Information Hazard: Understanding the Perils of True Knowledge

June 2025

Executive Summary

This report provides a comprehensive examination of Nick Bostrom's concept of Information Hazard, a critical framework for understanding how the dissemination of true information can inadvertently or intentionally lead to significant harm. Formalized in 2011, this concept challenges conventional notions of transparency by positing that certain verified truths may pose risks to individuals, societies, or even humanity itself. The report details Bostrom's foundational definition, explores his nuanced typology of information hazards—including data, idea, knowing-too-much, and attention hazards—and presents ten deeply explained examples. These examples are rigorously analyzed and ranked by their probability of occurrence and potential impact, ranging from personal psychological distress to global catastrophic risks.

A central theme emerging from this analysis is the inherent paradox that knowledge, traditionally viewed as an unmitigated good, can simultaneously create new vulnerabilities and avenues for severe harm. This necessitates a fundamental re-evaluation of information policy, moving beyond merely combating falsehoods to strategically managing the flow of verified truths. The report highlights the subtle and often overlooked nature of these hazards, particularly in rapidly advancing technological domains such as artificial intelligence and synthetic biology, where the potential for misuse or unintended consequences is amplified. The discussion extends to the complex ethical dilemmas involved in balancing the principle of freedom of information with the imperative for safety, revealing a crucial trust-risk trade-off in information governance. Ultimately, the report concludes by advocating for a holistic,

adaptive approach to knowledge management, emphasizing responsible innovation, robust ethical frameworks, and a shared societal understanding of knowledge's dual potential.

1. Introduction: The Perils of True Information

The conventional understanding of knowledge often equates it with progress, empowerment, and enlightenment. Yet, a growing body of philosophical inquiry suggests that true information, far from being universally beneficial, can harbor significant risks. This counter-intuitive notion forms the bedrock of the concept known as Information Hazard.

1.1. Defining Information Hazard: Nick Bostrom's Foundational Concept

The formal concept of an "Information Hazard," also referred to as an "infohazard" or "cognitohazard," was rigorously defined by the philosopher Nick Bostrom in 2011.[1] At its core, an information hazard is characterized as "a risk that arises from the dissemination of (true) information that may cause harm or enable some agent to cause harm".[1] This precise definition is paramount to understanding the concept, as it explicitly distinguishes information hazards from the more commonly discussed dangers of false information, such as misinformation or disinformation.[3] The focus here is exclusively on verified truths and their potential for detrimental outcomes.

This framework introduces a profound tension with the widely accepted principle of freedom of information. The very premise of an information hazard suggests that certain categories of true information might be too dangerous for unrestricted dissemination, thereby challenging conventional societal norms that champion openness and transparency.[2] The implication is that the act of acquiring and sharing knowledge, traditionally seen as a primary driver of human advancement, can simultaneously generate new vulnerabilities and pathways to catastrophic harm. This inherent conflict between the pursuit of knowledge and the imperative of safety forces a re-evaluation of the unconditional dissemination of information, particularly in domains where scientific and technological breakthroughs rapidly yield powerful

capabilities with dual-use potential.

Bostrom's work on information hazards emerged from his broader research at the Future of Humanity Institute (FHI) at the University of Oxford. FHI operated as a prominent interdisciplinary research center dedicated to exploring "big-picture questions about humanity and its prospects," with a significant focus on global catastrophic and existential risks.[5] Within this context, the study of information hazards played a crucial role, contributing to a deeper understanding of how knowledge itself could contribute to humanity's most profound challenges.

1.2. The Subtlety and Overlooked Nature of Information Hazards

Information hazards are frequently described as "often subtler than direct physical threats, and, as a consequence, are easily overlooked".[4] This subtlety arises because the information itself does not directly inflict harm; rather, it enables an agent to cause harm or triggers a sequence of events that lead to detrimental outcomes. Unlike immediate, tangible dangers, the mechanisms of harm associated with information hazards often operate indirectly, below the threshold of immediate perception or conventional risk assessment. This inherent lack of immediate tangibility makes them particularly insidious and challenging for individuals, organizations, and governments to proactively identify and manage. It points to a systemic blind spot in traditional security and risk management paradigms, which are often geared towards more direct, observable threats.

A classic illustration of this concept is the stringent classification of information pertaining to thermonuclear weapons. The inherent danger posed by the knowledge of how to construct such devices necessitates strict controls on who can access this information.[2] By limiting access, the potential for "massive amounts of harm to others" is directly mitigated.[2] This real-world example underscores the practical application of managing information hazards through restricted access, demonstrating that security measures focused solely on physical infrastructure or cyber vulnerabilities are insufficient without an equally rigorous focus on the content and context of information itself. Proactive identification of these subtle risks therefore requires a multidisciplinary approach, combining insights from philosophy, ethics, technology, and social sciences to expand existing risk assessment frameworks to include intangible assets like knowledge and ideas as potential sources of catastrophic risk.

2. Bostrom's Typology of Information Hazards

To systematically understand the diverse ways in which true information can lead to harm, Bostrom developed a structured framework, categorizing these risks beyond a monolithic understanding.

2.1. Core Categories: Adversarial Hazards vs. Unintended Consequences

Bostrom's primary classification of information hazards divides them into two major categories based on the nature of the harm and the intent involved [2]:

- **Adversarial Hazard**: This category describes situations where specific true information is purposefully acquired and utilized by a "bad actor" or adversary to inflict harm upon others. This aligns with traditional notions of intelligence and security threats, where knowledge empowers malicious intent.
- **Unintended Consequence**: In contrast, this category encompasses scenarios where harm arises not from deliberate malicious intent, but as an unforeseen or indirect outcome of the information's dissemination. The harm may affect the individual who learns the information, or a broader population, without any purposeful malevolent action.

This distinction between adversarial and unintended harm is fundamental for developing effective mitigation strategies. Adversarial hazards, driven by malicious intent, typically call for countermeasures such as strict access control, robust encryption, counter-intelligence operations, and deterrence. However, unintended consequences, which emerge from complex interactions between information, human psychology, and societal systems, demand a more nuanced approach. For instance, a groundbreaking scientific discovery published with purely benevolent intentions could, once widely known, trigger unforeseen societal anxieties, economic disruptions, or even psychological distress in individuals. This highlights that managing information hazards is not solely about thwarting malicious actors but also about anticipating and managing the complex, non-linear ripple effects of knowledge dissemination. This framework thus expands the boundaries of traditional risk management beyond a simplistic "good actor versus bad actor" dichotomy,

mandating the inclusion of systemic risks that arise from the inherent properties of information itself and the unpredictable ways humans interact with it. It suggests that even well-intentioned actions, such as open scientific publication, can lead to significant harm if the broader informational context and potential for unintended consequences are not thoroughly understood and proactively addressed.

2.2. Detailed Sub-types of Information Hazards

Beyond the core bifurcation, Bostrom's typology offers more granular sub-types, providing a comprehensive understanding of the diverse mechanisms through which information can become hazardous.

- **Data Hazards**: These hazards involve specific, concrete pieces of data that, if disseminated, create a direct and identifiable risk.[2] Examples include highly sensitive information such as the precise genetic sequence of a lethal pathogen or the detailed blueprint for constructing a thermonuclear weapon.[2] Such information is often "resource-intensive to acquire" due to its complexity and the stringent security measures surrounding it, and if obtained, it "allows you to do really bad things".[8] Data hazards are directly pertinent to critical areas like biosecurity, where they could enable the recreation of biohazards, and national security, where they might facilitate nuclear proliferation.[2]
- **Idea Hazards**: These hazards stem from the dissemination of a general idea or conceptual breakthrough that, even without detailed specifications, can enable harm.[3] A classic example is the fundamental scientific concept that a nuclear fission reaction can be used to create a bomb. Merely knowing this general idea can be sufficient for a well-resourced team to independently develop a nuclear device, as it provides the core "inspiration, knowledge, and processes" needed to guide extensive scientific and engineering efforts towards a destructive outcome.[2] Another example is the idea of simple methods for screening undercover police officers, which does not require esoteric knowledge.[8] Unlike data hazards that require specific, detailed information, idea hazards leverage abstract principles that can be independently developed or reverse-engineered once the core concept is understood.
- **Knowing-Too-Much / Spoiler Hazards**: This category describes information that, if known, directly causes danger or harm to the individual who possesses that knowledge.[2] Historically, women in the 16th and 17th centuries who possessed knowledge of the occult or birth control methods faced a heightened

risk of being accused of witchcraft, illustrating the concept of "forbidden knowledge".[2] In a more contemporary context, "spoiler hazards" occur when learning the ending of a movie or a significant plot twist prematurely diminishes the enjoyment or experience of entertainment, as "many forms of entertainment depend on the marshalling of ignorance".[3] While not physically harmful, it constitutes a genuine form of disappointment and a loss of a unique experiential value.[3] More profoundly, philosophical thought experiments like "Dead Kid Currency" or "The Drowning Child" can "radically change my view on value and my potential in the world" for some individuals, leading to significant moral distress, guilt, or an overwhelming sense of responsibility that can be psychologically debilitating.[7] The harm in these cases is existential or psychological, not physical. Bostrom also refers to this subset as "spoiler hazards," and proposed broader terms include "knowledge hazards" or "direct information hazards".[3]

- **Attention Hazards**: This subtle hazard arises when merely drawing public or professional attention to certain potent or relevant ideas or data increases risk, even if the information itself is already generally known.[3] Adversaries typically face a vast search space when seeking methods to cause harm. By focusing discourse on a specific threat (e.g., emphasizing "viral attacks" over conventional explosives), researchers or media can inadvertently "signal to an adversary that viral weapons... constitute an especially promising domain in which to search for destructive applications," thereby guiding their malicious efforts.[3] This is particularly relevant in strategic communication, intelligence analysis, and security research, where public discussions can unintentionally provide strategic guidance to malicious actors.
- **Other Related Concepts**:
    - **Willful Blindness**: This refers to the deliberate avoidance of knowledge of facts, often to evade legal or ethical responsibility. An example is a company intentionally avoiding information on unsafe work practices to lessen liability in case of injury.[2] This represents an inverse information hazard, where the *absence* of knowledge, rather than its dissemination, leads to harm.
    - **Social Contagion / Harmful Trends**: This phenomenon occurs where knowledge of certain trends, particularly physically dangerous viral trends, leads to their replication and widespread adoption.[2] This highlights the self-propagating and imitative nature of some information-driven harms.
    - **Partial Information Hazards**: The danger in this scenario lies not in complete knowledge, but in incomplete, fragmented, or misleading information. Sometimes, a lack of full context can be more volatile or dangerous than complete transparency.[8]

- **Template Hazards, Signaling Hazards, Evocation Hazards**: These are additional categories within Bostrom's broader typology [4], indicating a comprehensive classification of how information can serve as a "rate-limiting step"—the critical missing piece—that enables bad actors to deploy scientific capabilities for harmful purposes.[4]

The detailed typology of information hazards, encompassing Data, Idea, Knowing-Too-Much, Attention, and related concepts like Willful Blindness and Social Contagion, reveals that information's capacity to cause harm is far from monolithic. It spans a wide spectrum, from explicit, actionable blueprints to abstract conceptual breakthroughs. The harm can be external and widespread (adversarial) or deeply internal and psychological (knowing-too-much). Furthermore, the inclusion of attention hazards highlights a meta-level risk where the *focus* on certain information, even if already known, can amplify danger. The concept of "willful blindness" demonstrates harm arising from *avoiding* information, while "social contagion" points to information's self-replicating harmful potential. The subtle point about "partial information" being dangerous suggests that incomplete knowledge can be more volatile than full transparency. This multifaceted nature implies that a single, generic approach to information control is insufficient. This comprehensive typology provides an indispensable analytical framework for identifying, categorizing, and understanding the diverse pathways through which information can become hazardous. It moves beyond a simplistic "secret versus public" dichotomy to a granular appreciation of how different forms, contexts, and dynamics of information dissemination can lead to harm. This detailed understanding is absolutely essential for developing targeted, effective, and ethically sound mitigation strategies that are tailored to the specific nature of the information hazard at hand.

3. Case Studies: 10 Deeply Explained Information Hazard Examples

The following ten examples provide concrete illustrations of information hazards as defined by Nick Bostrom. They are drawn from the available material and elaborated upon to demonstrate the nuances of each hazard type and its potential real-world implications. Each case study adheres to a consistent structure for clarity and analytical depth.

Example 1: Blueprints for a Thermonuclear Weapon

- **Type of Hazard**: Data Hazard (Adversarial)
- **Description of Information**: This refers to the highly detailed, specific technical specifications, schematics, and operational instructions necessary for the design and construction of a thermonuclear (hydrogen) weapon.[2] This information is inherently "resource-intensive to acquire" due to its complexity and the stringent security measures surrounding it.[8]
- **Mechanism of Harm**: The dissemination of such blueprints directly enables a state or a highly resourced non-state actor to bypass years of research and development, accelerating their path to acquiring a weapon of mass destruction. This information acts as a critical "rate-limiting step" [4], providing the precise knowledge needed to operationalize destructive capabilities, thereby empowering malicious actors to inflict catastrophic harm on a global scale.[2]
- **Real-World Implications**: The most severe implication is the acceleration of nuclear proliferation, increasing the likelihood of nuclear warfare, regional conflicts escalating to nuclear exchanges, or nuclear terrorism. This poses a direct global catastrophic risk, potentially leading to widespread death, environmental devastation (e.g., nuclear winter), and geopolitical instability that threatens human civilization itself. This is why such information is universally classified at the highest levels.

Example 2: Genetic Sequence of a Highly Lethal Pathogen

- **Type of Hazard**: Data Hazard (Adversarial)
- **Description of Information**: This involves the complete and accurate genetic (DNA or RNA) sequence of a naturally occurring or engineered pathogen characterized by high virulence, transmissibility, and lethality (e.g., a highly weaponizable virus or bacterium).[2] This includes information that could facilitate the recreation or enhancement of such biohazards.
- **Mechanism of Harm**: Access to this specific data could enable a malicious actor—ranging from a rogue state to a well-funded terrorist group or even a highly skilled individual—to synthesize, modify, or recreate the pathogen using increasingly accessible synthetic biology tools. This directly facilitates the

development and deployment of biological weapons.[2] The existence of such information is a core "dual-use concern" in biosecurity.[4]

- **Real-World Implications**: The primary implication is the increased risk of a synthetic pandemic or a deliberate bioweapon attack. Such an event could lead to massive fatalities globally, overwhelm healthcare systems, trigger widespread societal panic and breakdown, and cause severe economic devastation. It represents a significant global catastrophic risk, potentially on par with nuclear threats.

Example 3: The General Idea of Using Fission for a Bomb

- **Type of Hazard**: Idea Hazard (Adversarial)
- **Description of Information**: This refers not to specific blueprints, but to the fundamental scientific concept or general idea that a nuclear fission chain reaction can release immense amounts of energy, making it a theoretical basis for a weapon.[2] This is distinct from detailed engineering plans.
- **Mechanism of Harm**: While abstract, this core idea provides the conceptual breakthrough necessary for weapon development. A sufficiently resourced team, even without specific data, can leverage this general principle to conduct the necessary research and development to create a nuclear bomb.[2] It serves as the "missing inspiration, knowledge, and processes" [4] that, once understood, can guide extensive scientific and engineering efforts towards a destructive outcome.
- **Real-World Implications**: The widespread knowledge of this idea lowers the conceptual barrier to nuclear weapon development for any nation or entity with the scientific and industrial capacity. It contributes to the overall risk of nuclear proliferation by making the foundational scientific principle accessible, thereby increasing the probability of new actors pursuing and eventually acquiring nuclear capabilities, with similar catastrophic implications as Example 1.

Example 4: Knowledge of Flaws in Critical Infrastructure Design (Chernobyl Case)

- **Type of Hazard**: Knowing-Too-Much / Unintended Consequence Hazard (with elements of Adversarial if exploited)
- **Description of Information**: This refers to true, critical information about

inherent design flaws or operational vulnerabilities within a vital system, such as a nuclear reactor's safety mechanisms.[9] In the Chernobyl case, the Soviet government knew of reactor flaws, but operators did not.[9]

- **Mechanism of Harm**: If this information is *not* disseminated to the appropriate operational personnel, as tragically occurred at Chernobyl, it can lead to catastrophic accidents due to ignorance of risks, even if the intent is not malicious.[9] Conversely, if such information *were* widely disseminated without proper context or mitigation, it could cause public panic, loss of trust in institutions, or even be exploited by adversaries for sabotage. The harm is an unintended consequence of information existing but being improperly managed (either withheld or over-disclosed).
- **Real-World Implications**: As tragically demonstrated by Chernobyl, the implications include massive loss of life, widespread environmental contamination, long-term health crises, significant economic disruption, and a severe erosion of public trust in government and industry oversight.[9] This example highlights the complex ethical dilemma of balancing the public's "right to know" against potential dangers and the importance of responsible information flow within organizations.

Example 5: Specific Methods for Screening Undercover Police Officers

- **Type of Hazard**: Idea Hazard (Adversarial, with potential for Unintended Social Harm)
- **Description of Information**: This involves the idea or specific techniques for identifying and screening out undercover law enforcement agents, such as requiring proof of employment with a known, legitimate organization.[8] This is described as not requiring "esoteric knowledge" or "lots of resources".[8]
- **Mechanism of Harm**: The widespread dissemination of such an idea could significantly enhance the ability of criminal organizations, terrorist groups, or other illicit networks to identify and neutralize law enforcement or intelligence efforts. By making it easier to detect undercover operatives, it undermines investigative capabilities, facilitates illegal activities, and potentially endangers agents.[8] The hazard lies in the ease with which this idea can be adopted and its direct utility for those seeking to evade justice.
- **Real-World Implications**: This could lead to a substantial increase in organized crime activities, drug trafficking, and other illicit operations by making it harder for authorities to infiltrate and disrupt them. It compromises public safety and

security at a local or national level, potentially leading to increased violence and a breakdown of law and order in affected areas.

Example 6: Drawing Attention to a Specific Vulnerability or Attack Vector (Attention Hazard)

- **Type of Hazard**: Attention Hazard (Adversarial)
- **Description of Information**: This hazard arises not from new information, but from the act of publicly highlighting or focusing significant discourse on a particular type of threat, vulnerability, or attack methodology (e.g., emphasizing "viral attacks" as distinct from conventional explosives).[3] The underlying information may already be generally known or discoverable.
- **Mechanism of Harm**: Adversaries, facing a vast array of potential harmful avenues, conduct a "vast search task" to identify the most effective methods. By drawing disproportionate attention to a specific domain (e.g., bioweapons, a particular cyber vulnerability), public discourse or research can inadvertently "signal to an adversary that viral weapons... constitute an especially promising domain in which to search for destructive applications," effectively guiding their efforts and increasing the likelihood of an attack in that area.[3]
- **Real-World Implications**: This meta-level information hazard can subtly but significantly influence the strategic decisions of malicious actors. It can lead to a misallocation of defensive resources (if attention is drawn to a less probable but highly impactful threat) or, more dangerously, direct adversaries towards optimal targets or methods, thereby increasing the efficiency and success rate of specific types of attacks (e.g., targeted cyberattacks, focused bioweapon development).

Example 7: The "Spoiler Hazard" (Knowing the End of a Story)

- **Type of Hazard**: Knowing-Too-Much / Spoiler Hazard (Unintended Consequence, harm to knower)
- **Description of Information**: This refers to learning critical plot points, twists, or the ending of a narrative work (e.g., a movie, book, or video game) before one has had the opportunity to experience it firsthand.[3] The information is true and accurate.

- **Mechanism of Harm**: The harm here is primarily subjective and directly experienced by the knower. "Many forms of entertainment depend on the marshalling of ignorance".[3] Knowing the outcome prematurely diminishes the suspense, surprise, emotional impact, and overall enjoyment of the narrative experience. While not physically harmful, it constitutes a genuine form of disappointment and a loss of a unique experiential value.[3]
- **Real-World Implications**: While seemingly trivial compared to other hazards, this example powerfully illustrates the principle that true information can directly cause harm to the individual knower. It underpins common social norms around content warnings and responsible media consumption, and even informs individual mitigation strategies like "refrain[ing] from reading reviews and plot summaries".[8] It highlights that "harm" can extend beyond physical or economic damage to include psychological or experiential detriment.

Example 8: "Dead Kid Currency" and "The Drowning Child" Thought Experiment

- **Type of Hazard**: Knowing-Too-Much / Idea Hazard (Unintended Consequence, harm to knower's worldview/psychology)
- **Description of Information**: This refers to powerful philosophical thought experiments or ethical concepts, such as Peter Singer's "Drowning Child" argument (which posits a strong moral obligation to aid suffering at significant personal cost) or the more stark concept of "Dead Kid Currency" (implying a moral imperative to prevent suffering even if it means sacrificing personal comfort or aspirations).[7] These are true, logically coherent ideas.
- **Mechanism of Harm**: For individuals who deeply engage with and internalize such concepts, the knowledge can "radically change my view on value and my potential in the world".[7] This can lead to profound moral distress, overwhelming guilt, a crippling sense of responsibility, or a feeling of moral paralysis in a world filled with suffering. The "harm" is existential, psychological, and can significantly impact an individual's well-being, life choices, and mental health, even if it does not involve physical danger.
- **Real-World Implications**: While not a direct societal threat, the widespread dissemination and internalization of such demanding ethical frameworks can lead to burnout among altruistic individuals, significant personal psychological burdens, and potentially a sense of futility or despair. It underscores how abstract philosophical ideas, when deeply understood, can have profound and sometimes detrimental direct impacts on an individual's inner world and capacity for

flourishing.

Example 9: Detailed Information on How to Commit Financial Fraud (LLM Output)

- **Type of Hazard**: Data Hazard / Idea Hazard (Adversarial)
- **Description of Information**: This encompasses specific, actionable instructions, detailed methodologies, and step-by-step guides for executing complex financial fraud schemes (e.g., phishing techniques, investment scams, identity theft processes). Such information can be generated and disseminated by large language models (LLMs).[4]
- **Mechanism of Harm**: LLMs, by providing "true information [that] can be used to create harm to others, such as how to build a bomb or commit fraud" [4], democratize access to sophisticated malicious knowledge. This significantly lowers the barrier to entry for individuals or groups seeking to commit fraud, enabling a wider range of actors to engage in illicit financial activities without requiring prior specialized expertise or extensive research.
- **Real-World Implications**: The widespread availability of such information can lead to a substantial increase in financial crimes, resulting in significant monetary losses for individuals, businesses, and financial institutions. It erodes public trust in digital systems, online transactions, and financial security. This represents a pervasive and growing threat in the digital age, posing new challenges for cybersecurity, law enforcement, and regulatory bodies globally.

Example 10: Public Disclosure of AI System Capabilities in a Competitive Race

- **Type of Hazard**: Signaling Hazard / Attention Hazard / Data Hazard (Complex Adversarial/Unintended Interplay)
- **Description of Information**: This refers to the precise sharing of information about one's own advanced AI system capabilities, benchmarks, and progress, along with insights or guesses about rivals' achievements, within a highly competitive AI development environment.[10]
- **Mechanism of Harm**: In "highly decisive races" to develop powerful new technologies like advanced AI, public knowledge of capabilities can paradoxically be more dangerous than private information. This is because it can intensify

competitive pressure, leading developers to "cut corners on safety" in a desperate bid for victory, thereby increasing the overall risk of a "disaster" that affects all actors.[10] This dynamic can lead to a neglect of "proper oversight" and an increased likelihood of "misaligned AI objective functions (the control problem)" or the "use of TAI by actors wishing to impose harms on others (the political problem)".[10]

- **Real-World Implications**: This complex information hazard directly contributes to the existential risks associated with advanced AI. It increases the probability of catastrophic outcomes such as uncontrollable AI systems, AI misuse by rogue actors, or an AI arms race leading to global instability. It highlights a critical dilemma for AI governance: while transparency is generally desirable, in certain competitive contexts, it can exacerbate risks, necessitating careful strategic communication and international cooperation to prevent a race to the bottom on safety.

4. Analysis: Probability and Impact Assessment

This section details the methodology for assessing the probability and impact of each information hazard example and presents a ranked table, followed by a comprehensive rationale for each assessment.

4.1. Methodology for Qualitative Assessment of Probability and Impact

Given the qualitative nature of the information and the absence of precise quantitative data, the assessment of probability and impact for each information hazard example is conducted using a structured qualitative methodology. This approach aims to provide reasoned judgments based on the available information and an expert understanding of risk dynamics.

Probability Assessment: This metric evaluates the likelihood of the information hazard manifesting and leading to harm. Categories are defined as:

- **Low**: Highly unlikely to occur; significant barriers (e.g., extreme secrecy, immense resource requirements, very narrow applicability) exist.

- **Medium**: Plausible; some barriers exist, but the conditions for manifestation are reasonably met or could be overcome.
- **High**: Very likely to occur; few barriers, widespread accessibility, or inherent susceptibility to the hazard.

Factors considered include: ease of access and dissemination of the information; the number and type of actors capable of using it for harm; the likelihood of independent rediscovery (for idea hazards); existing safeguards, classification levels, and regulatory environments; and the "obviousness" or common knowledge status of the idea.[8]

Impact Assessment: This metric evaluates the potential scale and severity of the harm if the information hazard manifests. Categories are defined as:

- **Low**: Minimal harm, primarily affecting individuals or small groups, with easily reversible consequences (e.g., personal inconvenience, minor financial loss).
- **Medium**: Moderate harm, affecting a significant number of individuals or a localized community, with reversible but notable consequences (e.g., significant financial loss, localized disruption).
- **High**: Severe harm, affecting a large population or a national scale, with long-lasting or irreversible consequences (e.g., widespread illness, significant economic damage, major societal disruption).
- **Catastrophic**: Extreme, widespread, or existential harm, affecting global populations, with potentially irreversible and civilization-altering consequences (e.g., mass fatalities, societal collapse, existential threat to humanity).

Factors considered include: the scale of potential harm (individual, local, national, global); the severity and nature of the harm (psychological, financial, physical, environmental, existential); and the potential for cascading failures or secondary effects.

4.2. Table: Ranked Information Hazard Examples by Probability and Impact

The following table systematically presents the assessment for each of the ten examples, providing a clear and immediately digestible overview of their comparative risk profiles.

Example	Type of Hazard	Brief Description	Probability	Impact	Overall Risk Ranking
1. Blueprints for a Thermonuclear Weapon	Data Hazard (Adversarial)	Detailed technical specs for a hydrogen bomb.	Low-Medium	Catastrophic	Extreme
2. Genetic Sequence of a Highly Lethal Pathogen	Data Hazard (Adversarial)	Full genetic code of a weaponizable virus/bacterium.	Medium-High	Catastrophic	Extreme
3. The General Idea of Using Fission for a Bomb	Idea Hazard (Adversarial)	Conceptual understanding of nuclear fission for weapons.	High	Catastrophic	High-Extreme
4. Knowledge of Flaws in Critical Infrastructure Design	Knowing-Too-Much / Unintended Consequence	Undisclosed design flaws in vital systems (e.g., nuclear reactors).	Medium-High	High-Catastrophic	High-Extreme
5. Specific Methods for Screening Undercover Police Officers	Idea Hazard (Adversarial)	Techniques to identify covert law enforcement agents.	Medium-High	Medium	Medium-High
6. Drawing Attention to a Specific Vulnerability or Attack Vector	Attention Hazard (Adversarial)	Publicly highlighting a particular threat domain.	High	Medium-High	High
7. The "Spoiler Hazard"	Knowing-Too-Much / Spoiler Hazard	Learning critical plot points of a story	High	Low	Low

	(Unintended Consequence)	prematurely.			
8. "Dead Kid Currency" and "The Drowning Child" Thought Experiment	Knowing-Too-Much / Idea Hazard (Unintended Consequence)	Profound philosophical concepts inducing moral distress.	Medium	Low-Medium	Medium
9. Detailed Information on How to Commit Financial Fraud	Data Hazard / Idea Hazard (Adversarial)	Step-by-step guides for executing complex financial scams (e.g., via LLMs).	High	Medium-High	High
10. Public Disclosure of AI System Capabilities in a Competitive Race	Signaling Hazard / Attention Hazard / Data Hazard (Complex)	Sharing detailed progress of advanced AI systems in a competitive environment.	Medium	High-Catastrophic	High-Extreme

4.3. Detailed Rationale for Probability and Impact Ranking of Each Example

For each of the ten examples, a comprehensive justification is provided for its assigned Probability and Impact scores, linking back to the defined methodology and drawing upon the nuances identified in the available information.

- **Example 1: Blueprints for a Thermonuclear Weapon**
  - **Probability: Low-Medium**. While the information itself is highly sensitive and classified, and the risk of espionage or insider threats is ever-present, the resources, specialized expertise, and vast infrastructure required to successfully act upon these blueprints are immense. This significantly limits the number of potential bad actors to sovereign states or exceptionally well-resourced non-state actors, making widespread misuse less probable. International treaties and non-proliferation efforts also act as strong

deterrents and control mechanisms, although the risk of a state actor acquiring or developing such capabilities remains a persistent concern.
    - **Impact: Catastrophic**. The successful use of a thermonuclear weapon, whether in a limited regional conflict or a full-scale global exchange, would lead to an existential or global catastrophic risk. This includes unimaginable loss of life, widespread destruction of infrastructure, long-term environmental devastation (e.g., nuclear winter), and profound geopolitical instability, threatening the very fabric of human civilization.
- **Example 2: Genetic Sequence of a Highly Lethal Pathogen**
    - **Probability: Medium-High**. The accessibility of genetic sequences is increasing due to public databases and open science initiatives. Furthermore, advancements in synthetic biology tools (e.g., CRISPR) are making the process of synthesizing or modifying pathogens cheaper and more widespread, lowering the barrier to entry compared to nuclear weapons. While significant expertise is still required, the "dual-use concern" [4] is very real, and the potential for a wider range of actors to misuse this information is growing. The digital nature of genetic information also makes its unauthorized dissemination easier to achieve than physical blueprints.
    - **Impact: Catastrophic**. A deliberate release of a highly lethal and transmissible synthetic pathogen could lead to a global pandemic far more severe than natural ones. This could result in billions of deaths, the collapse of global healthcare systems, widespread societal breakdown, and severe economic devastation, posing an existential threat to humanity.
- **Example 3: The General Idea of Using Fission for a Bomb**
    - **Probability: High**. The fundamental scientific principle of nuclear fission and its energy release is widely known and taught in advanced physics and engineering curricula globally. It is not esoteric or classified information.[8] This widespread accessibility means the conceptual barrier to nuclear weapon development is inherently low for any scientifically literate and resourced entity. The idea itself is abstract and cannot be "unlearned" by society.
    - **Impact: Catastrophic**. While the idea itself is not a direct weapon, its universal knowledge means that the conceptual hurdle for nuclear weapon development has been permanently removed. Any sufficiently resourced and determined entity can pursue the necessary research and engineering to develop such a weapon. This increases the overall risk of proliferation and, consequently, the probability of a nuclear catastrophe, as the foundational scientific principle is universally accessible.
- **Example 4: Knowledge of Flaws in Critical Infrastructure Design (Chernobyl Case)**

    - **Probability: Medium-High**. Such design flaws or critical vulnerabilities often exist in complex systems and may be known to some individuals or departments but not adequately communicated or addressed across the organization, or to the public. The probability of such vital information *not* reaching the right operational personnel, or being ignored due to organizational inertia, secrecy, or "willful blindness" [2], is significant.[9] This is a recurring issue in large, bureaucratic, or security-sensitive organizations.
    - **Impact: High-Catastrophic**. As tragically demonstrated by the Chernobyl disaster, the manifestation of this hazard can lead to regional devastation, long-term environmental contamination, severe health consequences for large populations, and massive economic and social disruption. It can also profoundly erode public trust in government and industry, leading to long-term societal instability and a loss of confidence in critical systems.
- **Example 5: Specific Methods for Screening Undercover Police Officers**
    - **Probability: Medium-High**. The idea is not "esoteric knowledge" and does not require "lots of resources for research".[8] It is a relatively straightforward concept that could be independently rediscovered or disseminated within criminal and subversive networks. The ease of communication and lack of significant technical barriers make its spread and adoption quite probable, especially given the continuous cat-and-mouse game between law enforcement and criminal elements.
    - **Impact: Medium**. While not globally catastrophic, the widespread application of such methods can significantly undermine law enforcement and intelligence operations. This can lead to an increase in organized crime, drug trafficking, and other illicit activities, compromising public safety and security at a local or national level. The harm is substantial but typically localized or national in scope, affecting community safety and the effectiveness of justice systems.
- **Example 6: Drawing Attention to a Specific Vulnerability or Attack Vector (Attention Hazard)**
    - **Probability: High**. This type of information hazard occurs frequently in the context of security research, public awareness campaigns, and media reporting. Researchers, in their efforts to raise awareness about potential threats, can inadvertently highlight promising attack vectors. The dynamics of public discourse and media sensationalism often lead to a disproportionate focus on specific, high-impact threats, effectively signaling their potency to potential adversaries.[3]
    - **Impact: Medium-High**. This hazard can directly guide and optimize the efforts of malicious actors. By signaling that a particular domain (e.g., a

specific type of cyber vulnerability, a bio-agent, or a critical infrastructure weakness) is "especially promising," it increases the efficiency and likelihood of targeted attacks. The ultimate impact depends on the severity of the vulnerability or the potency of the attack vector highlighted, potentially leading to significant economic disruption, data breaches, or even physical harm.

- **Example 7: The "Spoiler Hazard" (Knowing the End of a Story)**
  - **Probability: High**. In the age of pervasive social media, instant communication, and readily available online content, spoilers are ubiquitous. The probability of encountering a spoiler for popular media is extremely high, often occurring unintentionally through casual conversation or online browsing.
  - **Impact: Low**. The harm is primarily subjective and non-physical, affecting the individual knower's enjoyment and emotional experience. While it can cause significant personal disappointment or frustration, it does not pose a physical, financial, societal, or existential threat. It serves as a clear, relatable example of how true information can directly cause personal harm, even if minor, by diminishing an expected experience.
- **Example 8: "Dead Kid Currency" and "The Drowning Child" Thought Experiment**
  - **Probability: Medium**. These philosophical concepts are widely discussed within academic ethics, effective altruism communities, and broader intellectual circles. The probability of an intellectually curious individual encountering and deeply engaging with these ideas is moderate, especially within certain professional or academic contexts where such ethical dilemmas are explored.
  - **Impact: Low-Medium**. The harm is primarily psychological, moral, or existential to the individual who internalizes these concepts. It can lead to profound moral distress, guilt, a sense of overwhelming responsibility, or even moral paralysis in the face of global suffering. While not a direct physical or societal threat, it can significantly impact an individual's mental well-being, life choices, and overall capacity for happiness and flourishing, potentially leading to burnout or despair.
- **Example 9: Detailed Information on How to Commit Financial Fraud (LLM Output)**
  - **Probability: High**. Large Language Models (LLMs) are widely accessible to the public, and their ability to generate detailed instructions for illicit activities, including fraud, is a known and ongoing challenge for AI safety and ethics.[4] Despite mitigation efforts by AI developers, the sheer volume of LLM

usage and the ingenuity of malicious prompts make the generation and dissemination of such information highly probable, effectively lowering the barrier to entry for potential fraudsters who lack prior specialized expertise.
    - **Impact: Medium-High**. Widespread access to sophisticated fraud methodologies can lead to a significant increase in financial crimes, resulting in substantial monetary losses for individuals, businesses, and financial institutions globally. It erodes trust in online systems, digital transactions, and the overall financial ecosystem, posing a pervasive and evolving threat that requires continuous vigilance and adaptation from law enforcement and cybersecurity professionals.
- **Example 10: Public Disclosure of AI System Capabilities in a Competitive Race**
    - **Probability: Medium**. In highly competitive technological races, there is often a tension between the desire for secrecy to maintain a competitive edge and the pressure for transparency (e.g., for funding, talent attraction, or signaling progress). The probability of such information being publicly disclosed depends on the specific competitive dynamics, perceived strategic advantages, and the ethical frameworks guiding the actors.[10] The inherent drive for rapid advancement often outweighs caution in such high-stakes environments.
    - **Impact: High-Catastrophic**. As highlighted by Bostrom, in "highly decisive races," public knowledge of capabilities can lead to actors "cut[ting] corners on safety" in pursuit of victory, significantly increasing the risk of a "disaster" affecting all.[10] This could manifest as the development of misaligned AI systems (the "control problem") or the deployment of powerful AI by actors with malicious intent (the "political problem"). The ultimate impact is directly tied to the scale and power of advanced AI, potentially leading to an existential catastrophe or irreversible global harms.

The process of ranking these diverse examples by probability and impact reveals a critical underlying pattern: the likelihood of an information hazard manifesting is often inversely correlated with the resources and specialized knowledge required to act upon it, while its impact is frequently directly proportional to the destructive potential of the underlying technology or idea. For instance, easily accessible information, such as spoilers or LLM-generated fraud methods, has a high probability of causing harm, albeit often with a lower individual impact. Conversely, highly classified, resource-intensive information, like nuclear blueprints, has a lower probability of widespread misuse but carries a disproportionately catastrophic impact. The severity of the impact is also heavily modulated by the broader societal and geopolitical

context; for example, a nuclear blueprint's impact is catastrophic due to existing global tensions and the inherent nature of the weapon. This implies that effective risk mitigation must be highly tailored: for high-impact, low-probability events, extreme secrecy, international cooperation, and robust deterrence are paramount. For high-probability, lower-impact events, public education, ethical guidelines, technological safeguards, and rapid response mechanisms are more relevant. This comprehensive analysis underscores that managing information hazards is not a monolithic problem but a complex, multi-dimensional challenge requiring a strategic and adaptive blend of technical, policy, ethical, and social interventions. It moves the discussion beyond simple "information control" to the cultivation of a responsible "knowledge ecosystem" that understands the delicate balance between the pursuit of knowledge, technological progress, and the imperative of safety and long-term human flourishing.

5. Broader Implications and Mitigation

The concept of information hazards extends far beyond theoretical discussions, holding significant relevance for contemporary challenges, particularly those posed by emerging technologies and their potential to contribute to existential risks.

5.1. Information Hazards in the Context of Emerging Technologies and Existential Risk

The rapid advancement of certain technologies introduces novel and amplified forms of information hazards, demanding proactive consideration.

- **Artificial Intelligence (AI)**: Large Language Models (LLMs) are identified as direct sources of information hazards, capable of generating instructions for harmful activities such as bomb-making or financial fraud.[4] This capability democratizes access to dangerous knowledge, making it available to a wider array of individuals who might not otherwise possess the expertise to cause harm. The competitive dynamics within AI development races further complicate this landscape, introducing complex information hazards where public knowledge of capabilities can paradoxically increase risk by incentivizing developers to "cut

corners on safety" in pursuit of victory.[10] This suggests that AI not only creates novel types of information hazards by making harmful knowledge generation and dissemination highly scalable and accessible but also amplifies existing ones by accelerating the development of dangerous capabilities or by making it easier for adversaries to identify optimal attack vectors. Bostrom's broader work on AI risks includes the "control problem" (ensuring AI objectives align with human values) and the "political problem" (preventing powerful AI from being used by malicious actors).[10] Information hazards are intricately linked to both, as they can exacerbate misalignment or facilitate misuse. The competitive pressure in AI development implies that the usual checks and balances for safety might be bypassed, creating a volatile information environment. This points to a self-reinforcing loop where rapid advancements in AI capabilities, once known or inferred, can drive further risky development, making the information landscape increasingly perilous. AI development thus presents an unprecedented and urgent challenge for information governance. The rapid pace of AI progress, combined with its inherent dual-use nature and the competitive pressures, means that the window for identifying and mitigating new information hazards is shrinking. This calls for a proactive and deeply integrated approach to "responsible AI development" that incorporates information hazard considerations from the very earliest stages of research and deployment, rather than as an afterthought or reactive measure.

- **Biosecurity and Synthetic Biology**: The increasing availability of detailed genetic sequences of diseases or the chemical makeup of toxins poses significant adversarial hazards, as this information can be used to recreate or modify biohazards.[2] Certain forms of biological weapons research are already prohibited by international conventions, such as the Biological Weapons Convention, precisely because of the extreme information hazards they pose.[4] The concept of "dual-use concern" is central to biosecurity, highlighting that technologies and information developed for benevolent scientific or medical purposes can be deliberately misused for nefarious ends.[4] The danger of specific biological information, such as DNA sequences, is underscored by the continuous miniaturization, cost reduction, and increased accessibility of synthetic biology tools and genetic sequencing technologies. This implies that biological information hazards are becoming less "resource-intensive" for a broader range of actors, effectively democratizing the potential for biological harm. Unlike physical weapons, where materials are scarce and controlled, biological information can be replicated, shared, and disseminated digitally with relative ease, making traditional control mechanisms far more challenging. Biosecurity strategies must therefore urgently evolve beyond traditional physical containment

measures to include robust "information containment" and responsible data sharing protocols for biological research. This represents a global challenge in balancing the principles of open science for accelerating progress in medicine and biotechnology with the imperative to prevent deliberate misuse. It necessitates the development of new international norms, ethical guidelines, and potentially regulatory frameworks specifically for the management of biological information.

- **Existential Risk**: Information hazards are a core area of study for organizations like the Future of Humanity Institute (FHI), which explicitly focused on "global catastrophic risk, and in particular existential risk".[5] Bostrom's work on concepts like the "vulnerable world hypothesis" [6] is implicitly linked, suggesting that certain discoveries could fundamentally alter the world's vulnerability to catastrophic events. Bostrom's foundational work and the institutional context of FHI firmly situate information hazards within the broader framework of existential risk. This implies that information hazards are not merely isolated incidents of harm or security concerns, but can function as "rate-limiting steps" [4] that unlock pathways to catastrophic or even existential outcomes for humanity. The dissemination of certain critical information could be the trigger that pushes humanity past a "Great Filter" [5] or makes the "vulnerable world hypothesis" [6] a stark reality. This elevates the discussion of information hazards from a conventional risk management problem to a fundamental challenge for the long-term survival and flourishing of human civilization. Understanding, anticipating, and mitigating information hazards is not just about preventing immediate harm but about safeguarding humanity's long-term potential and ensuring its continued existence. This requires a proactive, foresight-driven approach to scientific discovery and technological development, where the potential for informational externalities—even from true and beneficial knowledge—is rigorously assessed alongside intended benefits. It calls for a deep commitment to responsible innovation.

5.2. Ethical Dilemmas: Balancing Transparency, Freedom of Information, and Risk

The very concept of information hazards directly challenges the widely held principle of freedom of information, asserting that some true information may be too dangerous for unrestricted dissemination.[2] This raises profound moral and policy questions regarding "who gets to decide what information should be kept secret" and

the extent of the public's "right to know" information, even if that knowledge could be dangerous.[9]

A critical tension arises: while restricting information may prevent harm, "hiding information from others even potential infohazards also risks hurting trust if people come to feel that they're being misled or kept in the dark".[9] This highlights the delicate balance between security and public trust. The fundamental ethical tension between the societal value of transparency and freedom of information and the imperative to prevent harm stemming from dangerous knowledge is a recurring theme. The explicit warning that "hiding information... risks hurting trust" reveals a complex trade-off: any decision to restrict information, or even to avoid discovering it, must be carefully weighed against the potential for eroding public trust, fostering suspicion, and potentially leading to unintended negative consequences, as seen in the Chernobyl example where a lack of information contributed to catastrophe. This implies that there is no simple, universally applicable solution, but rather a continuous ethical negotiation that requires balancing competing values. This ethical dilemma suggests that information policy in the age of information hazards cannot be purely utilitarian (focused solely on minimizing harm) but must also integrate deontological principles, such as rights, trust, and autonomy. It necessitates robust public discourse, transparent decision-making processes, and democratic oversight to navigate these complex trade-offs responsibly. It also implies a need for clear ethical guidelines for researchers, policymakers, and media professionals to manage information responsibly without unduly sacrificing fundamental societal values.

5.3. Strategies for Mitigating Information Hazards

Mitigating information hazards requires a multi-faceted approach, acknowledging the diverse nature of these risks.

- **Responsible Disclosure / Selective Dissemination**: The principle that "risky information needs to always be kept secret from everyone" is not absolute. Instead, access can be carefully limited to those with a "need to know," allowing for controlled dissemination while preventing widespread harm.[9] This involves careful gatekeeping and access control mechanisms, often seen in classified government information or proprietary corporate data.
- **Research Prioritization / Avoidance**: In certain high-risk areas, the information hazards can be so profound that the research itself "should not be conducted,"

as exemplified by prohibitions on specific forms of biological weapons research under international conventions.[4] This also includes a broader strategy of simply choosing to "invest less in discovering and disseminating certain kinds of information".[8] This proactive approach aims to prevent the creation of new hazards.

- **Encryption and Secrecy**: Traditional information security measures remain vital for protecting highly sensitive data hazards, involving robust encryption and strict secrecy protocols.[8] These technical measures serve as a primary line of defense against unauthorized access and dissemination.
- **"Info-lifeguards"**: While not explicitly detailed in the provided materials, the concept of "info-lifeguards" [7] implicitly suggests the need for individuals or entities specifically tasked with identifying, assessing, and managing dangerous information flows. These could be specialized roles within organizations or dedicated interdisciplinary teams.
- **Willful Blindness (Strategic Avoidance for Individuals)**: For certain types of hazards, particularly "spoiler hazards," individuals can employ personal mitigation strategies by consciously choosing to "refrain from reading reviews and plot summaries".[8] This highlights individual agency in managing personal information exposure and recognizing when ignorance can be beneficial.
- **Focus on General Ideas vs. Specific Data**: For certain discussions, limiting public conversation to "general ideas" rather than delving into detailed specifications can reduce the risk of enabling malicious actors.[9] This applies particularly to idea hazards, where the abstract concept is widely known, but detailed implementation knowledge remains restricted.
- **Addressing Partial Information Hazards**: Recognizing that danger can sometimes lie in incomplete or fragmented information, mitigation might involve providing more complete context or reframing information to reduce its potential for misuse or misunderstanding.[8] This moves beyond simple suppression to thoughtful contextualization.
- **Ethical Guidelines and Norms**: The importance of "institutionalizing ethics in AI through broader impact requirements" [6] and developing similar ethical frameworks across other sensitive scientific and technological domains is crucial for guiding responsible conduct and decision-making. These guidelines foster a culture of responsibility among researchers and developers.

The diverse range of mitigation strategies demonstrates that managing information hazards is far more complex than simple censorship or suppression. It encompasses proactive measures like avoiding certain research paths, strategic communication, and even individual responsibility. The emphasis on "partial information" being

dangerous is particularly illuminating, suggesting that sometimes *more* complete or contextualized information, rather than less, might be the solution to a hazard. This indicates that effective mitigation is not merely about blocking information but about cultivating a responsible and resilient "knowledge ecosystem" that understands the nuanced interplay of information, intent, and outcome. Effective mitigation of information hazards demands a holistic, adaptive, and context-dependent approach. It is not about a blanket policy of secrecy but about cultivating a responsible "knowledge ecology" that includes robust ethical frameworks, sophisticated foresight mechanisms, and a shared understanding across society of the delicate balance between the pursuit of knowledge, technological progress, and the imperative of safety. This implies a significant paradigm shift in how society manages scientific discovery and technological innovation, moving towards a more anticipatory and ethically informed model.

6. Conclusion: Navigating the Knowledge Landscape

Information hazards, as formalized by Nick Bostrom, represent a subtle yet profound category of risks arising from the dissemination of true information. Their increasing relevance in an era characterized by rapid technological advancement, particularly in fields like artificial intelligence and synthetic biology, underscores the urgent need for a nuanced understanding and proactive management of knowledge.

This report has highlighted the utility of Bostrom's typology in categorizing diverse information risks, demonstrating how true information can lead to harm through various mechanisms, from enabling malicious actors with specific data or ideas to causing unintended psychological distress or guiding adversarial attention. The analysis of ten distinct examples, ranked by their probability and impact, revealed a complex interplay between the accessibility of information, the resources required to act upon it, and the potential scale of harm. This underscores that risk mitigation strategies must be highly tailored to the specific nature of the information hazard. Furthermore, the inherent ethical dilemmas involved in balancing the principle of freedom of information with the imperative to prevent harm present a continuous challenge, demanding careful navigation of the trust-risk trade-off in information governance.

The ongoing and evolving challenge of managing dangerous knowledge necessitates

continued interdisciplinary research, fostering collaboration among philosophers, scientists, policymakers, and ethicists. There is a critical need for developing robust governance frameworks, establishing responsible innovation principles, and cultivating societal norms that acknowledge knowledge's dual potential. As humanity continues to expand its understanding of the world and develop increasingly powerful technologies, it bears a collective responsibility to wield this knowledge wisely, acknowledging its capacity for both immense good and catastrophic harm. This responsible stewardship of information is paramount to shaping a safer and more prosperous future for all.

Works cited

1. en.wikipedia.org, accessed June 26, 2025, https://en.wikipedia.org/wiki/Information_hazard#:~:text=An%20information%20hazard%2C%20infohazard%2C%20or,philosopher%20Nick%20Bostrom%20in%202011.
2. Information hazard - Wikipedia, accessed June 26, 2025, https://en.wikipedia.org/wiki/Information_hazard
3. What are information hazards? — EA Forum, accessed June 26, 2025, https://forum.effectivealtruism.org/posts/Nc5EjccDTfmcrG93j/what-are-information-hazards
4. Information Hazards: A Typology of Potential Harms from Knowledge - ResearchGate, accessed June 26, 2025, https://www.researchgate.net/publication/266404727_Information_Hazards_A_Typology_of_Potential_Harms_from_Knowledge
5. Future of Humanity Institute - Wikipedia, accessed June 26, 2025, https://en.wikipedia.org/wiki/Future_of_Humanity_Institute
6. Future of Humanity Institute, accessed June 26, 2025, https://www.futureofhumanityinstitute.org/
7. Info Lifeguards — EA Forum, accessed June 26, 2025, https://forum.effectivealtruism.org/posts/kzGKvyeyqBCoiCiMr/info-lifeguards
8. Are there any good writings on information hazards? : r/slatestarcodex - Reddit, accessed June 26, 2025, https://www.reddit.com/r/slatestarcodex/comments/s4tsd8/are_there_any_good_writings_on_information_hazards/
9. Can knowledge hurt you? The danger of infohazards (and exfohazards) - YouTube, accessed June 26, 2025, https://www.youtube.com/watch?v=sfgcg2bW8TI
10. Information Hazards in Races for Advanced Artificial Intelligence - Centre for the Governance of AI, accessed June 26, 2025, https://cdn.governance.ai/Information_Hazards_in_AI_Races__current_ver_.pdf