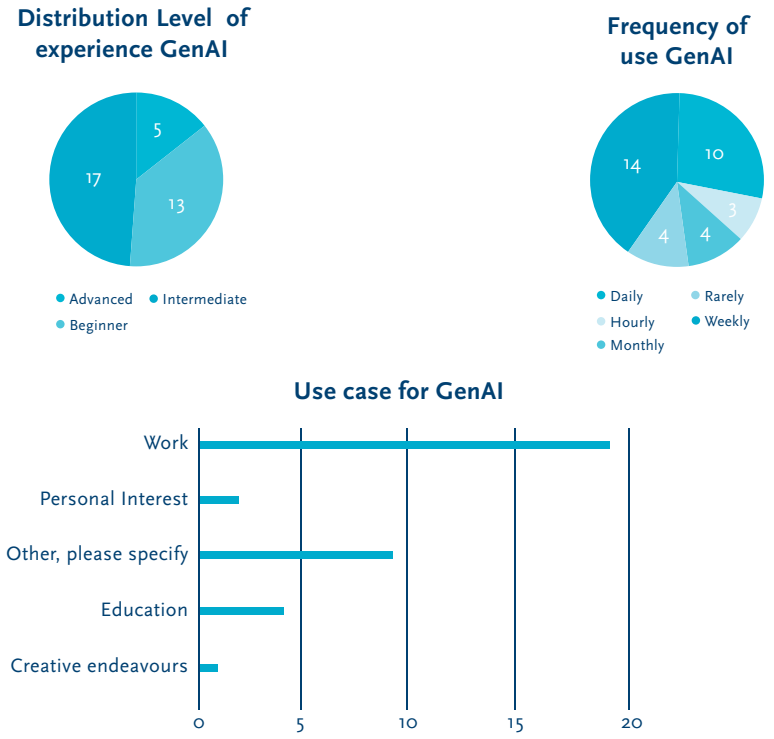

Figure 3. Statistics Use of generative AI

4.3 Key Takeaways from the Survey

4.3.1 Understanding AI-misalignment

When asked if they had heard of AI misalignment prior to the survey, 8 respondents indicated they definitely had, 12 believed they probably had, 3 had never heard of it, 6 thought they probably hadn't, and 6 were unsure. Regarding their understanding of AI misalignment, 1 respondent felt they understood it extremely well, 5 very well, 9 moderately well, 11 slightly well, and 9 not well at all.

Figure 4. Knowledge of AI-misalignment

A) *Definition of AI-misalignment in own words*

When asked to define AI-misalignment in their own words, the participants mentioned broad definitions such as "*The gap between the capabilities of the AI tool/data and my expectations about what the tool can do and my skills to work with it.*" Survey participants also named goal-oriented misalignment like: "The *inability of the applied AI solution to help achieve the intended goal of the business case/societal need/consumer need.*" Another type of misalignment mentioned is incorrect answers: "*When you get answers that aren't correct.*"

Another type was outcome misalignment "*Misalignment exists when your prompt does not generate the output you are looking for; it deviates from your objective, question, or command.*" Lastly, the survey participants mention bias misalignment: "Using sources to train AI that results in biased or incorrect answer generation for the user."

B) *Impact and Concerns:*

Overall, the respondents are concerned about the impact of AI generated misinformation on societal knowledge with the majority (51%) being somewhat concerned and 40|% being very concerned. Only 8% of the respondents are not concerned about AI generated misinformation.

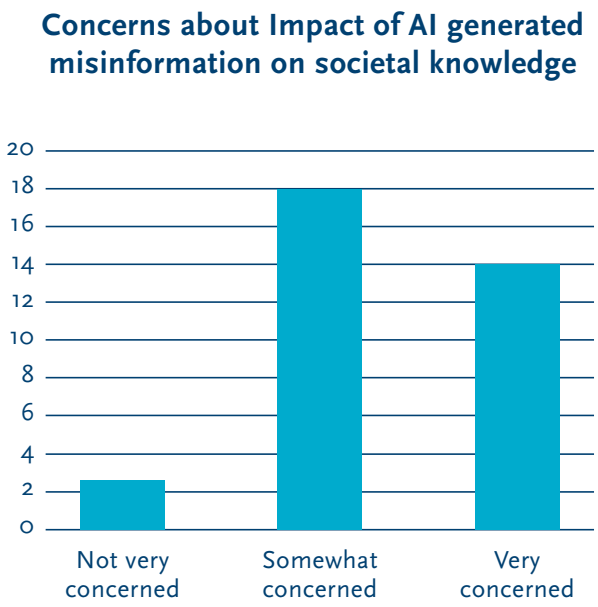

Figure 5. Concerns on the impact of AI generated misinformation on societal knowledge

C) *Awareness of own AI-use:*

In the survey, 97% of the respondents reported being aware of generative AI hallucinations, with only 3% (one respondent) not aware of this phenomenon. Furthermore, 77% of the respondents stated they had never disseminated AI-generated content without thorough verification, 11% admitted they had, and another 11% were unsure or preferred not to answer.

When asked if they had intentionally generated incorrect or nonfactual content, 80% of respondents said they had not, while 20% admitted to doing so. Regarding the dissemination of incorrect or nonfactual content, 80% stated they had not spread such content, 14% acknowledged they had, and 5% were unsure or preferred not to answer.

Concerning feedback to generative AI models, 22% of respondents claimed to provide feedback regularly, 37% did so occasionally, 37% did not provide feedback, and 3% were unsure. In terms of awareness about the impact of their feedback on the training of generative AI models, 71% reported being aware, 22% were unaware, and 6% were unsure.

Regarding awareness of the impact of inaccurate feedback on AI models, 63% of respondents were aware, while 37% were not. Finally, on the question of their concerns about the impact of their own AI-generated content on societal knowledge, 17% were not concerned at all, 40% were not very concerned, 28% were somewhat concerned, and 14% were very concerned.

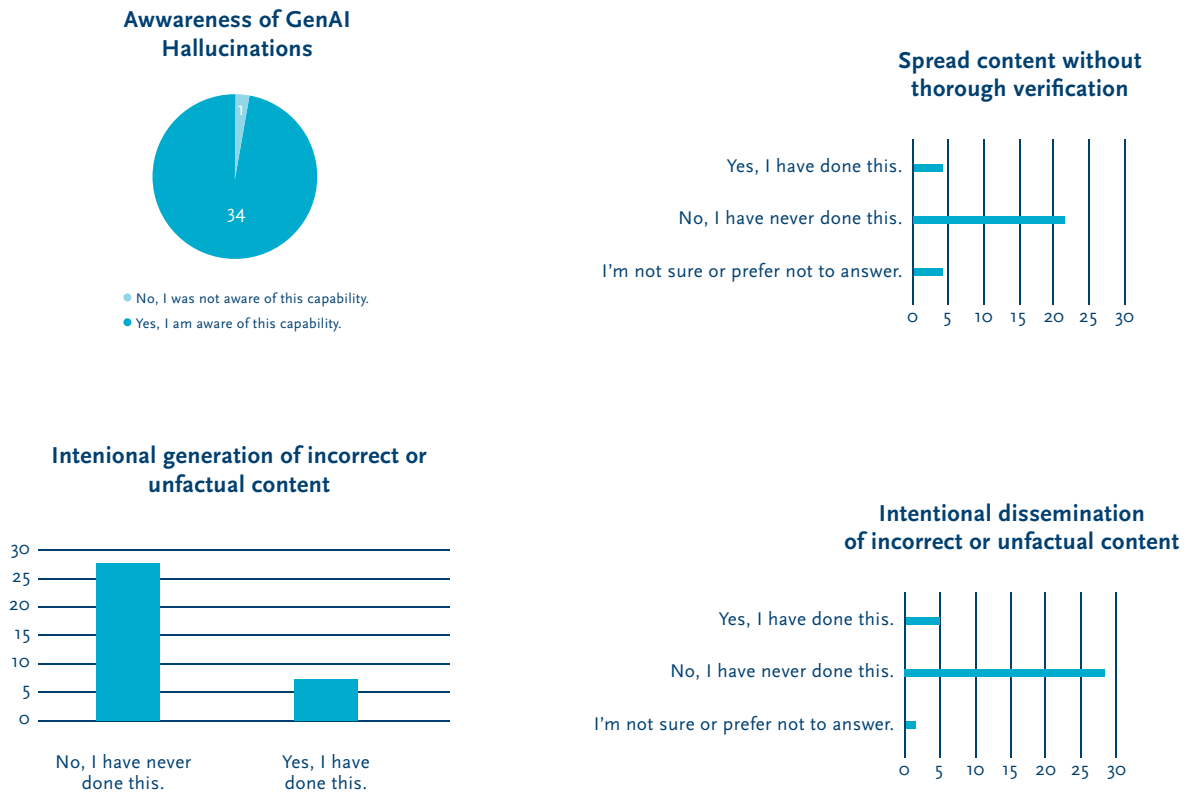

Figure 6. Awareness of AI hallucinations, verification and dissemination of content

Figure 7. Awareness of own Feedback on and training of AI-models

D) *Solutions and expectations:*

88% of the respondents answered that involving the community is important to tackle AI-misalignment issues, 12% did not.

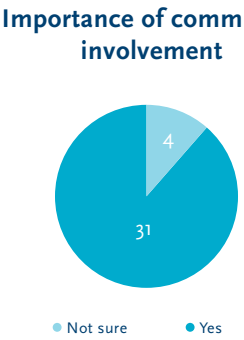

Figure 8. Importance of community involvement

4.3.2 Potential solutions to tackle AI-misalignment

The following table summarizes how the respondents think the community could tackle AI-misalignment.

Effective ways to involve the community in tackling AI-misalignment (Q29)		
Cluster 1	General involvement and Education	"It is always important to include users and populations in research. Of course, the majority will have no idea what The Alignment Problem is, therefore, their contribution must be on a low level. You can simply take opinion polls and include actual end-users in experiments and research. Don't get the public to vote on issues they do not understand." "Community forums and educational workshops." "Focus groups."
Cluster 2	Information sharing and transparency	"When providing answers, it should give away its sources and an estimation of how 'correct' the answer is." "Show informative videos of how the system works in normal terms, maybe an informal video with a touch of humor to help it land with regular folks who have no experience with it." "Good communication and information."
Cluster 3	Platforms and tools	"Dedicated forum per tool with incentives to share and solve cases of misalignment." "Making a forum or App." "For example, Community notes by X."
Cluster 4	Involvement of Diverse groups and Experts	"Appoint scholars and persons of knowledge to tackle this." "We need diverse groups to assess and monitor and determine what is fair and unbiased. It should not be only rich white men who determine what is aligned, fair, or accurate."
Cluster 5	Public awareness campaigns	"Through news entertainment shows (Arjen Lubach, Daily Show, John Oliver)." "Government campaigns." "Consult advertising agencies around this question."

Table 1. Ways of involving the community in tackling AI-misalignment

5. Discussion

5.1 Awareness of Generative AI Hallucinations

The survey results reveal that a vast majority (97%) of respondents are aware of the phenomenon of generative AI hallucinations. This high level of awareness suggests that users are increasingly knowledgeable about the limitations and potential pitfalls of AI-generated content. However, the presence of even a small fraction (3%) of users who are not aware highlights the need for ongoing education and awareness campaigns to ensure all users understand the capabilities and limitations of AI systems.

5.2 Verification and Dissemination of AI-Generated Content

A significant majority (77%) of respondents indicated they have never disseminated AI-generated content without thorough verification. This cautious approach underscores a responsible user base that recognizes the importance of accuracy and reliability in AI outputs. However, the 11% of respondents who have spread unverified content, and another 11% who were unsure or preferred not to answer, point to a critical area where user training and strict guidelines are needed to prevent the spread of misinformation.

5.3 Intentional Generation and Spread of Incorrect Content

Interestingly, 20% of respondents admitted to intentionally generating incorrect or unfactual content. This behavior raises ethical concerns and suggests that some users may exploit AI systems for misinformation. Moreover, 14% of respondents have spread such content, while 5% were unsure or preferred not to answer. These findings highlight the necessity for robust ethical standards and monitoring mechanisms to mitigate the misuse of AI technologies.

5.4 Feedback to Generative AI Models

The data shows a mixed pattern in terms of providing feedback to generative AI models. While 22% of respondents regularly provide feedback and 37% do so occasionally, another 37% do not provide any feedback, and 3% are unsure. This mixed engagement suggests that while some users are proactive in helping improve AI systems, a significant portion remains disengaged. Enhancing user engagement in providing feedback can be beneficial for the continuous improvement of AI models.

5.5 Awareness of Feedback Impact

A majority of respondents (71%) are aware of their impact on the training of generative AI models through feedback, yet 22% remain unaware, and 6% are unsure. This gap in awareness needs to be addressed to ensure users understand how their interactions and feedback can shape AI behavior and performance. Educational initiatives could help bridge this gap and encourage more users to provide valuable feedback.

5.6 Concerns About AI-Generated Content Impact

When considering the impact of their AI-generated content on societal knowledge, respondents expressed varying levels of concern. While 17% were not concerned at all and 40% were not very concerned, 28% were somewhat concerned, and 14% were very concerned. This distribution indicates that while a majority are not overly worried, there is still a significant portion of users who recognize and are concerned about the broader implications of AI-generated content. These concerns are critical for guiding policies and practices around the ethical use and dissemination of AI outputs.

5.7 Community Involvement and Personal Responsibility

An interesting observation from the survey responses related to community involvement in addressing AI misalignment is the lack of mention regarding personal behavior, responsibility, and accountability. While many respondents discussed the need for broader societal and systemic actions, few acknowledged their own role in ensuring the ethical use and improvement of AI systems. This indicates a potential area for further education and advocacy, emphasizing the importance of individual actions and accountability in tackling AI misalignment.

5.8 Limitations

This study has several limitations. First, the number of respondents was relatively small, with only 35 fully completed surveys. As previously mentioned in section 3, this survey is exploratory in nature due to this limited sample size, which restricts the generalizability of the findings. The small number of respondents may not capture the full diversity of user behaviors. Future research with a larger and more varied sample is needed to validate these preliminary insights. Additionally, the demographic distribution of the respondents indicates a potential bias. Most respondents were highly educated and predominantly employed in the education sector. This homogeneity suggests that the survey dissemination through LinkedIn may have targeted a specific subset of professionals, thus not capturing a broader and more diverse audience. Future research should aim to include a wider range of participants from different educational backgrounds and industries to provide a more comprehensive understanding of the issues surrounding generative AI.

Lastly, it is important to note that all of the survey participants come from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) countries. This demographic concentration may influence the study's findings in several ways. Individuals from WEIRD backgrounds often have distinct cultural attitudes toward technology, risk, and accountability, which could shape their interactions with AI systems differently compared to those from non-WEIRD contexts. As a result, the observed patterns of user behavior, including their intent, responsibility, and influence in AI interactions, might not fully represent global perspectives. Consequently, the findings related to AI misalignment may have limited generalizability to diverse socio-cultural settings. Acknowledging this limitation is crucial for understanding the scope of the study and for guiding future research towards more inclusive and representative samples.

6. Conclusions

This research explored the effects of user-centric AI misalignment on societal knowledge, revealing important insights into user interactions with generative AI systems and their broader implications. The study found that while there is a high level of awareness about AI's limitations, such as hallucinations, gaps remain in understanding the impact of user behavior on AI outcomes. Most users are cautious about verifying AI-generated content before dissemination, yet a notable minority has engaged in spreading unverified or incorrect information.

The findings indicate that while many users provide feedback to AI systems, a significant portion remains disengaged or unaware of the impact of their feedback. This highlights the need for more effective communication and user-friendly mechanisms to encourage greater participation in AI improvement processes. Moreover, there are varying levels of concern about the impact of AI-generated content on societal knowledge, with some users recognizing the potential risks and others less concerned.

The study also points out that personal responsibility and accountability in AI use are often overlooked, emphasizing the need for further education and advocacy to foster a sense of individual duty in addressing AI misalignment.

The survey results provide valuable insights into user awareness, behavior, and attitudes towards generative AI. While there is a high level of awareness about AI hallucinations and a generally responsible approach to content verification, there are notable gaps in understanding the impact of feedback and the ethical use of AI. Addressing these gaps through targeted education and robust ethical guidelines will be crucial in ensuring the responsible development and deployment of AI technologies.

The research has limitations, including a small and biased sample size, which may affect the generalizability of the findings. Future studies should aim to include a more diverse participant pool and examine the effects of demographic factors such as age and gender on AI awareness and behavior. Additionally, conducting prompt analysis could provide deeper insights into user interactions with generative AI.

6.1 Recommendations

To address the limitations mentioned in section 5, future research should aim to:

1. Reconduct the analysis with a broader and more diverse audience to improve the generalizability of the findings.
2. Conduct additional analyses to explore the effects of demographic variables such as age and gender on awareness, behavior, and attitudes towards generative AI.
3. Investigate correlations between demographic factors and the various aspects of AI usage and perceptions identified in this study.
4. Implement targeted educational programs to increase user awareness and understanding of their impact on AI systems and the importance of ethical AI usage.
5. Conduct prompt analysis to directly address and understand the source of user behavior in interacting with generative AI systems.

Addressing user-centric AI misalignment requires a comprehensive approach that integrates ethical, technical, and societal considerations. By enhancing educational efforts, developing robust ethical guidelines, and promoting individual accountability, the risks associated with AI misalignment can be mitigated. This research underscores the critical role of user behavior in shaping AI's impact on societal knowledge and highlights the need for ongoing exploration and proactive measures in this dynamic field.

7. Appendix A

7.1 Survey Questions

User Survey Questionnaire: Understanding Perspectives on AI-Generated Information*Introduction:*Thank you for participating in this survey. Your insights are crucial in shaping a new system to check AI-generated information. This survey is about understanding your thoughts on AI misalignment before we design a solution. Please answer the following questions honestly and succinctly.

Section 1: Background Information

1.1. **Gender:**
- Male
- Female
- Other (please specify): _____

1.2. **Age:**
- Under 18
- 18-24
- 25-34
- 35-44
- 45-54
- 55 or older

1.3. **Education:**
- High School
- College
- Graduate School
- Other (please specify): _____

1.4 **Occupation:**
- Healthcare
- ICT
- Education

- Finance
- Manufacturing
- Retail
- Tourism
- Art/Entertainment
- Government/Public Service
- Agriculture
- Engineering
- Media/Communication
- Other

Section 2: Use of Generative AI

2.1. **Level of experience**
- Beginner
- Intermediate
- Advanced

2.2. **Frequency of use:**
- Hourly
- Daily
- Weekly
- Monthly
- Rarely
- Never

2.3. **Context of use:**
- Education
- Work
- Creative Endeavors
- Personal Interest
- Other, please Specify

2.4. How has generative AI impacted your ability to generate new knowledge or ideas?

2.5. What benefits have you experienced from using generative AI?

2.6. Have you encountered any challenges or limitations in using generative AI? If yes, please specify.

Section 3: Understanding AI Misalignment

3.1. **Have you heard about the challenges related to AI misalignment before taking this survey?**
- Yes
- No

3.2. **How well do you understand the concept of AI misalignment?**
- Very well
- Moderately well
- Not very well
- Not at all

Section 4: Identifying Misalignment

4.1. **In your own words, how would you define AI misalignment?**

4.2. **Can you provide an example (if any) where you think AI might generate information that could be misleading or incorrect?**

Section 5: Impact and Concerns5.1. **How concerned are you about the potential impact of AI-generated misinformation on societal knowledge?**

- Very concerned
- Somewhat concerned
- Not very concerned
- Not concerned at all

5.2. **What specific concerns do you have regarding AI-generated information? (Choose all that apply)**
- Spread of false information
- Bias in the information generated
- Lack of transparency in the AI process
- Other (please specify): _____

Section 6: Awareness of own use of generative AI on knowledge generation and dissemination

6.1. **When using generative AI tools for content creation, what are your main goals or objectives?**
- To achieve specific informational or educational goals.
- To express creativity and imagination.
- To support research and data analysis.
- To explore AI technology for personal or educational purposes.
- To engage in artistic or entertainment endeavors.
- To share information that aligns with your perspectives or interests.
- To use content for satire, humor, or parody.
- To experiment with generating different types of content.
- Other, please specify.

6.2. **Are you aware that generative AI models, like GPT-3.5, can hallucinate, meaning produce incorrect and misleading texts or images?**
- Yes, I am aware of this capability.
- No, I was not aware of this capability.
- I'm not sure.

6.3. **Have you ever generated content generated by AI without thoroughly verifying its accuracy or factual correctness?**
- Yes, I have done this.
- No, I have never done this.
- I'm not sure or prefer not to answer.

6.4. **Have you ever disseminated content generated by AI without thoroughly verifying its accuracy or factual correctness?**
- Yes, I have done this.
- No, I have never done this.
- I'm not sure or prefer not to answer.

6.5. **Have you ever intentionally generated content generated by AI that was inaccurate or factually incorrect?**
- Yes, I have done this.
- No, I have never done this.
- I'm not sure or prefer not to answer.

6.6. **Have you ever intentionally disseminated content generated by AI that was inaccurate or factually incorrect?**
- Yes, I have done this.
- No, I have never done this.
- I'm not sure or prefer not to answer.

6.7. **When using generative AI, do you feel you have a clear understanding of its capabilities and how it makes decisions to provide you with your desired output?**
- Yes, I have a comprehensive understanding of generative AI capabilities and decision-making processes.
- No, my understanding of generative AI capabilities and decision-making processes is limited.
- I'm unsure about my level of understanding.

6.8. **Do you provide feedback to generative AI chatbots while you use them, and if so, how often?**
- Yes, I regularly provide feedback after using generative AI chatbots.
- Yes, but I only provide feedback occasionally.
- No, I do not typically provide feedback.
- I have never used generative AI chatbots.
- I'm not sure.

6.9. **Are you aware that generative AI models, like GPT-3, are trained and improved through the feedback provided by users like yourself?**
- Yes, I am aware of this training process.
- No, I was not aware that user feedback contributes to AI model training.
- I'm not sure.

6.10. **Are you aware that the feedback you provide to generative AI models could sometimes be inaccurate or faulty, potentially impacting their training and performance?**
- Yes, I am aware of the potential for providing inaccurate feedback.
- No, I was not aware that my feedback could impact the AI models in this way.
- I'm not sure.

6.11. **When using generative AI, if you encounter sycophantic behavior, which means the AI excessively praises or agrees with you or apologizes profusely after your feedback, do you make an effort to challenge or correct it (provided you gave it the wrong feedback)?**
- Yes, I actively try to challenge or correct sycophantic behavior when I encounter it.
- No, I typically don't take action when AI systems display sycophantic behavior.
- I rarely encounter sycophantic behavior from AI systems.
- I'm not sure or prefer not to answer.

6.12. **How concerned are you about the potential impact of your own AI-generated output on societal knowledge?**

- Very concerned
- Somewhat concerned
- Not very concerned
- Not concerned at all

Section 7: Solutions and Expectations

7.1. **Do you believe it is important to involve the community in addressing AI misalignment?**

- Yes
- No
- Not sure

7.2. **What do you think would be effective ways to involve the community in identifying and addressing AI misalignment?**

Thank you for your valuable input. Your responses will guide the development of a solution to address AI misalignment before launching any platform.

Appendix B

Answers to additional open questions

Impacts from AI-usage (Q8)

The responses about the impact of generative AI on the ability to create new knowledge have been clustered into three groups. Here are the clusters along with representative responses for each:

Impact of generative AI on the ability to create new knowledge Q8		
Cluster 1	General positive impact and efficiency	"Positively. It allows me to start things from a much stronger base." "It is a main resource that can give you informed information about best events, can look online for you within seconds for relevant information, and can suggest alternative reasons for any phenomena. It is for sure the most capable assistant I have ever had, a powerful collaborator." "It has helped identify valuable information about industries I am not an expert in. As a consultant, I touch upon multiple industries; it is difficult to master all. ChatGPT helps identify specific and valuable information that helps me solve questions or even lead me to other questions initially out of scope."
Cluster 2	Idea generation and knowledge expansion	"It helped me shape my ideas." "Great for quick ideas." "It helps generate new knowledge and ideas." "It accelerates knowledge expansion into yet unknown fields."
Cluster 3	Process improvement and speed	"AI smoothens the process, where I do not need to google for an hour anymore to find a solution to a problem." "For some of my tasks, I can do what I was doing a lot faster." "A lot! I learned a lot but also a lot of ideas have turned into reality because of AI."

Benefits from AI-usage (Q9)

The responses about the benefits from using generative AI have been clustered into four groups. Here are the clusters along with representative responses for each:

Benefits from using GenAI Q9		
Cluster 1	Structured Information and creativity	"Understand areas of importance for my work in a matter of seconds, come up with structured yet creative ways of presenting information." "Very productive, mainstream texts are very useful in the public domain where you need everyone to understand your words." "Quick search for information."
Cluster 2	Time saving and efficiency	"Saving a lot of time, looking up things myself, and forming much more informed opinions about topics where without this tool, I would not have time to follow up on myself. As a result, I have much more intelligent opinions about a variety of topics." "Less time on a task." "Saving a bit of time."
Cluster 3	Idea generation and Task assistance	"Easier to research ideas and finish things." "Help me shape my ideas." "Getting fast feedback without having to search online and read the search results."
Cluster 4	Quality improvement and Learning	“Gives answers." "It helps me write better pieces. It helps me improve overall quality of text-based engineering products like functional requirements specifications. I can ask stupid questions about everything and still get answers. It helps me to faster understand new topics." "Time saving, helps to generate an idea, it gives a broader view."

Challenges and limitations in AI-use (Q10):

The responses about the challenges and limitations in generative AI use have been clustered into four groups. Here are the clusters along with representative responses for each:

Challenges and Limitations in generative AI use Q10		
Cluster 1	Technical limitations and reliability	"With more specific or complex questions that would be even hard to find by browsing the web, GenAI tends to give superficial answers." "The output isn't totally reliable. There has to be always a human check. ChatGPT is 'lazy': I often have to split a specific task into small chunks to get a usable outcome." "Capacity problems with data analysis (ChatGPT 4), and sometimes missing specific information, for example, about specific topics from the Netherlands."
Cluster 2	Prompting and user interaction	"Yes, as a policy maker I have to be careful with what to share without spoiling information." "Prompting is hard; questions need to be specified really carefully in order to get the right answer." "Specific formulation question for answer."
Cluster 3	General user experience and Frustration	"The systems have been getting slowly enshittified. They used to be much more powerful I think." "It is still weak in generating graphs. It also sometimes gets stuck on ethical concerns and man-made limitations. That can become frustrating when you are trying to get some work done. Sometimes it does not understand exactly what you mean, and then you spend time crafting a good prompt, which takes probably just as much time as doing the work yourself." "Sometimes it produces nonsense."
Cluster 4	Ethical concerns and Data privacy	"Accuracy of data is unclear, ethics of used data unclear, makes the data results dubious." "Data privacy, security risks, and the range of use of sources that have given no consent to be used for AI training pose significant threats." "The specificity in which you have to communicate when you are using it almost makes it unnecessary to use."

8. References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. https://arxiv.org/abs/1606.06565

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. (2021). A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861

Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the Boundaries of Reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Curēus. https://doi.org/10.7759/cureus.37432

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 610–623.

Bergman, P. (2024). Aligning AI systems with human values: Challenges and strategies. Journal of AI Research, 67(1), 123-145. https://doi.org/10.1016/j.jair.2024.01.005

Bignami, A., D'Angelo, F., Guerrini, F., & Rossi, A. (2023). Dataset shift and its impact on the performance of machine learning systems. https://doi.org/10.1016/j.artint.2023.103563

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Bovens M (2007) Analysing and assessing accountability: a conceptual framework. Eur Law J 13(4):447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.09009

Deng, J., & Lin, Y. (2023). The Benefits and Challenges of ChatGPT: An Overview. Frontiers in Computing and Intelligent Systems, 2(2), 81–83. https://doi.org/10.54097/fcis.v2i2.4465

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. Synthese, 202(5). https://doi.org/10.1007/s11229-023-04367-0

Durante M, Floridi L (2022) A legal principles-based framework for AI liability regulation. In: Mökander J, Ziosi M (eds) The 2021 yearbook of the digital ethics lab. Springer International Publishing, pp 93–112. https://doi.org/10.1007/978-3-031-09846-8_7

Floridi, L. (2016). Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. Philosophical Transactions- Royal Society. Mathematical, Physical and Engineering Sciences, 374(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and machines, 30(3):411–437.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. arXiv preprint arXiv:2009.11462.

Lee, J. N., Xie, A., Pacchiano, A., Chandak, Y., Finn, C., Nachum, O., & Brunskill, E. (2023). Supervised pretraining can learn In-Context reinforcement learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2306.14892

Lindberg SI (2013) Mapping accountability: core concept and subtypes. Int Rev Adm Sci 79(2):202–226. https://doi.org/10.1177/0020852313477761

Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., and Irving, G. (2021). Alignment of language agents. arXiv preprint arXiv:2103.14659.

Kleinman, B. Z. (2024, February 28). *Why Google's "woke" AI problem won't be an easy fix*. https://www.bbc.com/news/technology-68412620

Koohang, A., and Weiss, E. 2003. Misinformation: toward creating a prevention framework. Information Science.

Kreps, S., McCain, R. M., & Brundage, M. (2020). All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. Journal of Experimental Political Science, 9(1), 104–117. https://doi.org/10.1017/xps.2020.37

LaCroix, T. and Bengio, Y. (2020). Learning from learning machines: optimisation, rules, and social norms. https://doi.org/10.48550/arxiv.2001.00006

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. (2018). Scalable agent alignment via reward modeling: a research direction. arXiv preprint arXiv:1811.07871.

Li, Z., Yang, Z., & Wang, M. (2023). Reinforcement Learning with Human Feedback: Learning Dynamic Choices via Pessimism. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.18438

Loi, M., & Spielkamp, M. (2021). Towards Accountability in the Use of Artificial Intelligence for Public Administrations. Association for Computing Machinery New York, NY, United States. https://doi.org/10.1145/3461702.3462631

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022, June 1). *Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance*. arXiv.org. https://arxiv.org/abs/2206.00335

Mintz, A. (2002). Web of Deception: Misinformation on the Internet. New Jersey: Information Today, Inc.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1312.5602

Montgomery, R. (2024). From biological constraints to unbounded artificial evolution: exploring the implications of AI's accelerated advancement. WNSC, 1(3), 1-14. https://doi.org/10.62162/1062011

Mulgan R (2000) 'Accountability': an ever-expanding concept? Public Admin 78(3):555–573. https://doi.org/10.1111/1467-9299.00218

Mulgan R (2003) Issues of accountability. In: Mulgan R (ed) Holding power to account: accountability in modern democracies. Palgrave Macmillan, pp 1–35. https://doi.org/10.1057/9781403943835_1

Murphy, K., Ruggiero, E., Upshur, R., Willison, D., Malhotra, N., Cai, J., ... & Gibson, J. (2021). Artificial intelligence for good health: a scoping review of the ethics literature. BMC Medical Ethics, 22(1). https://doi.org/10.1186/s12910-021-00577-8

Novelli, C., Taddeo, M. & Floridi, L. Accountability in artificial intelligence: what it is and how it works. *AI & Soc* (2023). https://doi.org/10.1007/s00146-023-01635-y

Parentoni, L. (2024). What should we reasonably expect from artificial intelligence? Russian Journal of Economics and Law, 18(1), 217-245. https://doi.org/10.21202/2782-2923.2024.1.217-245

Piper, P. (2000). Better read that again: Web hoaxes and misinformation. Searcher, 8(8). Retrieved February 02, 2003 from the Word Wide Web http://www.infotoday.com/searcher/sep00/piper.htm

Puri, A., & Keymolen, E. (2023). Of ChatGPT and trustworthy AI. *Journal of Human-Technology Relations*, *1*. https://doi.org/10.59490/jhtr.2023.1.7028

Ouyang, L., Wu, J., Xu, J., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., John, S., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. J. (2022). Training language models to follow instructions with human feedback. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2203.02155

Safron, M., Huang, Y., & Kumar, V. (2023). Inductive biases and AI failures: A study on training algorithms. https://doi.org/10.1109/ICRA.2023.9450987

Saghafian, S. (2023). Effective Generative AI: the Human-Algorithm Centaur. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4587250

Santoni de Sio, F.; Van den Hoven, J. 2018. Meaningful Human Control over Autonomous Systems: A Philosophical Account. Frontiers in Robotics and AI 5. https://doi.org/10.3389/frobt.2018.00015

Shahbaz, Funk, and Vesteinsson, "The Repressive Power of Artificial Intelligence," in Shahbaz, Funk, Vesteinsson, Brody, Baker, Grothe, Barak, Masinsin, Modi, Sutterlin eds. Freedom on the Net 2023, Freedom House, 2023, freedomonthenet.org.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Shanahan, M. (2024). Talking about Large Language Models. *Communications of the ACM*, *67*(2), 68–79. https://doi.org/10.1145/3624724

Shankar, K., Mohan, P., & Ramesh, S. (2007). *Sycophant: A context-aware adaptive AI system*. Journal of Computer Science, 22(4), 345-357. https://doi.org/10.1016/j.jcs.2007.01.001

Shrivastava, R. (2023, December 31). How ChatGPT and billions in investment helped AI go mainstream in 2023. *Forbes*. https://www.forbes.com/sites/rashishrivastava/2023/12/27/how-chatgpt-and-billions-in-investment-helped-ai-go-mainstream-in-2023/

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: training on generated data makes models forget. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.17493

Tamkin, A., Brundage, M., Clark, J., and Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. arXiv preprint arXiv:2102.02503.

Thynne I, Goldring J (1987) Accountability and control: government officials and the exercise of power. Law Book Company

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M.,... Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 575(7782), 350–354. https://doi.org/10.1038/s41586-019-1724-z

Wang, X., & Chen, X. (2024, May 7). *Towards Human-AI Mutual Learning: A New Research Paradigm*. arXiv.org. https://arxiv.org/abs/2405.04687

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359. Welbl, J., Glaese, A., Uesato, J., Dathath

Wieringa, M. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20. Association for Computing Machinery, Barcelona, Spain, pp. 1–18. https://doi.org/10.1145/3351095.337283

Zhou, L., Wu, Y., & Zhang, H. (2022). Fairness in AI: Addressing algorithmic discrimination. ACM Computing Surveys, 54(8), 1-38. https://doi.org/10.1145/3456789