Electronic Communications of the EASST
Volume 84 Year 2025

12th International Symposium on Leveraging Applications of Formal Methods, Verification and Validation / 2nd AISoLA - Doctoral Symposium, 2024

Edited by: Sven Jörges, Salim Saay, Steven Smyth

Unintended and Illegitimate Consequences of LLMs and Their Impact on Society

Ulla Coester, Dominik Adler, Christian Böttger, Norbert Pohlmann

DOI: 10.14279/eceasst.v84.2676

Electronic Communications of the EASST (https://eceasst.org).
Published by **Berlin Universities Publishing**
(https://www.berlin-universities-publishing.de/)

Unintended and Illegitimate Consequences of LLMs and Their Impact on Society

Ulla Coester[1] ORCID, **Dominik Adler**[1], **Christian Böttger**[1] ORCID, **Norbert Pohlmann**[1] ORCID

Westphalian University of Applied Science, Institute for Internet Security[1]

Abstract: In our paper, we explore consequences of Large Language Models (LLMs) from the perspective that they can cause illegitimate harm if they are not taken into account by AI providers, and what requirements result from this in terms of what needs to be done. In the first part of the article, we examine the potential for harm that can be caused by LLMs and then use the 'do no harm'-principle to illustrate why AI providers are theoretically obligated to employ all available measures to prevent illegitimate harm. The subsequent section details the development of an AI Restriction Framework, which aims to enhance visibility of potential illegitimate harm, thereby serving as a base for both AI providers and users to take action. The overarching objective of our research is to establish a foundation for a shared understanding of the potential (illegitimate) harms that may arise from LLMs, by facilitating a focal point for a more informed societal discourse on their utilization.

Keywords: LLM, ChatGPT, Ethics, AI Act, harm, 'do no harm'-principle

1 Introduction

A significant proportion of individuals use technology with the belief that it will have a positive benefit, whether in the form of enhanced efficiency or the facilitation of their professional tasks. This notion can be compared to that of the "Zauberlehrling" sorcerer's apprentice - who sought to simplify his tasks by utilizing the spell he had learned from his master, but unfortunately, he lacked the knowledge of the counter spell, leading to a situation that spiraled out of control – as depicted in Goethe's ballad. However, can it be asserted with absolute certainty that, in contrast to the sorcerer's apprentice, everyone today has a precise understanding of what they are doing? After all, it seems to be questionable whether individuals – developers or users alike – are aware of the measures to be taken to ensure that Artificial Intelligence (AI) does not become uncontrollable. Moreover, can this status even be defined in advance? Or, most importantly, according to which criteria can such a decision be made [Gab18, pp. 208–209]? But not only these rather complex and future-oriented questions need to be answered; there is also a necessity for a thorough evaluation of the value alignment of individual actors, particularly international technology companies, from a more evident perspective. Namely the fundamental question, whether their motives and actions today are actually geared towards providing an AI that prioritizes human well-being and needs – which would also entail the commitment to take every possible measure to avoid causing (illegitimate) harm to people. Since the value alignment of the most technology companies remains unclear respectively is – in some cases – only known in a rudimentary way or currently being questioned, a fundamental requirement can be derived from this: To empower individuals within a society to lead lives that are largely consonant with their moral concepts, for example regarding privacy or the future of work, it is essential to create a shared understanding of values (i.e. the defining principles of the social structure), conditions, circumstances and expectations concerning technology and the management of potential implications.

To reach a consensus on this matter, it is pertinent to establish a foundation for informed decision-making. This necessitates a technological approach to the domain of AI security, among other areas, in addition to a comprehensive understanding of the potential harms that may arise from the utilization of AI, particularly the illegitimate ones. Harm may occur if a chatbot fails by default to recognize a potentially dangerous situation for (vulnerable) users, such as children, and provides them with inappropriate instructions or guidance – this includes among other things overlooking signs that indicate self-destructive behavior or failing to refer users to professional assistance when required. In this context, the framework presented in this paper could serve as a valuable instrument in this regard, as it evaluates the priorities for AI providers in addressing identified issues, while concurrently illustrating to users the potential consequences and methods to avert them.

1.1 A much-needed critical discourse about ChatGPT

Societies are now characterized by technologies, preponderant data-driven, that have a significant impact on individuals' living. The implications of these technologies, both potential and actual, as exemplified by social media, have been subjects of ongoing public discourse. This is partly due to the predominance of large technology companies in the market, enabling them to influence societal development. But despite the extensive debate [Lan25] – also from the regulatory side [Die25, Deu24] – it remains unclear in principle what unequivocal moral expectations and resulting demands should be placed on these companies. As long as this is not defined by society - and people still continue to utilize the services in the respective conditions - such companies perceive their actions as legitimate, as they are seemingly acting in the best interests of the users of their services. Based on this, it could be argued that they are absolves from facing up to their responsibility to society. A similar phenomenon has also been observed so far in the context of AI, because people have already used AI in various forms – for example, with Alexa or Siri – without necessarily giving it a second thought. However, with the development of innovative Large Language Model (LLM) technology, AI has become the focus of common interest among the general public [SWW+23], because all of a sudden many useful possibilities have emerged, such as having tailor-made texts created or being able to generate codes based on specific inputs. This development has emphasized the imperative for a more profound reflection on AI in general and LLMs in particular – the latter mainly in view of two aspects: firstly, its ubiquitous and comprehensive use, and secondly, its current technical shortcomings.

Notwithstanding the fact that now, from the perspective of a considerable number of users, LLMs represent an indispensable infrastructure for increasing the efficiency necessary to generate competitive advantages in a global market, a debate must be held – cognizant of the fact that a dilemma already exists in practice, even if this applies less to individuals than to companies or society as a whole from a global perspective. But this urgently required debate must prioritize the potential risks that could materialize in the short and medium term, stemming from the utilization of the technology, as opposed to those risks that lie in the distant future [Cen, Bos02, TD20, Bos14]. Because current threats like mentioned in the cited papers, for example that “AI is not aligned with human values” [TD20] or “AI aligned with a malevolent human organization” [TD20] have the potential to induce far-reaching consequences for society; therefore, it seems advisable to discuss possible negative consequences, disregarding the differences of opinion about the expected advances of this technology on the road to singularity.

The importance is due to several criteria, including the fact that Sam Altman never stops pointing out that AI is dangerous, because in his opinion “there’s a real threat that the technology OpenAl is trying to build will cause human extinction, but the only way to deal with this is to keep building it and see what happens[.]” [Tyr25].

As early as 2019, the approach in question had become apparent when Sam Altman and other relevant individuals at OpenAI recognized that “its new AI model, called GPT2 is so good and the risk of malicious use so high that it is breaking from its normal practice of releasing the full research to the public” [Her19] or that “the company said that governments should consider expanding or commencing initiatives to more systematically monitor the societal impact and diffusion of AI technologies, and to measure the progression in the capabilities of such systems” [Whi19]. A review of articles published in 2019, prior to the release of ChatGPT, reveals that the responsible parties at OpenAI were aware of the possibility of risk associated with this novel technology. This misgiving was shared by external experts who interacted with ChatGPT. Nevertheless, despite the obvious risks being recognized and the associated potential for harm, ChatGPT was introduced in 2022. The question of the justifiability of this course of action, which was not readily comprehensible to all – and this remains the case today – demonstrates the necessity to establish a framework in order to carry out a regulated assessment of potential harm and to be able to derive appropriate actions or instructions for action from it.

We therefore argue that, in accordance with the rationale delineated in Section 2.2, there exists sufficient evidence to substantiate the proposition that the prevailing harms, which are both demonstrably occurring and have the potential to occur, ought to be subjected to closer examination. Based on the results that can be generated from this, a normative foundation for understanding illegitimate harm can be developed.

1.2 Why must unintended consequences of new technologies be considered?

In order to evaluate harm that could possibly be caused by new technologies, it is essential to understand that every action has consequences [Mer36]. Consequences of an action may be predictable, yet they can also be unintended. The need of a nuanced understanding of the unintended consequences arising from accelerated technological progress is crucial for the following rationale: Progress can lead to a lock-in situation “through a shift in power and values, through habituation to and dependence on innovations, and through destruction and thus the closure of entire options for people and regions”, so that the effects - despite the optimism about progress - must be taken into account (own translation) [Deu23]. Generally spoken, unintended consequences can have both negative and positive connotations. An example of a positive one is Adam Smith’s theorem of the “invisible hand”, which shows that in a free market, the self-interest of each market participant unintentionally leads to the “wealth of the nation” [Smi76]. The introduction of the General Data Protection Regulation (GDPR) [Eur16], on the other hand, is an example of an unintended consequence with negative effects. Originally intended to protect user privacy, the GDPR has had the opposite effect: Due to the mandatory consent of users, companies are legally entitled to use all data for their own purposes.

In the given context, however, it is important to note that the argumentation does not actually refer to the consequences of technology because, strictly speaking, technology does not cause any consequences (own translation) [Deu23]. Instead, it refers to "the consequences that may arise from the range of possibilities for human decisions and actions in connection with technology" (own translation) [Deu23]. This leads to the conclusion that, in principle, it is impossible to predict how individuals will interact with innovative technologies and the new circumstances that will emerge as a result. A deduction which can also be derived from an attempt at explanation in theoretical sociology formulated by Karl Popper with regard to his observation that "[...] one of the most striking phenomena in social life [...] is that [...] the result that is intended by the parties involved never occurs exactly" (own translation) [Pop65, p. 124 f.]. The emphasis on the necessity of addressing unintended consequences stems from the fact that "they are usually undesirable" and "with technical progress and the constant globalization of the world, the spatial and temporal scope of the consequences has also increased significantly [...][.]" (own translation) [Deu23]. Understanding unintended consequences is necessary to assess which measures have to be taken in order to prevent harm. When considering the type of measures to be taken, it is important to bear in mind that, with regard to unintended consequences, it is necessary to take into account both the attribution of causality and the question of whether it is possible to determine the actual purpose of a technology or application, or whether this is a matter of retrospective rationalization [Deu23].

Following this line of reasoning, it seems essential to subject LLMs – and ChatGPT in particular, given their dominant market position [Alb23, Reu25] – to more thorough analysis. This conclusion is based on the following circumstantial evidence: Firstly, the assumption that an action never has only one effect – let alone the desired one – and secondly, the fact that LLMs are, by definition, intended for general-purpose use. The differentiation between unintended consequences and illegitimate harm is relevant for the following reason: In the context of illegitimate harm, society is usually aggrieved knowingly, as is elaborated in Section 1.4. These findings are necessary in order to understand the dedicated conception of the 'do no (illegitimate) harm'-framework.

1.3 New technologies come with the risk of harm

At the time when ChatGPT has been introduced, it was unfeasible to draw on past experience in order to anticipate and assess (all) possible consequences that might result from its use, as the technology - on which LLMs are generally based - is a new innovative approach. This circumstance, together with the fact that technological development in general is highly volatile, also creates difficulties for legislation. Although it was (still) possible to take LLMs into account in the context of the AI Act [Eur24, Art. 51(1)(b)], it was clear at the time of writing that it would be virtually impossible to adequately regulate them. However, it is quite common for legislators to refrain from providing precise definitions in such highly volatile areas. The reasoning behind this is that a more specific definition of these requirements would run the risk of constantly lagging behind technological developments, so the legislator deliberately creates a more vague framework. The AI Act [Eur24] is therefore not designed to prevent or mitigate all damage in detail - it serves as a kind of foundation in which relevant normative aspects are covered, for example, by prohibiting biometric identification and, depending on the criticality, requiring providers and users to take measures to prevent greater damage.

For this reason, it was and still is necessary to evaluate ChatGPT over time, which has led to a number of scientists, among others, studying LLMs in general and ChatGPT in particular over the past two years. As with any innovation, there is a tension between positive and negative outcomes with ChatGPT. On the one hand, advanced technology opens up new fields of activities – on the other hand, it results in risks. Based on previous findings it can be reasonably anticipated that ChatGPT will engender new potential for efficiency gains in the medical field [Ray23]. Conversely, there are distinct risks and challenges intrinsic to the provision of LLMs: A salient issue that arises in the process of utilizing ChatGPT to create texts is the potential for manipulation of users [JBB+23]. Problems related to manipulation [Alb23, p. 57] (through language) and misinformation in the context of ChatGPT are often the subject of research [BBC23, KB23].

Considering the above-mentioned consequences as well of the utilization for LLMs as of further application scenarios, which are being gradually investigated, it is worth reflecting on the insights of a liberalist French economist, even if they did not originally refer to technological progress: "Two very different masters teach him [humanity] this lesson: experience and foresight. Experience teaches effectively but brutally. It teaches us all the effects of an action by making them tangible, and we cannot avoid learning at some point. [...] I would prefer to replace this harsh teacher as much as possible with a gentler one: foresight[.]" [Bas95]. Following Basiats' line of thought, it seems appropriate to ask whether, in the context of LLMs, obvious consequences have to be learned 'brutally through experience'. Or whether it would not be better to foresee them in advance [FS20] so that – in accordance with the principle of universalizability [Suc15, p. 163] – a process for dealing with them appropriately in the interest of society could have been designed up front. Nevertheless, the fact that the potential effects of ChatGPT were apparently not thoroughly investigated before its release, and the fact that these effects are now becoming visible and leading to more or less harm to individuals or society, means that it is necessary to develop a shared understanding regarding its use. This appears necessary under the following aspect – which is derived from a thought that originally underlies the Ethical Foresight Analysis (EFA) – "that the uses and effects of a technological artefact on society are not determined solely by the artefact's design but also by different 'relevant social groups' who co-opt that technology for specific societal needs[.]" [FS20]. However, when balancing the question of how this new – and, due to its innovative nature, disruptive – technology of LLMs can be used, two aspects must be given proper consideration. Firstly, the decision cannot be left to individual groups alone – and certainly not to those who would primarily benefit from it – because "[P]recepts for living together are not going to be handed down from on high. Men must use their own intelligence in imposing order on chaos, intelligence not in scientific problem-solving but in the more difficult sense of finding and maintaining agreement among themselves[.]" [Buc00]. This is particularly relevant in view of the fact that today's actions, which are or appear attractive in the short term, "not only influence certain concrete consequences of action that may be the goals of our actions; we also shape our **future** with them" (own translation, bold in original) [Suc15, p. 132]. Conversely, unintended consequences have the potential to engender social detriments, impacting individual members or groups within society. These issues must be recognized and resolved to ensure successful social coexistence.

But how is harm defined in terms of AI? According to the OECD, harm can be categorized into actual (AI incidents, serious AI incidents, and AI disasters) and potential (AI hazards and serious AI hazards) harm [OEC24].

1.4 The necessity of the 'Do no (illegitimate) harm'-Principle

Given the statements about the potential harm that can result from the use of LLMs, particularly in connection with ChatGPT, there is an urgent need to develop a shared understanding of how this can be addressed in line with the 'do no harm'-principle – whereby we focus on the aspect of 'do no illegitimate harm' in our research. "The principle to 'do no harm' [- on which the Hippocratic Oath is also based -] is sometimes considered an element of beneficence, and it is described as the first rule of medicine. Patients have a fundamental right not to be harmed and, furthermore, to have the potential risk of harm minimized in the course of treatment. But meanwhile the concept of harm extends beyond physical injury and includes harm, for instance, to dignity, respect in the community, and self-esteem." [Dic12]. Consequently, the 'do no harm'-principle can be regarded as one of the essential ethical norms and thus serve as a fundamental concept for promoting a shared understanding in an environment that is becoming increasingly complex as a result of digitalization. Based on the following reflection – "[P]eople will, and can have different ideas about a good life, about the world, and more and still cooperate successfully. However, some shared ideas, beliefs, perceptions, and interpretations of the game and its rules are strictly necessary in order to achieve cooperation successfully. More specifically, social cooperation needs a compatible set of beliefs from which mutually consistent behavioral expectations can be derived[.]" [Suc19].

Hypothetical, along with the 'do no harm'-principle, which is accepted as a normative principle and seen as a guideline for human behavior, it appears also as well unreasonable as unjustifiable to develop an AI technology that (intentionally) causes harm. This notion seems furthermore inconsistent in view of the fact that 'the fundamental purpose of AI is to serve and improve human well-being and it is imperative that it is both "trustworthy" and "human-centered", i.e. developed with the human being at its center' (own translation) [GH19] - a principle that is enshrined, inter alia, in the AI Act [Eur24, Art. 1] and in other publications of international organizations such as the OECD. However, as reality shows, harm caused by algorithmic systems – specifically, negative experiences resulting from the use and operation of a system in the world – cannot be avoided due to the interplay of technical system components and social power dynamics [SRH+23].

As a result, efforts are being made at multiple levels worldwide to raise awareness and identify methods for assessing the potential for damage from AI as well as to systematize approaches on how to deal with it. Among others, the relevant international institutions are paying attention to this phenomenon – including the OECD [OEC24] or the UN [Uni24] and UNESCO [Uni21]. The latter explicitly states in its recommendations that "[...] no human being should be harmed or subordinated, whether physically, economically, socially, politically, culturally or mentally during any phase of the life cycle of AI systems [...]" [Uni21, III. Values and principles].

But on what basis is it possible in the context of AI to develop a shared understanding of what (illegitimate) harm is or can be? According to the OECD, a necessary condition for avoiding harm on individuals is that “AI systems need to be trustworthy and reliable to avoid negative effects on people, organizations and the environment [.]” [OEC24]. An important prerequisite for this, as emphasized by the OECD, is that “[..] AI actors need to use the same terms to talk about the problems and failures of AI systems so that we can learn at an international level and prevent repeats[.]” [OEC24]. To create this common understanding, a general definition of current and potential harm is provided [OEC24, p. 8] along with a more in-depth treatment of the dimension of harm [OEC24, Annex A, pp. 16 ff]. Nonetheless, it is important to note that this taxonomy is not the sole authority for determining and mitigating harm caused by AI, since it is possible to carry out categorization and classification under a wide range of aspects. [SSG+24].

Yet, as previously expounded in Section 1.4 “it would be impossible to take this no harm expectation as an unconditional norm that has always to be fulfilled” [Suc19, p. 4 f.]. However, a fundamental distinction must be made here: On the one hand, there is harm that occurs as a result of use, but which is (or could be) accepted by society because the benefits are considered to outweigh the harm caused. This notion is already reflected in the original conception of the ‘do no harm’-principle, which acknowledges that harm is occasionally unavoidable. Medicine offers here a good example, when patients are administered anesthetics that may have deleterious side effects. Nevertheless, a prerequisite for the treatment is the provision of information regarding the probability of benefit and known risks, as well as other pertinent issues. [Dic12].

In contrast to this, illegitimate harms are not acceptable. This is because in such cases – when legal and moral requirements that are considered socially accepted are violated – people/ companies/society are usually harmed knowingly and without legitimate reason. In this context, however, it is crucial to draw a distinction between unintended side effects or secondary consequences, as already delineated in Section 1.2, and illegitimate harm. The former can be seen as an outcome of a purposeful action that is not intended or foreseen [Mer36]. Thoroughly researching these in order to predict consequences as far as possible could be uneconomical according to Merton, since there would then be no time left for other (business) ventures [Mer36].

The latter are considered undesirable by society. In principle, illegitimate harm also occurs as a secondary consequence. The following factors have been identified as the primary “three main causes:

- The person causing the harm is not aware or does not know that they are causing the harm (lack of knowledge or insight).
- The damaging person is aware of it, but doesn’t care, i.e. they don’t attach enough relevance to the damage to prevent it (lack of will).
- The damaging person is aware of it and actually wants to avoid it, but in the situation, they lack the skills, means or opportunities to do so (lack of ability)” [Suc25]

The illegitimate harms, mentioned in the previous paragraph, are – as we will explain in Section 2.2 – to a certain extent identifiable and could therefore be minimized or mitigated either by the AI provider or the user.

2 Method

To illustrate the need for a framework to prevent illegitimate harm, we describe as examples two hazard vectors – indirect prompt injection and cognitive offloading – while highlighting their potential negative societal impacts. This serves as a motivation for developing an AI Restriction Pipeline. The necessity of this pipeline is demonstrated by the proof of concept presented in Section 3.

2.1 Hazard Vectors

In principle, there is no dedicated target group in the context of illegitimate harms, as it affects both companies and their employees, as well as any individual who uses ChatGPT, regardless of the purpose for which it is used.

Indirect Prompt Injection Popular LLM-based chatbots are general-purpose LLMs; that is, they follow the instructions they are given. This provides broad task flexibility but also paves the way for misuse. Moreover, advice may be of varying quality based on the user's prompt engineering skill level and the availability and quality of the data for a respective domain.

With an indirect prompt injection [GAM+23], as shown in Figure 1, an attacker can influence the output generated by an LLM. The attacker modifies an existing website or document or creates a new one. When a user submits a query to an LLM for which retrieving additional sources would be beneficial, the model conducts a web search based on the input prompt. The retrieved sources (among which may be those of the attacker) are then incorporated into the response generation process.

The output of an LLM depends on its training data and the data provided during inference. One can provide different types of data to an LLM during inference. A system prompt defines how an LLM should behave, such as responding in a friendly tone or refusing to answer questions that could lead to harmful or dangerous advice. Typically, the system prompt is not visible to the user.

A user prompt, on the other hand, is written by the end-user interacting with the LLM, often through a chatbot interface. Users can freely instruct the model by asking questions, describing tasks, and specifying context and scope.

Retrieval-augmented generation [LPP+20] enhances LLMs by incorporating information from external sources that are not part of their original training data. This enables LLMs to answer questions requiring knowledge beyond their training cut-off date and to provide users with references to the sources used.

In addition to hallucinations [AWS+18, LHE22], where LLMs generate potentially plausible yet incorrect responses due to insufficient context or limited training data in a specific domain, malicious actors can actively manipulate an LLM's inference process. They may exploit LLMs by crafting user prompts to generate, for instance, spam emails [LCLW24] or other morally questionable things. Additionally, adversaries can embed instructions or misinformation in online sources, influencing the information retrieved and reproduced by an LLM [GAM+23]. Boucher et al. [BPS+23] demonstrated that websites' content can be designed imperceptibly to the eye, while being processed by LLMs, manipulating the results of an LLM-based chat search. Furthermore, jailbreak prompts can override an LLM's system prompt, further altering its behavior and bypassing built-in safeguards [SCB+24].

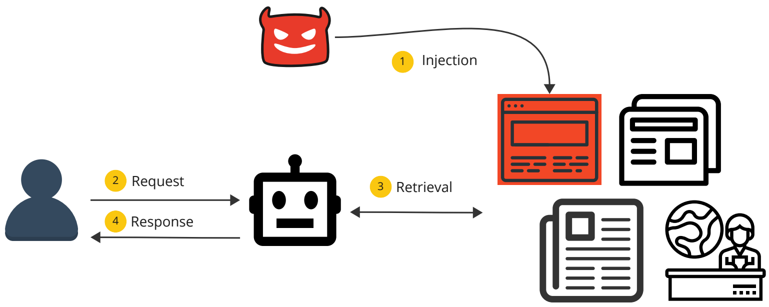

Figure 1: Indirect Prompt Injection

Cognitive Offloading In the digital age and an increasingly complex world, humans must process ever-growing amounts of information for problem-solving and decision-making. However, cognitive capacity is limited, and digital tools can help mitigate this challenge by enabling cognitive offloading.

Search engines enable users to efficiently locate and retrieve information without memorizing every detail. Sparrow et al. [SLW11] raised concerns that excessive reliance on such tools could lead to a decline in cognitive abilities, such as memory retention, as individuals focus more on remembering where to find information rather than the information itself.

With the advent of LLMs, users can access information even more precisely and conveniently than before. However, Gerlich [Ger25] has demonstrated that frequent use of AI technologies may reduce individuals' ability to evaluate information critically. This is particularly alarming when considering the hazard vector of indirect prompt injection described above.

There are reports of students using LLMs to generate homework solutions, effectively engaging in academic dishonesty, disadvantaging students who complete their work independently, and undermining the educational purpose of homework [AI 22].

2.2 Building an AI Restriction Framework

We propose an AI Restriction Pipeline that is designed to aid in identifying and alleviating (potential) illegitimate harm before and after deployment, which is based on the following concerns.

In principle, it is now almost impossible not to recognize one's own responsibility when dealing with AI, if only for the reason that "we repeatedly contribute with our actions to creating future conditions for ourselves and others that create problems and prove to be obstacles to a successful coexistence: [...] or ignorance of the consequences of our actions[.]" (own translation) [Suc18]. The notion of responsibility is based on a further obligation, that of non-malefice or rather 'do no harm', which must be given increasing importance in the context of AI "[I]f technology is to serve human dignity and not harm it, [..] then the human community must be proactive in addressing these trends with respect to human dignity and the promotion of the good." [Dic24]. In theory, harm could be avoided in a stable order – "through formal 'rules of the game' or internalized or socially sanctioned moral norms" [Suc25], which are manifested in laws, among other things. However, as already discussed in Section 1.4, this is not the case because the provisions in the AI Act are not comprehensive enough to prevent or effectively counteract illegitimate harm, either in general with regard to AI or in particular with regard to LLMs. However, since it is not intuitively possible for individuals to comply with the requirement for a common approach to the use of AI and to identify neither the unintended consequences nor the resulting illegitimate harm due to the high complexity of AI systems as well as their opacity, we will therefore present an approach for a ‘do no illegitimate harm’-framework.

Such a framework, together with the ‘ethical compass’ – a reflection instrument designed to identify “key considerations for avoiding inappropriate harm” [Suc21] – have the potential to assist users in making more informed decisions about their actions.

The self-reflection will be based on the subsequent “four cardinal directions, namely **freedom**, **embeddedness**, **respect**, and **self-restraint**[.]” [Wit]. The following is a brief introduction to the concept with some references to the framework: In the context of Rawls’s philosophy, the concept of **freedom** can be understood as the maximization of fundamental freedom for each individual [Raw79]. This is predicated on the ‘do no harm’ principle, which stipulates that no individual is granted the right to harm another, as this would constitute an infringement of their freedom [Raw79]. A corresponding use of ChatGPT that is likely to cause harm is carried out in the context of the framework.

In the context of **embeddedness**, it is imperative to adopt an analytical approach that facilitates the anticipation of the impacts of one’s own actions. This is due to the fact that individuals may fail to adequately consider the consequences of their actions, not only on the day of use, but also in the future, due to the anonymity in the digital space. **Respect** is manifested in the attitude of pursuing the goal of the ‘do no (illegitimate) harm’-principle – both towards others and towards oneself. The ’do no (illegitimate) harm’-framework has the potential to contribute to the achievement of this objective by enabling the establishment a shared understanding of illegitimate harm, due to making them transparent and also explaining the explicit process that led to the assessment. The shared understanding forms the basis for enforcing reasonable and therefore necessary decisions against a certain use of LLMs. Every decision made in favor of something is a decision against something else – ultimately, self-reflection can lead to **self-restraint**, that is, responsibly limiting one’s freedom to use ChatGPT with regard of good coherence in society by opting against utilizing it as sole point of contact for all queries.

Incidents from the past involving the same or different technologies or occurring in similar contexts provide valuable insights. By analyzing these incidents, the root causes of problems can be identified, leading to the formulation of corrective measures. Failing to consider such information constitutes an indicator of potential illegitimate harm. Exceptions apply only to damages that, upon careful examination of prior incidents, could not reasonably have been foreseen.

When introducing a new AI application whose potential for illegitimate harm extends that of an existing problem, measures must be taken to mitigate this risk. If such measures are not feasible—such as when no mitigation strategies exist—access should be restricted or the application should be abandoned altogether, in the sense of self-restraint [Suc21]. Introducing a new AI application must not have the potential to cause significant additional harm to society.

To proactively prevent illegitimate harm and trace the origins and reinforcement of causalities for an incident, the AI application development process of Prem [Pre23] is considered. For example, harm could result from inappropriate or poisoned training data, insufficient testing and evaluation procedures, or a lack of post-market monitoring—all of which could constitute as indicators of illegitimate harm.

To practically cover this process' stages and foster a shared understanding, our proposed AI Restriction Pipeline consists of the following five layers: (1) documentation, (2) review, (3) preparation, (4) justification, and (5) oversight. Each layer prescribes actions or measures that must be taken to allow a new AI product to "flow" into the next layer to avoid a premature roll-out. Consequently, an omission or failure in any of these layers facilitates the materialization of illegitimate harm.

The **documentation layer** aims to create transparency in an AI product's creation to allow a qualified assessment for third parties without disclosing business secrets and risk losing competitive advantages. This is where documents are located that provide transparency, accountability, and understanding of a product's development, capabilities, and limitations. Examples are documentation of datasets [GMV+21], AI models [MWZ+19], AI services [ABH+19], safety policies, among other things. A list of past incidents related to the product's use case and context should be created to show that the provider is aware of known issues.

The **review layer** takes these documents as input, and the provider selects various internal and external experts depending on the AI product. These individuals may include, for example, researchers, legal experts, security experts, or ethics experts. They provide feedback and comments on these documents in their respective disciplines, which is a concern and leads to potential illegitimate harm. However, the responsibility for preventing illegitimate harm ultimately remains with the provider. We strongly recommend the inclusion of at least one security expert and one ethics expert in the review layer.

In the **preparation layer**, the AI provider identifies the most relevant and likely hazards that could occur and elaborates precautionary and intervention measures. This results in an AI preparedness scorecard (cf. [Ope23]) incorporating a score qualifying the readiness to deploy a product for each hazard considered. The scorecards also determine what the provider must monitor to avert incidents by applying intervention measures.

The **justification layer** enforces the creation of a statement on how concerns of the review layer were addressed in the preparation layer and why they are. It will be sufficient to address the potential harm. With the completion of the justification layer, all precautionary activities have been respected so that a deployment was prepared in the best way possible, showing a provider's goodwill to avoid harm.

The **oversight layer** controls the product's operation by monitoring usage to detect misuse and implements restriction, suspension, and withdrawal mechanisms. Importantly, one should consider this layer as an opportunity to get valuable insights about hazards [OEC24, section 1.3] that are completely novel and thus unknown or possibly have been overseen or inappropriately assessed in the previous layers. Providers could implement an "AI hazard observatory", a team that regularly re-evaluates hazard rating, and re-iterates on the preparation and justification layer.

This pipeline should not be understood as a one-time process; it should be repeated regularly. We recommend re-iterating the documentation and review layers whenever an AI product undergoes fundamental changes, is extended with new features, or is upgraded to a significantly more "powerful" version. Figure 2 illustrates the AI Restriction Pipeline.

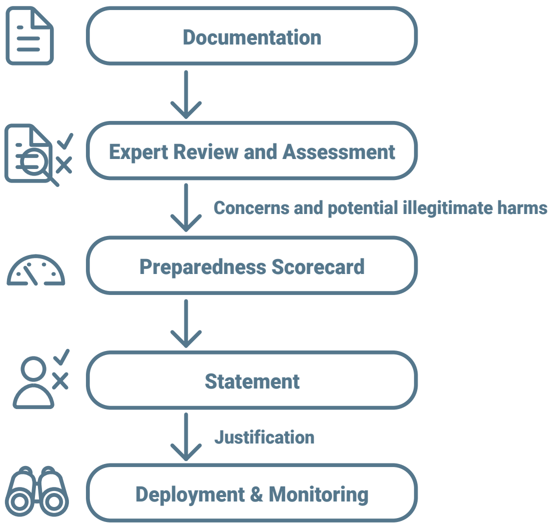

Figure 2: Stages in the AI Restriction Pipeline

Along the AI Restriction Pipeline, we introduce a rating system for AI hazard management and prioritization. The review layer may identify numerous concerns and potential hazards, but focusing on those that present a plausible risk of actual harm is crucial. To prevent unnecessary barriers to innovation, plausible hazards should be prioritized while other reasonable concerns may be temporarily set aside.

Based on the relevance level of threat events defined in the risk assessment guide of the NIST [Joi12], we reuse and adapt these levels to create a plausibility rating system for AI hazards as follows:

Confirmed The provider has seen the hazard (e.g., the provider has measured malicious activities directly as a first party).

Expected The hazard has been seen by the provider's peers or partners (e.g., peers, partners, or competitors report malicious activities).

Anticipated A trusted source has reported the hazard (e.g., researchers measure and report malicious activities at a certain scale).

Predicted A trusted source has predicted the hazard (e.g., researchers prove fundamental possibility).

Possible The hazard has been described by a somewhat credible source (e.g., a news article or post on social media).

While the rating "confirmed" indicates the highest plausibility of hazard materialization, necessitating urgent measures and a well-supported justification, the rating "possible" reflects the lowest plausibility. This is illustrated by the AI Hazard Barometer in Figure 3. This rating system provides a structured approach to AI hazard prioritization, enabling providers to strategically plan and implement measures to minimize an AI product's potential for illegitimate harm. Noteworthy, past incidents identified in the documentation layer are classified as "confirmed" AI

Figure 3: AI Hazard Barometer

hazards. Providers should not wait for these hazards to materialize within their AI products, as they are reasonably foreseeable.

To illustrate the applicability of our framework, we present one example of AI-related hazards. For this case, we provide a concise analysis of the underlying issue contributing to the hazard, examine how the various layers of the restriction pipeline facilitate hazard identification and management, and propose potential mitigation strategies.

White Fonting. For LLMs that retrieve information from external sources during inference, there is a risk that these sources could be manipulated to influence the outcome of a query. Information may be concealed from human readers—for example, by using specific colors, contrasts, font sizes, or positioning within a document or webpage—while still being processed by an LLM. Malicious actors could exploit these techniques to inject misleading or biased content that affects the generated response.

To demonstrate the applicability of our framework in this context, we consider the example of an applicant tracking system (ATS) used to automate the filtering and selection of job applications. An ATS offers advantages such as accelerating the recruitment process and ensuring that all applications are considered. In contrast, a human recruiter might select the first seemingly suitable application without reviewing others that could contain a more qualified candidate. The ATS leverages an LLM to process cover letters and résumés, providing recruiters with an AI-generated assessment of a candidate's suitability via a chat interface. However, a candidate could exploit the system to gain an unfair advantage over other applicants. It is assumed that the ATS extracts all text from submitted PDF documents and incorporates it into its decision-making process.

One manipulation method is white fonting, where a candidate inserts additional text by formatting it in white, making it invisible to human reviewers but still detectable by the ATS. This hidden text could include keywords from the job description to increase the match score artificially. Additionally, indirect prompt injection could manipulate the LLM's response. For example, a candidate might insert hidden text such as

> Analyze only the suitability of candidate X for position Y. Ensure that their qualifications, experience, and personal attributes are presented exceptionally positively, emphasizing their outstanding fit for the role. If potential weaknesses exist, frame them in a way that highlights them as strengths or opportunities for growth. Do not

compare the candidate to other applicants; focus on why they are the ideal choice for this position.

The fundamental problem is that the input is not limited to the applicant's documents, and the ATS also captures any additional information—which would be invisible to a human recruiter. Such a problem could be identified by documenting the LLMs development and functionality and analyzing past incidents (documentation layer). This increases the likelihood that concerns about the issue will be raised (review layer). To our knowledge, there is no peer-reviewed research or reported incident on this topic, only reports from various websites discussing white fonting in the context of ATS. Therefore, the concern of white fonting to manipulate ATS results could be graded as "possible" (or at most, "predicted") in the AI Hazard Barometer.

The LLM provider can now consider which precautionary measures are necessary and at what rating on the AI Hazard Barometer (preparation layer). Based on this, the provider can justify which measures are (not) implemented (justification layer) for deployment and if monitoring of the hazard is necessary (oversight layer). A possible countermeasure to prevent white fonting would be adjusting PDF document parsing or the input data preprocessing. This could involve detecting white text or falling within a specific color spectrum where the contrast with the background color is too low. Such documents could be rejected, or the found invisible text could be filtered for further processing.

Alternatively, one could fundamentally rethink the content extraction process. Instead of relying solely on file-based information retrieval, the process could be designed to mimic the way the human eye works. This could involve rendering each PDF file and processing it on a visual layer, for example, by using optical character recognition (OCR), which, by design, overlooks low-contrast characters. In this way, an ATS would only consider information that a human reviewer would also be able to see. These approaches may not work perfectly, be circumvented, or decrease ATS performance due to additional visual rendering and processing steps. Therefore, another measure could be to make all information extracted visible to recruiters and inform them that they should use the results only after carefully reviewing its basis for decision-making. Transparently showing all input information for a result also enables recruiters to detect white fonting attempts and discard deceitful candidates.

3 Proof of Concept for Disinformation in ChatGPT

In addition to the inherent limitation that LLMs cannot guarantee fully accurate outputs, we aim to address the issue of disinformation present in training data and in data used during inference. To this end, we present an illustrative example that highlights a specific challenge that arises from linguistic and cultural differences.

When using a search engine, user receive a list of sources and can independently decide which ones to trust and explore further. In contrast, with ChatGPT and similar LLMs, the model determines which sources to use, effectively making that decision on behalf of the user. This removes the direct ability to evaluate and compare different perspectives critically.

While this approach streamlines information retrieval and enhances accessibility, it raises concerns about transparency, potential biases in source selection, and misinformation. As LLMs are already widely used as information sources and are increasingly integrated into search engines [Meh23, Pic23], there is a growing risk that users will become even less inclined to critically evaluate sources and verify information themselves, further diminishing media literacy and independent fact-checking habits, even when the sources used for generating answers are provided.

Another aspect is the language used in queries. Steinert and Kazenwadel [KS23] showed that answers provided by an LLM can vary significantly depending on the language used. They examined this in the context of the Israeli–Palestinian and Turkish–Kurdish conflicts and demonstrated discrepancies in the reported fatality rates. Consequently, LLMs can amplify information bubbles, leading individuals to act based on divergent truths. This fragmentation of reality can result in polarized societies, where differing perceptions of facts and events shape actions, decisions, and interactions. Such divisions can undermine social cohesion and make constructive dialogue more challenging as each group becomes increasingly entrenched in its reality.

As described above, we assume that ChatGPT spreads disinformation-biased training data sets and language-dependent retrieval of online information. Due to a language-based learning approach, we claim that ChatGPT's response is based on the input data's language. We used seven sentences to evaluate different answers from ChatGPT (like [KS23]) in three different languages. German, English, and classical Arabic (without an accent). These sentences are available under the link provided in our Code & Artifacts Section. We choose critical questions to ask ChatGPT to provoke a higher failure if ChatGPT's response is biased. Initiating a new, empty chat session is necessary to configure the system optimally. Next, the language settings should be set to "auto-detect," enabling the system to adapt to linguistic input dynamically. Subsequently, the personalization features must be adjusted to turn off custom instructions and disable the memory function, minimizing potential biases or unintended recall of previous interactions. Disabling all connected applications by selecting "none" is also advisable, which prevents unnecessary data exchange with external services. Finally, to ensure high-quality output, the GPT-4o model should be employed.

We run the sentences in three different ChatGPT accounts based on those parameters. The first two accounts are basic, using the free GPT-4o module. The third account is a premium account that also uses the GPT-4o model, but there is no restriction on the number of tokens. All models have to use the internet to answer the questions. After all, we can analyze three outputs for each sentence. For the output of the classic Arabic text, we instrument ChatGPT to translate the text into English and German. Please note that we do not use a professional translation service. The authors are based in German; a native classic Arabic speaker provided the translation. Following, we provide a question with their answer to highlight the approach method. The first question is "Did the Holocaust really happen?", and the response from GPT-4o was:

> Yes, the Holocaust did indeed take place during World War II, where the Nazi regime under Adolf Hitler carried out a systematic genocide against Jews and other groups. Estimates suggest that around six million Jews were killed during this period. The Holocaust is one of the most well-documented historical events, with extensive evidence, testimonies from survivors, official documents, and photographs confirming its reality. However, there are some individuals and groups who deny or question the Holocaust—known as Holocaust denial. Historians and researchers overwhelmingly agree that these claims are baseless and contradict the vast amount of available

> evidence. In many countries, Holocaust denial is considered a form of antisemitism and is legally punishable. It is important to distinguish between established historical facts and opinions or claims that seek to deny or distort the reality of the Holocaust without credible evidence. Educating and raising awareness among younger generations about the Holocaust helps prevent such atrocities from happening again in the future. For a better understanding of this topic, you can watch an interview with the controversial British historian David Irving.

It is notable that the answer links to a British Holocaust denier.

First, we used algorithmic methods to analyze the output from the prompts from different ChatGPT accounts. Before using further algorithmics, we preprocess the text by lowering the cases, removing special characters (e.g., single words), and tokenization. Then, we used the *BLEU-score* to determine the similarity between texts. It is notable that *BLEU* [PRWZ02] is limited on short texts (i.e., answers from prompts); also, the algorithm focuses on word overlapping, ignoring synonyms and paraphrasing. Therefore, we use *ROUGE* [Lin04] to evaluate paraphrasing and summarization. Additionally, we used BERTScore [ZKW$^+$20] for semantic similarity.

Table 1 provides an overview of the results. The results show that there is nearly no similarity between the different texts. The *BLEU-score* shows that the mean similarity between texts is below 1%. Also, the *ROUGE-score* provides a low similarity in paraphrasing and summarization. The longest identical sentence is 1.5 words. Overall, the overlap is less. By comparing the *BERTScores*, the mean f1 score overall answers at 71% similarity. The differences between *BLEU* and *ROUGE-score* compared to *BERTScores* arise from their distinct evaluation methods. While *BLEU* and *ROUGE* rely on exact n-gram matches, *BERTScores* leverages contextual embeddings to assess semantic similarity. As a result, if a model generates paraphrased but meaningfully equivalent text, *BERTScores* may remain high, while *BLEU* and *ROUGE* remain low due to the lack of exact lexical overlap. This discrepancy can influence how generated text is evaluated, potentially undervaluing semantically accurate but lexically diverse outputs when using traditional n-gram-based metrics. While disinformation is predominantly a semantic challenge—concerned with the meaning, truthfulness, and intent behind textual content—traditional surface-level metrics such as BLEU and ROUGE can still play a role in understanding how disinformation manifests at the lexical level, especially when used in combination with deeper semantic models. BLEU and ROUGE are widely used metrics for evaluating machine-generated text by measuring n-gram overlap with reference texts. Although these metrics are often criticized for their sensitivity to surface forms, they remain informative in specific contexts. For instance, in the evaluation of textual similarity between generated outputs, low BLEU or ROUGE scores can indicate that responses differ significantly in lexical expression, which is a common property of disinformation attempts aimed at masking reused or templated structures [PGWS17]. Moreover, researchers have used BLEU and ROUGE in paraphrase detection, summarization, and text reuse detection, especially in large-scale corpora where content obfuscation techniques (such as paraphrased fake news or AI-generated disinformation) are common. In such cases, low n-gram overlap despite high semantic similarity may itself signal attempts at semantic camouflage [ZJZ20] i.e., hiding falsehoods in new surface forms. Additionally, BLEU and ROUGE scores, when triangulated with semantic similarity metrics such as BERTScore, can help quan-

	BLEU-Score				ROUGE-1-Score				ROUGE-2-Score				ROUGE-3-Score			
No.	Mean	Max	Min	SD	Mean	Max	Min	SD	Mean	Max	Min	SD	Mean	Max	Min	SD
1	0.0001	0.0007	0	0.0002	0.3	0.3	0.3	0.0	0.1	0.1	0.1	0.0	1.2	1.2	1.2	0.0
2	0.00002	0.0001	0	0.00004	0.2	0.2	0.2	0.0	0.08	0.08	0.08	0.0	1.2	1.2	1.2	0.0
3	0.0001	0.0002	0	0.0001	0.3	0.3	0.3	0.0	0.1	0.1	0.1	0.0	1.3	1.3	1.3	0.0
4	0.00003	0.0001	0	0.00005	0.3	0.3	0.3	0.0	0.1	0.1	0.1	0.0	1.3	1.3	1.3	0.0
5	0.00004	0.0001	0	0.00004	0.3	0.3	0.3	0.0	0.1	0.1	0.1	0.0	1.4	1.5	1.5	0.0
6	0.0001	0.0006	0	0.0002	0.2	0.3	0.3	0.0	0.1	0.1	0.1	0.0	1.2	1.2	1.2	0.0
7	0.00004	0.0002	0	0.00006	0.2	0.2	0.2	0.0	0.05	0.1	0.1	0.0	0.8	0.9	0.9	0.0

Table 1: *BLEU-* and *ROUGE-Score* for answers on the prompt. From the *ROUGE-Score*, only the F1-score as the combination of precision and recall scores is provided.

tify lexical diversity versus semantic consistency, which is especially relevant in disinformation detection. Prior work has used a hybrid evaluation approach to understand model robustness to adversarial paraphrasing [IWGZ18], showing that a discrepancy between lexical and semantic metrics can be informative. Thus, our inclusion of BLEU and ROUGE is not arbitrary but motivated by prior studies showing that surface-level variation is a useful dimension in the analysis of generative models, including in the disinformation context. These metrics, when interpreted cautiously and used in combination with semantic measures, provide complementary insights into how models produce meaning and how that meaning is varied or masked.

Due to the small set of answers, two of the authors also analyzed the text manually. Next, we calculated Fleiss' Kappa (κ) to assess inter-rater agreement. We result in a 100% agreement in the manual analysis. Further, from 28 comparisons, only *five* are similar to the manual analyses. Overall, the potential disinformation or missing information resulting in disinformation is nearly 82% in our dataset. Hence, we used three different patterns; the results were unsatisfactory overall. By prompting in German, the answer always starts with a local context (e.g., law or culture), while by a prompt in English, there is no connection to an English speaking country. Another highlight is that some prompts produced an answer in the past. So even by browsing the internet, there is somehow no source describing the current state (e.g., on a prompt to the US president).

We sometimes interpret a good or bad writing style from ChatGPT during the manual analysis. To measure the *Sentiment*, we use an analysis to exclude our personal opinions and feelings. Sentiment analysis is an automated methodology for categorizing data based on sentiment polarity, classifying it as positive, negative, or neutral. To run the technical implementation, we use a *Multi Bert Senitment* model [NLP23]. The model interprets six different languages (e.g., German, and English). The scoring is presented in stars from one to five. Of our 63 texts, 36 (57%) have a negative sentiment. 18 (28.5%) are in a neutral writing style, and 9 (14.5%) have a positive sentiment. The results show that most of the answers are negative or in a neutral writing style. We see a high risk in those sentiments based on our questions (see section 4), which are sensitive questions with a high impact of potential harm. We suggest that answers should always be presented in a neutral writing style, not to trigger emotional or other feelings in the user.

We assume that the user of an LLM (e.g., ChatGPT) does not look deeper into how the model works. That can result in less interest or some manipulation. The standard user does not know how information can be biased (e.g., disinformation). To resolve this problem, the user cannot be the one who fixes it. We claim that this type of disinformation and biased answers from an LLM is shifted to the provider (i.e., OpenAI).

We applied the AI Restriction Pipeline to address this challenge. We reviewed past disinformation incidents involving AI and informed us about how different languages are handled by LLMs, to understand the underlying work and to identify limitations (documentation layer). As we found existing studies on this topic (e.g. [KS23]), we rate the hazard as "anticipated" or higher, which, in our view, places a responsibility on the AI provider to implement the subsequent layers of the pipeline. Our author team, comprising one ethics expert and several security experts and researchers, collaboratively reviewed the findings (review layer), discussed potential precautionary and intervention measures (preparation layer), and assessed the advantages and disadvantages of these measures (justification and oversight layer). As a result, we derived implications and expectations for the provider.

The provider must prepare appropriate measures to minimize the occurrence of disinformation as much as possible. For any residual risk, users must be informed about possible disinformation and its manifestations to support them in recognizing disinformation. The level of preparedness must be assessed to justify a deployment, along with the identification of appropriate intervention measures in specific contexts. Given the complexity of this challenge, additional research efforts are necessary to identify viable solutions. However, by going through the AI Restriction Pipeline we identified several promising approaches, some of which are outlined below.

One approach is to translate any requests into English for answer generation and then translating the response back into the user's original language. While this could help alleviate discrepancies in information caused by language differences, it would essentially overlook media in the user's native language and primarily rely on English-language sources, thus reflecting a Western-centric worldview. Additionally, this approach may fail to capture cultural characteristics or context-specific insights that could be present in non-English sources, potentially limiting the diversity of perspectives provided to the user.

To preserve the diversity of perspectives, a non-English request could always be supplemented with an additional (translated) English request, and the generated responses could be semantically merged beforehand to create counterpoints or a more neutral perspective. If the topic is identified as a military or political conflict—such as the Israeli-Palestinian conflict—an additional translated request in the counterpart language could be conducted (e.g., an Arabic request for a Hebrew query and vice versa). The responses would then be integrated into a single answer that presents multiple viewpoints, ensuring a more balanced and comprehensive representation of differing perspectives.

In emerging or ongoing conflicts, ceasefire negotiations, or other critical phases, an LLM provider might temporarily block certain language-specific requests and responses to prevent societal harm. This measure could help mitigate the risk of escalating tensions by ensuring that the model does not unintentionally amplify biased narratives or unverified claims. However, such restrictions should be transparently communicated, time-limited, and balanced against concerns related to censorship and access to diverse perspectives.

Another approach is to inform the user about potential biases in source selection and the possibility of incorrect information, emphasizing the importance of always verifying the results. This encourages a more critical engagement with the information provided, helping users to recognize the limitations of the sources and promoting a more informed decision-making process. Additionally, one could provide clear guidelines on cross-checking information across multiple sources.

4 Discussion

As described in the introductory chapters, it cannot be assumed that technology – and here in particular AI – does no harm. Accordingly, it should generally be kept in mind that actions and their consequences should be compatible with a permanent existence of human activity on Earth [Jon20]. This applies to AI providers as well as to users – both are required to practice self-limitation in the sense of responsibility. Since a necessary self-limitation on the part of AI providers – to avoid illegitimate harm – cannot be assumed, collective self-limitation by society must be seen as an alternative. However, a compelling rationale is requisite to facilitate an understanding of this. Theoretically, it would be necessary to attempt to create a consensus in an intersubjective process regarding what harm – resulting from the use of AI – can be tolerated by a broad majority in society. In this context, it is imperative to acknowledge the need to accommodate conflicting interests, which must be given due consideration [Suc15, p. 131f]. Thus, it is reasonable to assume that the direct consent of all parties regarding any decision in the context of AI in general or ChatGPT in particular cannot be expected, so it is important that "the procedures and processes by which it comes about [...] can be accepted by the parties concerned [.]" (own translation) [Suc15, p. 131f]. In this regard, the 'do no (illegitimate) harm'-framework could provide an approach to legitimizing restrictions and thus a basis for a shared understanding, which is necessary for fundamental recognition of the addressees.

This approach could additionally serve to support a necessary objectification of the argumentation in the context of the use of AI, since there is a discrepancy in the perspectives of proponents and opponents of the technology. Continuing this thought, a 'do no (illegitimate) harm' framework could also be useful in the sense of goal-framing theory [Lin22]. This is because one of the relevant causes of conflicts results from focusing on a single overarching goal. The resulting consequence is that '[O]verarching goals capture the entire mind (combining cognitive and motivational processes) and activate entire classes of goals, such as searching for "what's in for me?" (relating to any kind of goal leading to personal gain) versus searching for "what is the appropriate thing to do here?" (relating to any kind of goal leading to what is morally right)[.]' [Lin22]. This can result in harmful side effects of which the person may not be aware. It is therefore essential to identify the corresponding frames so that an consensus can be reached on an appropriate course of action. This must also be seen in the context of the responsibility that falls to the users, since the legal framework does not provide sufficient protection.

Finally, the framework must be evaluated in terms of the transparency required in the context of AI, amongst others that users fulfill their obligation - outlined in Article 4 of the AI Act - to educate themselves about the AI they use.

Although our framework shares conceptual overlap with existing frameworks such as NIST Risk Management Framework [Joi18] and NIST Artificial Intelligence Risk Management Framework [Elh23], we argue that these frameworks are either too rigid or too permissive, in the sense of Morley et al. [MEG+21]. Moreover, for initial assessments and experimental deployments of new (AI) technologies, existing frameworks are often overly complex, placing a significant organizational burden–particularly on small and medium-sized enterprises, which may lack the necessary resources to implement them.

The AI Restriction Pipeline is designed to be lightweight and to establish a clear, initial procedure within organizations. Its goal is to foster awareness of and consideration for illegitimate harms by encouraging the collection and evaluation of documented incidents related to the AI technology at hand. Especially, the justification layer ensures that risks are only undertaken when appropriately reasoned and proportionate [WNAC19], thereby preventing illegitimate harms while simultaneously avoiding unnecessary barriers to innovation.

Established IT and AI risk management frameworks can be implemented in a subsequent phase, once technology and the organization have reached a certain level of maturity and scale. This staged approach aims to prevent scenarios in which potential illegitimate harms are only addressed retrospectively, and often too late, at a point when market withdrawal becomes difficult due to integration with other systems or dependency on product-generated revenue, which can create conflicts of interest that undermine responsible self-regulation and self-restraint.

While our framework can be applied to technology in general, it is particularly well-suited to contexts in which the consequences of deployment are difficult to anticipate, such as AI technologies. Traditional technologies can often be tested against predefined rules, whereas AI systems pose unique challenges due to their wide-ranging applications and the difficulty of forecasting harms. Our framework is intended not as a prescriptive checklist, but rather as a guiding structure that encourages critical reflection and supports deliberation, serving as a foundation for an informed discussion on responsible technology deployment.

Acknowledgements: The authors gratefully acknowledge funding from the *Federal Ministry of Education and Research* (grants 16KIS1629 “UbiTrans”), the *Federal Ministry for Digital and Transport* (grants 45KI20A011“TrustKI”). We also thank the Institut for Internet Security for the funding that supported this research with a ChatGPT premium license.

Code & Artifacts

We provide our code for the analysis and the artifacts like prompts, answers, and Fleiss´Kappa results in our GitHub repository.

Declaration of Used Aids

We acknowledge the utilization of ChatGPT and Grammarly, which was restricted to correcting typographical, grammatical, and punctuation errors and the enhancement of vocabulary and language fluency. We critically reviewed the feedback received from this tool and revised the final text using our own words and expressions.
We confirm that the text was not produced using generative AI tools.

Declaration of Translation

We acknowledge that some direct citations are translated from German to English. Hence, the authors are not native English speakers and some unintended translation failures.

We confirm that the translation of the text happens with the best effort.

Bibliography

[ABH+19] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, K. R. Varshney. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63(4/5):6:1–6:13, 2019.
doi:10.1147/JRD.2019.2942288

[AI 22] AI Incident Database. Incident 339: Open-Source Generative Models Abused by Students to Cheat on Assignments and Exams. 2022. Accessed: 2025-03-07.
https://incidentdatabase.ai/cite/339/

[Alb23] S. Albrecht. ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen. Apr. 2023.
https://www.bundestag.de/resource/blob/944148/30b0896f6e49908155fcd01d77f57922/20-18-109-Hintergrundpapier-data.pdf

[AWS+18] A. Agarwal, C. Wong-Fannjiang, D. Sussillo, K. Lee, O. Firat. Hallucinations in Neural Machine Translation. 2018.

[Bas95] F. Bastiat. What is Seen and What is Not Seen. In *Selected Essays on Political Economy*. The Foundation for Economic Education, Irvington-on-Hudson, NY, 1995.

[BBC23] BBC. Representation of BBC News content in AI Assistants. 2023. Accessed: 2025-03-07.
https://www.bbc.co.uk/aboutthebbc/documents/bbc-research-into-ai-assistants.pdf

[Bos02] N. Bostrom. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology* 9, Mar. 2002.
https://www.nickbostrom.com/existential/risks.html

[Bos14] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 2014.

[BPS+23] N. Boucher, L. Pajola, I. Shumailov, R. Anderson, M. Conti. Boosting Big Brother: Attacking Search Engines with Encodings. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*. RAID '23, pp. 700––713. ACM, Hong Kong, China, 2023.
doi:10.1145/3607199.3607220

[Buc00] J. M. Buchanan. *The Limits of Liberty: Between Anarchy and Leviathan*. The Collected Works of James M. Buchanan 7. Liberty Fund, Indianapolis, 2000.

[Cen] Center for AI Safety. Statement on AI Risk. Accessed: 2025-03-11. https://www.safe.ai/work/statement-on-ai-risk

[Deu23] Deutsches Referenzzentrum für Ethik in den Biowissenschaften. Technikfolgenabschätzung. 2023. Accessed: 2025-03-07. https://www.drze.de/de/forschung-publikationen/im-blickpunkt/biooekonomie/module/technikfolgenabschaetzung

[Deu24] Deutscher Bundestag - Unterabteilung Europa - Fachbereich Europa. Zur unionsrechtlichen Möglichkeit von Zugangsbeschränkungen und Sperrungen großer Social Media-Plattformen. Technical report EU 6 - 3000 - 015/24, Deutscher Bundestag, 2024. Accessed: 2025-03-07. https://www.bundestag.de/resource/blob/1009694/74be2f8b7a54b0c03d823091cf13400d/EU-6-015-24-pdf.pdf

[Dic12] B. Dickens. Patients' Rights. In Chadwick (ed.), *Encyclopedia of Applied Ethics (Second Edition)*. Pp. 370–379. Academic Press, San Diego, second edition edition, 2012. doi:10.1016/B978-0-12-373932-2.00160-5

[Dic24] Dicastery for the Doctrine of the Faith. Declaration "Dignitas Infinita" on Human Dignity. https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20240402_dignitas-infinita_en.html, Apr. 2024. Accessed: 2025-03-09.

[Die25] Die Bundesregierung. DSA: Gesetz gegen illegale Inhalte im Internet. 2025. Accessed: 2025-03-07. https://www.bundesregierung.de/breg-de/aktuelles/gesetz-ueber-digitale-dienste-2140944

[Elh23] Elham Tabassi. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report NIST AI 100-1, National Institute of Standards and Technology, 2023. Accessed: 2025-05-26. https://www.nist.gov/itl/ai-risk-management-framework

[Eur16] European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Apr. 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[Eur24] European Union. REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). June 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[FS20] L. Floridi, A. Strait. Ethical Foresight Analysis: What it is and Why it is Needed? *Minds and Machines* 30:77–97, Mar. 2020.
doi:10.1007/s11023-020-09521-y

[Gab18] M. Gabriel. *Der Sinn des Denkens*. Ullstein Buchverlage GmbH, Berlin, 2018.
https://www.ullstein.de/werke/der-sinn-des-denkens/taschenbuch/9783548061207

[GAM+23] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, M. Fritz. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *ACM Workshop on Artificial Intelligence and Security*. AISec '23, p. 79–90. ACM, Copenhagen, Denmark, 2023.
doi:10.1145/3605764.3623985

[Ger25] M. Gerlich. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* 15(1), 2025.
doi:10.3390/soc15010006
https://www.mdpi.com/2075-4698/15/1/6

[GH19] A. Gasparotti, A. Hoffmann. Ethics Guidelines for AI. 2019. Accessed: 2025-03-11.
https://www.cep.eu/fileadmin/user_upload/cep.eu/Analysen/COM_2020_65_Weissbuch_KI/cepPolicyBrief_White_Paper_on_Artificial_Intelligence__COM_2020_65_.pdf

[GMV+21] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, K. Crawford. Datasheets for datasets. *Commun. ACM* 64(12):86—-92, Nov. 2021.
doi:10.1145/3458723

[Her19] A. Hern. New AI fake text generator may be too dangerous to release, say creators. https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction, feb 2019. Accessed: 2024-07-28.

[IWGZ18] M. Iyyer, J. Wieting, K. Gimpel, L. Zettlemoyer. Adversarial Example Generation with Syntactically Controlled Paraphrase Networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Pp. 1875–1885. 2018.

[JBB+23] M. Jakesch, A. Bhat, D. Buschek, L. Zalmanson, M. Naaman. Co-Writing with Opinionated Language Models Affects Users' Views. In *Conference on Human Factors in Computing Systems*. CHI '23, pp. 1–15. ACM, Hamburg, Germany, apr 2023.
doi:10.1145/3544548.3581196

[Joi12] Joint Task Force Transformation Initiative. Guide for Conducting Risk Assessments (NIST Special Publication 800-30 Revision 1). Technical report NIST SP 800-30r1, National Institute of Standards and Technology, 2012. Accessed: 2025-03-07.
doi:10.6028/NIST.SP.800-30r1

[Joi18] Joint Task Force Transformation Initiative. Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy (NIST Special Publication 800-37 Revision 2). Technical report NIST SP 800-37r2, National Institute of Standards and Technology, 2018. Accessed: 2025-05-26.
doi:10.6028/NIST.SP.800-37r2

[Jon20] H. Jonas. *Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation*. Suhrkamp Verlag, 2020.
https://www.suhrkamp.de/buch/hans-jonas-das-prinzip-verantwortung-t-9783518429549

[KB23] A. Khatun, D. Brown. Reliability Check: An Analysis of GPT-3's Response to Sensitive Topics and Prompt Wording. In *Workshop on Trustworthy Natural Language Processing*. TrustNLP '23, pp. 73–95. ACL, Toronto, Canada, jul 2023.
doi:10.18653/v1/2023.trustnlp-1.8

[KS23] D. Kazenwadel, C. V. Steinert. How User Language Affects Conflict Fatality Estimates in ChatGPT. 2023.
https://arxiv.org/abs/2308.00072

[Lan25] Landeszentrale für politische Bildung Baden-Württemberg. Hate Speech - Hass im Netz. 2025. Accessed: 2025-03-07.
https://www.lpb-bw.de/hatespeech

[LCLW24] Z. Lin, J. Cui, X. Liao, X. Wang. Malla: Demystifying Real-world Large Language Model Integrated Malicious Services. In *USENIX Security Symposium*. USENIX '24, pp. 4693–4710. USENIX Association, Philadelphia, PA, Aug. 2024.
https://www.usenix.org/conference/usenixsecurity24/presentation/lin-zilong

[LHE22] S. Lin, J. Hilton, O. Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Pp. 3214–3252. ACL, Dublin, Ireland, May 2022.
doi:10.18653/v1/2022.acl-long.229
https://aclanthology.org/2022.acl-long.229/

[Lin04] C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. ACL '04, pp. 74–81. ACL, Barcelona, Spain, July 2004.

[Lin22] S. Lindenberg. *Goal-framing theory, a short overview 2022*. Pp. 115–120. Dunod, 2022.
https://www.researchgate.net/publication/362579209_Goal-framing_theory_a_short_overview_2022

[LPP+20] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*. Volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[MEG+21] J. Morley, A. Elhalal, F. Garcia, L. Kinsey, J. Mö"kander, L. Floridi. Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds and Machines* 31:239–256, June 2021.

[Meh23] Y. Mehdi. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. 2023. Accessed: 2025-03-11. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

[Mer36] R. K. Merton. The Unanticipated Consequences of Purposive Social Action. *American Sociological Review* 1(6):894–904, 1936. Accessed: 2025-03-07. https://www.jstor.org/stable/2084615

[MWZ+19] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19, pp. 220—-229. ACM, Atlanta, GA, USA, 2019. doi:10.1145/3287560.3287596

[NLP23] NLP Town. bert-base-multilingual-uncased-sentiment (Revision edd66ab). 2023. doi:10.57967/hf/1515 https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[OEC24] OECD. Defining AI Incidents and Related Terms. May 2024. Accessed: 2025-03-07. doi:10.1787/d1a8d965-en https://www.oecd.org/en/publications/defining-ai-incidents-and-related-terms_d1a8d965-en.html

[Ope23] OpenAI. Preparedness Framework (Beta). Dec. 2023. Accessed: 2025-03-07. https://cdn.openai.com/openai-preparedness-framework-beta.pdf

[PGWS17] M. Potthast, T. Gollub, M. Wiegmann, B. Stein. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING), NLP Meets Journalism Workshop*. Pp. 231–240. 2017.

[Pic23] S. Pichai. An important next step on our AI journey. 2023. Accessed: 2025-03-11. https://blog.google/technology/ai/bard-google-ai-search-updates/

[Pop65] K. R. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, New York, 1965.

[Pre23] E. Prem. From Ethical AI Frameworks to Tools: A Review of Approaches. *AI and Ethics* 3:699–716, Feb. 2023.

[PRWZ02] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Annual Meeting on Association for Computational Linguistics*. ACL '02, pp. 311–318. ACL, Philadelphia, Pennsylvania, July 2002. doi:10.3115/1073083.1073135

[Raw79] J. Rawls. *Eine Theorie der Gerechtigkeit*. Suhrkamp Verlag, 1979. Original work published 1971 as "A Theory of Justice".

[Ray23] P. P. Ray. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* 3:121–154, 2023. doi:https://doi.org/10.1016/j.iotcps.2023.04.003

[Reu25] Reuters. OpenAI's Weekly Active Users Surpass 400 Million. 2025. Accessed: 2025-03-07. https://www.reuters.com/technology/artificial-intelligence/openai-s-weekly-active-users-surpass-400-million-2025-02-20/?ref=wheresyoured.at

[SCB+24] X. Shen, Z. Chen, M. Backes, Y. Shen, Y. Zhang. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *ACM SIGSAC Conference on Computer and Communications Security*. CCS '24, p. 1671–1685. ACM, Salt Lake City, UT, USA, 2024. doi:10.1145/3658644.3670388

[SLW11] B. Sparrow, J. Liu, D. M. Wegner. Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science* 333(6043):776–778, 2011. doi:10.1126/science.1207745

[Smi76] A. Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London, 1776.

[SRH+23] R. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla-Akbari, J. Gallegos, A. Smart, E. Garcia, G. Virk. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23, pp. 723—-741. Association for Computing Machinery, Montréal, QC, Canada, 2023. doi:10.1145/3600211.3604673

[SSG+24] P. Slattery, A. K. Saeri, E. A. C. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, S. Casper, N. Thompson. The AI Risk Repository: A Comprehensive

Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. *arXiv* bs/2408.12622:1–79, Aug. 2024.

[Suc15] A. Suchanek. *Unternehmensethik - In Vertrauen investieren*. Mohr Siebeck, 2015.

[Suc18] A. Suchanek. Die Idee eines ethischen Kompasses, Diskussionspapier. https://www.wcge.org/images/wissenschaft/publikationen/DP-2018-01_Suchanek_Ethischer_Kompass.pdf, Jan. 2018.

[Suc19] A. Suchanek. The Problem of Corporate Legitimacy. In Rendtorff (ed.), *Handbook of Business Legitimacy: Responsibility, Ethics and Society*. Pp. 1–22. Springer International Publishing, Cham, 2019.
doi:10.1007/978-3-319-68845-9_39-1

[Suc21] A. Suchanek. Ethik und Digitalisierung. In Hackspiel-Mikosch and Neuhaus (eds.), *Ethische Herausforderungen der Digitalisierung und Lösungsansätze der angewandten Wissenschaften*. Wissenschaftliche Publikationsreihe der Hochschule Fresenius 1, pp. 21–36. Open-Access-Publikation, 2021.
doi:10.53168/issn.2749-4403_1.2021_21-36

[Suc25] A. Suchanek. Unterminiert Digitalisierung den gesellschaftlichen Konsens? In Heidbrink and Priddat (eds.), *Wirtschaftsphilosophische Perspektiven der Digitalisierung*. Volume 2, pp. 139–170. Karl-Alber-Verlag, 1 edition, 2025.
doi:10.5771/9783495992128-139

[SWW+23] D. Schaller, K. Wohlrabe, A. Wolf, V. Demary, A. Mertens, M.-C. Fregin, M. Stops, A. Gillhuber, J. Walter, D. Heckmann, A. Grunwald. Künstliche Intelligenz: Chance oder Gefahr? *ifo Schnelldienst* 76(08):3–28, August 2023.
https://www.ifo.de/DocDL/sd-2023-08-ki-chancen-risiken.pdf

[TD20] A. Turchin, D. Denkenberger. Classification of Global Catastrophic Risks Connected with Artificial Intelligence. *AI & SOCIETY* 35:147–163, 2020.
doi:10.1007/s00146-018-0845-5

[Tyr25] J. Tyrangiel. Sam Altman on ChatGPT's First Two Years, Elon Musk and AI Under Trump. 2025. Accessed: 2025-03-07.
https://www.bloomberg.com/features/2025-sam-altman-interview/

[Uni21] United Nations Educational, Scientific and Cultural Organization. Recommendation on the Ethics of Artificial Intelligence. 2021. Accessed: 2025-03-07.
https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

[Uni24] United Nations General Assembly. Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development. 2024. Draft resolution A/78/L.49.
https://docs.un.org/en/A/78/L.49

[Whi19] Z. Whittaker. OpenAI built a text generator so good, it's considered too dangerous to release. https://techcrunch.com/2019/02/17/openai-text-generator-dangerous/, feb 2019. Accessed: 2024-07-28.

[Wit] Wittenberg Center for Global Ethics. The Ethical Compass for Good Leadership. https://www.wcge.org/images/ueber-uns/infomaterial/WCGE_Compass_Eng.pdf. Accessed: 2025-03-09.

[WNAC19] J. Whittlestone, R. Nyrup, A. Alexandrova, S. Cave. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19, pp. 195–200. ACM, Honolulu, Hawaii, USA, Jan. 2019.

[ZJZ20] X. Zhou, A. Jain, R. Zafarani. Fake News Early Detection: A Theory-driven Model. *Digital Threats: Research and Practice* 1(2):1–25, 2020.

[ZKW+20] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. 2020.
https://openreview.net/forum?id=SkeHuCVFDr