Can We Still Know?

How Generative AI Undermines the Epistemic Role of Justification

By
Jim Dalton

Submitted to Central European University - Private University

Undergraduate Studies Department

In partial fulfillment of the requirements for the degree of Bachelor of Arts in Philosophy, Politics and Economics

Supervisor: Matteo Pascucci

Vienna, Austria
2025

AUTHOR'S DECLARATION

I, the undersigned, Jim Dalton, candidate for the BA degree in Philosophy, Politics and Economics declare herewith that the present thesis titled "*Can We Still Know?: How Generative Al Undermines the Epistemic Role of Justification*" is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright.

I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 26 May 2025

Jim Dalton

ABSTRACT

This thesis investigates how generative artificial intelligence undermines the epistemic role of justification. The methodology consists of analytic investigation and engagement with contemporary epistemological literature. The thesis introduces a conceptual framework consisting of three key elements (justifiers, justifications, and beliefs) that clarifies how justification works with reference to empirical claims and is compatible with existing epistemological theories. Central to the thesis is the observation that generative AI can produce multimodal content (e.g. images, texts, audio, and video) that is perceptually and structurally indistinguishable from genuine human-generated content. This indistinguishability disrupts our ability to assess whether a justification is based on an appropriate justifier, thereby compromising two core epistemic dimensions: subjective belief formation and intersubjective evaluation of justification. The thesis further argues that traditional theories of epistemic justification (e.g. foundationalism, coherentism, evidentialism, and reliabilism) are mostly ineffective at addressing these challenges in practice. As a result, our capacity to function as responsible epistemic agents collapses, giving rise to a triadic form of skepticism: about reality, others, and ourselves. The culmination of these disruptions leads to what the thesis defines as the "post-knowledge era", a novel epistemic condition in which even responsible knowers can no longer confidently determine whether they truly know something. The implications are urgent, raising significant philosophical and societal concerns.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Matteo Pascucci, for his insightful feedback, unwavering support, and, most importantly, his kindness and enthusiasm throughout the course of this thesis, which would have likely never seen the light of day without him.

I would also like to thank my parents for their constant love, patience, and encouragement, for always believing in me, and simply for being the wonderful people they are. I am and will forever be grateful to have you as my parents. *Maman, Papa, je vous aime.*

TABLE OF CONTENTS

I. INTRODUCTION

The Modern Pandora's Box

Pandora opened the box, and the world was never the same. Some blamed her for unleashing suffering, others saw it as a parable of disobedience. But at its heart, the myth is about curiosity, about the irresistible pull toward the unknown, and the irreversible consequences of crossing certain thresholds. Curiosity is what drives us to explore, to learn and to question the world as we know it; it is the driving force behind human progress, discovery and growth. But, as the Greek myth warns us, there are moments in history when human curiosity surpasses caution, opening boxes that should have remained closed. Today, the rapid development and spread of generative artificial intelligence feels like one of those moments. In our pursuit of progress, we have built systems capable of generating synthetic texts, images, voices, and videos with astonishing realism. These technologies are now deeply embedded in our social, cultural, and informational ecosystems, powerful tools that blur the line between the real and the artificial. While the promises of generative AI are numerous and compelling, the dangers it introduces are subtle, complex, and increasingly urgent. Indeed, like the evils released from Pandora's box, its societal impacts are already unfolding: creative work is being replaced by synthetic imitations, jobs across industries are at risk and slowly vanishing, and trust in media and public discourse is eroding. But among these many dangers, one stands out for its quiet, pervasive impact: the threat generative AI poses to our capacity to know. More precisely, it threatens the very mechanisms by which we form, justify, and evaluate beliefs, destabilizing the foundational processes of knowledge acquisition and ultimately challenging our ability to trust what we see, hear, and believe. This thesis explores these issues, arguing that generative AI undermines the epistemic role of justification. Understanding this danger of generative AI is crucial, because when our ability to know collapses, so does our capacity to

act, decide, and live together in a shared reality. In a world where nothing can be confidently known, trust erodes, dialogue breaks down, and the very foundations of democracy and collective life begin to crumble.

I begin by reviewing the epistemic role of justification, examining key existing theories and proposing a threefold conceptual framework (comprising justifiers, justification, and belief) that aims to clarify how justification functions. Following this, I provide a concise overview of generative AI technologies, focusing on their capacity to produce synthetic content that is indistinguishable from authentic one. I then provide my argument as to how generative AI undermines the epistemic role of justification. At the core of this disruption lies the indistinguishability between AI-generated and genuine justifiers: AI content can so convincingly mimic the perceptual and structural features of authentic justifiers that agents cannot reliably differentiate between them. This results in inappropriate justifiers being accepted as appropriate, undermining the basis on which beliefs are formed and evaluated. At the subjective level, individuals may perceive inappropriate justifications as unusually strong because the justifiers supporting them closely match their cognitive and perceptual expectations. At the intersubjective level, the traditional role of justification as a normative filter breaks down: even when faced with seemingly credible justifications, we can no longer determine their appropriateness. As a result, the distinction between justified and unjustified beliefs becomes blurred, making it increasingly challenging to separate falsehood from truth within our epistemic practices. Furthermore, I demonstrate how this indistinguishability challenges epistemological theories such as foundationalism, coherentism, evidentialism, and reliabilism, revealing their limitations in analyzing the notion of justification in the context of generative AI. In my final section, I explore the broader consequences of this disruption for epistemic agency and the rise of skepticism, before concluding by proposing the concept of the 'post-knowledge era' to capture the profound epistemic crisis we face: a structural breakdown

not only in our ability to identify knowledge but in the very foundations that make knowledge possible.

II. FOUNDATIONS OF KNOWLEDGE

Belief, Truth and Justification

Every day, without thinking much about it, we claim to know countless things: that the Earth revolves around the Sun, that the second World War ended in 1945, that Beethoven wrote Moonlight Sonata, or even that you are reading this sentence right now. Yet, behind these seemingly simple claims lies a deep question: what does it actually mean to know something? Philosophers have tried to answer this for thousands of years, coming up with various theories of knowledge. One of the most prominent theories, which arguably emerged with Plato is the theory of "justified true belief" (JTB). This theory argues that a person *S* knows *p* if and only if:

1) *p* is true
2) *S* believes that *p*
3) *S* is justified in believing *p*

In other words, according to this theory, knowing something means having a justified true belief about the thing in question (Ichikawa and Steup 2024). For example, Peter knows that there is an apple on the table in front of him if and only if:

1) It is true that there is an apple on the table in front of him
2) Peter believes that there is an apple on the table in front of him
3) Peter is justified in believing that there is an apple on the table in front of him

The validity of the first clause of the JTB model seems indisputable, as it cannot be argued that one knows something that is false. Indeed, in our example, it would be incorrect to assert that Peter knows there is an apple on the table in front of him if there were no apple at all. Similarly, it would be incorrect to assert that someone knows Paris is the capital of Japan, as this is not the case. The second clause seems equally undeniable: Peter would be unable to conclude that there is an apple on the table in front of him if he instead believed it to be a pineapple. The

third clause, justification, is also necessary. Without justification, a true belief could still be mere luck. For example, suppose Peter randomly guesses that there is an apple on the table without even looking at the table, and by coincidence, it turns out to be true. In such a case, although Peter's belief is true, he has not arrived at it through any reasoning or evidence. We would therefore argue that Peter does not genuinely know that there is an apple on the table in front of him. On the other hand, if Peter had seen the apple placed there earlier, or recognized its distinctive shape and color, these would constitute relevant reasons, which would constitute a form of justification for his belief. Without such justification, his belief, though true, would not count as knowledge.

The JTB theory appears to offer a compelling account of what it means to know something. However, in 1963, American philosopher Edmund Gettier raised a crucial objection to this view, arguing that a justified true belief can sometimes be insufficient to constitute knowledge. Gettier's argument is rooted in the analysis of knowledge as a propositional attitude, namely the idea that an agent knows (or does not know) the thought (proposition) expressed by a certain sentence when uttered in a specific context. To demonstrate his objection, Gettier presented two scenarios in which an individual has a justified true belief that does not amount to knowledge. One of these scenarios involves two men, Smith and Jones, at a job interview. While being interviewed, Smith is informed by the recruiter that Jones will be offered the job. Smith also noticed earlier that Jones had ten coins in his pocket. Based on these two pieces of information, he infers that "the person who will get the job has ten coins in his pockets". However, it is Smith who actually ends up getting the job. Furthermore, by coincidence, Smith also appears to have ten coins in his pockets, meaning that his conclusion that "the person who will get the job has ten coins in his pockets" turns out to be true. According to the JTB theory, since all three requirements for knowledge ascription are met in this scenario, one would have to say that Smith knows that "the person who will get the job has ten coins in

his pockets". Yet, this conclusion seems to go against our intuition, as we recognize that Smith's belief is only true by accident (he thought Jones was the one who would get the job), meaning he does not genuinely "know" the proposition expressed by the sentence "the person who will get the job has ten coins in his pockets" (Gettier 1963). Applying this problem to our earlier example with Peter and the apple, a similar case can be imagined. Suppose Peter looks at a table and believes there is an apple in front of him, based on his sensory perception which provides the justification for his belief. However, it turns out that what Peter is seeing is actually a cleverly disguised mango, painted and shaped to look like an apple. Nevertheless, a real apple is concealed behind the mango, but out of Peter's view. In this case, Peter's belief that there is an apple on the table in front of him is true, and he is justified in believing so based on his sensory evidence, yet it would seem wrong to say he genuinely "knows" it, as his belief is true only by coincidence and not because he correctly identified what he saw.

These Gettier problems appeared to demonstrate that, in specific instances, a justified true belief does not equate to knowledge, primarily due to the fact that the justification on which the individual formed their belief was partially flawed. Indeed, in the examples discussed, both Smith and Peter's justification for forming their belief is defective. Smith was lied to by the recruiter and Peter was visually tricked by a fake apple. Although the final belief in each case turns out to be true, it is supported by a justification that includes a false component (e.g. the assumption that Jones will get the job), thereby illustrating how justification can sometimes be fallible even when the belief it supports is true. It is important to note that Gettier's examples highlight how justification can fail primarily in regard to empirical claims. In the case of formal claims, which are prevalent in mathematics and logic, justification, in the sense of the availability of a proof, is robust and not subject to such flaws; rendering the JTB model still valid and reliable in those domains. For instance, if someone has a true belief, namely that $2 + 2 = 4$, together with a rigorous proof of it, then their belief amounts to

knowledge. There is no chance of the belief being accidentally true in the same way it could happen with empirical claims, as the justification here is obtained through deductive reasoning from the definition of the concepts involved. This distinction between formal and empirical claims is reflected in A.J. Ayer's principle of verificationism, which holds that a statement is meaningful only if it can be empirically verified or is analytically true (Ayer 1936). In this sense, mathematical and logical truths are secure, because they are verifiable through pure reasoning alone. By contrast, empirical claims require justification that relies on sensory input or observation, both of which can be flawed, as demonstrated by Gettier.

Even though justification in empirical settings can sometimes be fallible, it still plays a crucial role in how we form and evaluate knowledge, especially in everyday life. Appropriate justification is a filter between naive beliefs and robust beliefs. When someone says, "I know that the sun is shining today because I saw it myself" or "I know that Gandhi was Indian because I read it from a credible source," they invoke a certain type of justification as the key factor that separates fact from opinion. Without a way to determine and rely on appropriate forms of justification, we would be forced to treat all beliefs, regardless of how well-justified they are, as if they carried the same weight, even though they do not. Indeed, suppose that someone holds the belief that the Earth is flat. In a world where all justifications are equivalent, their belief would carry the same weight as the widely accepted belief that the Earth is spherical, even though their belief is clearly false, or at the very least based on a weaker form of justification. It is appropriate justification that enables us to distinguish the two contradicting beliefs: photographs of the Earth from space, the way gravity works, and the scientific understanding of the Earth's shape all serve as evidence to prove that the flat Earth belief is false. Without appropriate justification, there would be no way to differentiate between a well-supported belief and a false one, leaving all beliefs, regardless of their truth or falsity, treated equally. Furthermore, appropriate justification also plays a central role in belief formation.

People often adopt new beliefs based on what they perceive as credible justification, ranging from direct experience, such as seeing something with one's own eyes, to more formal sources like news reports or expert testimony. For instance, an individual might initially doubt that a certain medication is effective. But after reading about several clinical trials published in a reputable medical journal, hearing a physician explain the scientific evidence behind it, or even witnessing a close friend recover after using it with their own eyes, they may form the new belief that the treatment does, in fact, work. Even our most fundamental beliefs, such as the notion that touching fire causes pain or that our parents care for us, are rooted in some form of appropriate justification, whether through direct experience, repeated observation or trusted testimony.

It seems clear that, despite the fallibility of justification in certain empirical contexts, as highlighted by Gettier cases, justification remains a pillar of our understanding of knowledge. It serves an essential evaluative function, helping us distinguish between beliefs that are credible, appropriate, or reasonable to hold and those that are not. In other words, justification is what gives a belief its "right standing" in relation to knowledge (Watson, n.d.). Furthermore, as argued above, justification plays also a pivotal role in the formation of our beliefs. Due to all of this, philosophers have thoroughly investigated what kinds of justification can be considered adequate for grounding knowledge. This resulted in the emergence of two major approaches to justification: internalism and externalism.

Before discussing these, I propose a conceptual framework that clarifies how justification works with reference to empirical claims. This framework is compatible with any of the rival theories and consists of three ingredients: justifiers, justifications, and beliefs. At the foundation are justifiers, the raw experiential or sensory inputs upon which a justification is built. Justifiers can take many forms: perceptual experience, empirical data, memory,

testimony, or other accessible forms of evidence. They are the immediate sources, data provided by our interaction with the world on which we draw in the formation of justifications. What unites them is their grounding in our sensory experience of reality, representing the raw cognitive or sensory materials available to the subject.[1] Justification, then, is the structured, often inferential process that organizes and interprets these justifiers to support a particular belief. It is not merely the presence of evidence, but the way in which that evidence is cognitively processed, integrated, and articulated as a reason for holding a belief. This process of justification can be understood differently depending on one's theoretical stance, such as internalism or externalism. Finally, the belief is the propositional content that the agent holds to be true on the basis of formed justifications. To illustrate this framework, consider the example that "I hold the belief that the coffee shop on the corner is closed because I saw a sign on the door that said closed for maintenance." My belief is the propositional content that I hold to be true (that the coffee shop is closed). The justification is the thought that I saw a sign on the door that said closed for maintenance. Finally, the justifier consists of the actual sensorial experience of seeing the sign on the door that said closed for maintenance.

The epistemic quality of justification depends on the appropriateness of the justifiers involved. An inappropriate justifier might consist of misleading or defective sensory input (e.g. hallucinations, illusions, misheard statements, or false testimony). These inappropriate justifiers lead to inappropriate justification. In such cases, a person might hold a belief that is in fact true, but for the wrong reasons, or worse, a belief that is false altogether. For example, if someone believes it is raining because they hear water sounds outside their window, but in

[1] The term *justifier* is used in epistemology to refer to the basis upon which a belief is justified, often a state, experience, or belief that supports another belief. As Pappas in the *Stanford Encyclopedia of Philosophy* explains, "The justifiers for a given justified belief are those items that make up or constitute the person's justification for that belief at that time. They are those items, whether experiences, states of affairs, or other beliefs, on which the person's current justification is based" (Pappas 2023). In this thesis, I build on this notion by framing justifiers as the raw experiential or informational inputs that serve as the immediate basis for forming justifications. This approach emphasizes their empirical character while remaining compatible with both internalist and externalist theories of justification. Unlike accounts that treat justifiers as only internally accessible to the agent, my model allows justifiers to function across epistemological frameworks, including those where the justificatory force may stem from external or process-based features.

fact the sound comes from a leaking pipe, their belief may coincidentally be true, but it is based on an inappropriate justifier. By contrast, an appropriate justifier is based on accurate, trustworthy sensory experience or evidence. When our justification is built upon appropriate justifiers, it will count as appropriate itself, and the belief it supports will be most likely true. In this way, justification retains its normative role: although it does not guarantee the truth of an empirical claim, it is structured to aim at it. In sum, the conceptual framework can be represented as follows:

- Appropriate justifier → Appropriate justification → Most likely true belief[2]
- Inappropriate justifier → Inappropriate justification → Possibly false belief or true belief held for the wrong reasons

This framework also enables us to highlight a crucial distinction between the strength of a justification and its appropriateness. In a subjective perspective, the strength of a justification is determined by how likely the agent takes it to support the truth of a belief. More precisely, when presented with two justifications "J1" and "J2" for a same belief P, J1 will be a stronger justification than J2 in support of the belief P for a subject S if, according to S, the likelihood of P being true on the basis of J1 is higher than on the basis of J2. While a justification will either be strong or weak based on the subjective judgment of an agent, the appropriateness of the justification can be assessed in a non-subjective manner. More precisely, strength reflects how convincing a justification seems to an agent, while appropriateness refers to whether the justification meets intersubjective[3] or objective epistemic standards, such as being grounded in accurate justifiers. Therefore, strength belongs to the domain of personal evaluation, while

[2] It is important to note that an appropriate justifier is necessary for the justification to be appropriate. However, an appropriate justifier alone may not be sufficient to constitute an appropriate justification as the full appropriateness of the justification will depend on additional elements which vary on based on the theory of justification one subscribes to. Similarly, an inappropriate justifier is not necessary for one to hold an inappropriate justification, as one can have an appropriate justifier and still hold an inappropriate justification. However, an inappropriate justifier will always lead to an inappropriate justification.

[3] Intersubjectivity refers to the shared, mutual accessibility of experience and understanding between individuals, which underpins our ability to evaluate others' beliefs, justifications, and claims to knowledge (Zahavi 2005, 130)

appropriateness belongs to the normative standards of intersubjective epistemic practices. In that sense, an agent may regard J1 as stronger than J2 in supporting a belief, even if J1 is in fact an inappropriate justification and J2 an appropriate one. More simply put, in some cases, a justification perceived as strong by the agent may in fact be inappropriate, while a justification they view as weaker could be appropriate. Thus our framework stresses that subjective judgments of strength can be misaligned with epistemic adequacy, and sheds light on why individuals may hold beliefs that seem well supported but rest on flawed justifiers. It allows us to ask not only whether a belief is held, but whether it is held for the right reasons, based on appropriate justifiers and appropriate justification. This naturally leads to a further question: what counts as an appropriate justification? This is precisely where theories of epistemic justification, most notably internalism and externalism, begin to diverge. Both aim to explain what it takes for a belief to be justified; they differ in what they take to be the relevant conditions.

Internalist theories of justification hold that "what makes for the justification of a belief is necessarily internal to the believer" (Swinburne 2001, 9). In other words, the factors which justify a belief must be internally accessible to the believer, that is, a person must have cognitive access to the reasons or evidence that justify their belief. Suppose that a person holds the belief that the pH of a strong acid is typically below 1, but they are unable to provide reasons as to why they believe this to be true. According to the internalist account, since this person is unable to access the reasons behind their belief, the latter would therefore not be justified. Now, contrast this with the same person, who later recalls learning this fact in their high school chemistry class. In this case, the individual has access to the reasons for their belief, which makes their belief justified according to the internalist account. Put very simply, if someone were to ask them, "How do you know that the pH of a strong acid is typically below 1?", they would be able to respond with reasons they are aware of, such as "I learned it during

this specific chemistry unit in high school". There is no consensus among internalists on precisely which internal factors can be used to form justifications. This diversity of opinion has resulted in the emergence of several influential internalist theories, including foundationalism, coherentism and evidentialism.

Foundationalism holds that some beliefs, called basic or foundational beliefs, are justified independently of other beliefs. These mostly include beliefs based on our sensory experiences, for example seeing or feeling something (e.g. I have a headache). Such beliefs provide the grounding for all non-basic beliefs, which are justified through their connection to this foundation. For example, the belief "I have a heart disease" is a non-basic belief. It may be justified by several basic beliefs, such as "I feel a tightness in my chest", "I hear a doctor say there is an abnormality", "Doctors are generally reliable experts", and so on. Together, these basic beliefs justify the non-basic belief of having a heart disease (Poston, n.d.).

Coherentism argues that a belief is justified just in case it stands in a system of mutually supporting relationships with other beliefs in a person's system of beliefs. Justification arises from the overall harmony, consistency, and explanatory power of the system of beliefs. For instance, my belief that the cat is on the mat involves a complicated set of beliefs: I am seeing a cat, I am seeing a mat, I am seeing a cat on a mat, a cat is a particular kind of mammal, a mat is a particular type of floor covering, my vision is generally reliable under normal circumstances, these are normal circumstances, and so forth (Watson, n.d.). These beliefs support and reinforce each other, creating a coherent web of beliefs. In this system, the belief "the cat is on the mat" is justified not in isolation but because it fits with and is supported by this broader network of beliefs.

Evidentialism claims that a belief is justified if and only if one has evidence which supports the belief in question. The core criterion for justification is therefore evidential support, meaning that a belief is justified when the evidence for it is strong and relevant (Feldman and Conee 1985). To illustrate this point, consider the following example: a detective

investigating a murder. The victim has been found dead in his apartment, and the detective initially suspects the victim's business partner, John. In order to justify this belief, the detective gathers evidence, including a bloodstained hammer found near the body, a neighbor who saw John leaving the building at the time of the murder, a motive involving a business dispute, and forensic evidence showing John's fingerprints on the hammer. On this account, the belief is justified not because it is coherent with other beliefs or grounded in basic beliefs, but because the evidence the detective has provides a justification for his belief that John is most likely guilty.

While internalist theories of justification appear elegant due to their intuitive nature, they continue to face significant criticism in the context of Gettier cases, where an individual might have a justified belief from an internalist perspective that nevertheless does not constitute knowledge (Watson, n.d.). This resulted in the development of another significant account of epistemic justification: externalism. Very simply put, externalism holds that what makes a belief justified is not necessarily internal to the believer; rather, it maintains that certain factors external to one's mental life play a necessary role in justification. The most prominent externalist account consists of Goldman's reliabilism which argues that one's belief is justified if it is formed by a reliable process, that is one that tends to produce mostly true beliefs (Goldman 1979). For instance, visual perception in good lighting or memory under normal conditions are generally considered reliable, whereas methods like guessing or wishful thinking are not. Therefore, a belief acquired through guesswork would not be justified, whereas a belief formed through careful visual observation under normal conditions would be. It is important to note that reliabilism does not require the holder of a justified belief to know that it was reliably produced. The belief can be justified without the agent knowing that it is; all that matters is that the process through which the belief is formed is reliable from a relevant intersubjective viewpoint (Swinburne 2001, 11-12). For example, imagine you hold the belief that smoking causes lung cancer, a belief you had for a while. Now, someone asks you to

explain how you know this. You cannot recall the specific sources or details of the information. But let us suppose that, some time ago, you read about the link between smoking and lung cancer in a reputable scientific article. Despite not remembering how you came to hold this belief, externalism argues that it is still justified because it was originally formed through a reliable process (consulting credible scientific research). Most of us hold beliefs like this: even though we may not always remember the exact reasons behind them, we do not question their veracity all the time. While reliabilism (and externalism, more broadly) can avoid Gettier problems, it faces a serious objection known as the "generality problem". This objection questions how we determine which type of process must be reliable for a belief to be justified. Since any belief can be described as the product of countless different processes, ranging from the very broad (e.g., "visual perception") to the extremely specific (e.g., "perceptual identification of an apple under indoor lighting from two meters away"), reliabilism struggles to specify which level of generality is relevant (Swinburne 2001, 14-15).

III. "FAKE IT TILL YOU MAKE IT"

Overview of Generative Artificial Intelligence

"Artificial intelligence" (AI), like the smell of campfire smoke on clothing, seems to linger in our lives, present no matter where we go, no matter where we look. It is on everyone's lips, everybody's mind. It is both a scientific discipline and a type of technology; this thesis is more concerned with the latter. As a technology, AI refers to systems that can perform tasks associated with human intelligence such as learning, reasoning, problem-solving, visual perception, natural language processing, and decision-making[4] (Independent High-Level Expert Group on Artificial Intelligence 2019). In other words, AI is designed to gather information, make sense of it, and then respond accordingly, whether that means solving a problem, processing and generating language, storing information about a certain domain, recognizing what it sees or hears, drawing conclusions on the basis of certain premisses, or learning from past experience. The term itself was coined in the 1950s by mathematician and computer scientist John McCarthy, and the 1956 Dartmouth Summer Research Project on Artificial Intelligence is widely regarded as the official birth of AI. Over the decades, significant advancements have been made in the field, with notable breakthroughs such as IBM's Deep Blue defeating the chess grandmaster Garry Kasparov (Mucci 2025). These developments have resulted in an increased presence of AI in our daily lives, as illustrated by the following examples: directions applications (e.g. Google Maps), self-driving cars, various home electronics (e.g. robot vacuum cleaners and security systems), search and recommendation algorithms (e.g. for watching a movie, shopping online, or when using social networks), facial recognition to unlock phones, and early diagnosis in healthcare, among countless other things.

[4] Consult the full report by the Independent High-Level Expert Group on Artificial Intelligence established by the European Commission for a comprehensive and thorough definition of AI.

This thesis focuses on "generative artificial intelligence", a particular branch of artificial intelligence which has attracted significant attention in recent years. In short, generative artificial intelligence can create new original content such as text, images, video, audio or software code in response to a user's prompt. Generative AI is based on "foundation models", an advanced form of deep learning with multilayered neural networks that are structured similarly to neurons in the human brain. Deep learning models work with vast amounts of unstructured data (e.g. text documents, images, audio files, and social media posts) and, while doing this, they learn to discover patterns and make informed decisions. This ability facilitates the execution of highly complex tasks, like translation, image recognition, or text generation, that were unattainable with earlier forms of machine learning (Microsoft. n.d.; Stryker and Scapicchio 2025; Zewe 2023; Copeland 2025). Foundation models differ from traditional learning models in that they are designed for generality and adaptability. As defined in a report to the European Parliament, a foundation model is "an AI system model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks" (Ivanova and European Commission 2021).

One of the most famous examples of a foundation model is the large language model (LLM), which is designed to process and generate text in natural language. During the training stage, it learns syntactic patterns, grammar, and word relationships, and becomes able to predict the most likely next word or phrase in a sentence. This allows LLMs to generate text that is coherent, contextually relevant, and often so convincingly human-like that it can be nearly indistinguishable from content written by actual people (Cao et al. 2023). Indeed, the latest version of ChaGPT (GPT-4.5), which is one the most widely known examples of this type of generative AI, recently passed a modern version of the Turing Test by convincing judges it was human 73% of the time (Landymore 2025). The potential applications of this type of generative AI are extensive, ranging from the production of academic papers and journalistic articles to

language translations, scientific reports, legal documents, medical recommendations and personal messages such as emails and social media posts, all having the capacity to appear as authentic and credible as if written by a human.

Another example of foundation model is the generative adversarial network (GAN). It consists of two neural networks called the generator and the discriminator, which work against each other to create data that appears authentic. The generator's role is to produce convincing, realistic content, while the discriminator's role is to evaluate the authenticity of that content and to classify it as either real or fake. The generator is continually improving its outputs based on the feedback received from the discriminator. As the generator's capabilities in producing realistic content improve, the discriminator must also enhance its ability to detect fakes, creating a cycle of mutual improvement. Over time, this process results in generated outputs that are almost indistinguishable from real ones. In short, the generator learns to produce ever more convincing content by trying to fool the discriminator. This type of generative AI is predominantly employed to generate images, audio content or videos[5] (Cao et al. 2023). For example, it can create entirely new images from textual descriptions, ranging from surreal artworks to images that appear indistinguishable from real photographs. For instance, an AI-generated image featuring a meeting between Elon Musk, Donald Trump and Benjamin Netanyahu last February was mistaken for a real photograph and was even used by news outlets before being discredited. Recent studies have further confirmed the increasing challenges in distinguishing between human-generated and AI-generated images. In a 2022 study by Nightingale and Farid, some participants were asked to distinguish between real and AI-generated faces. Participants not only failed to reliably distinguish between real and AI-generated faces but also rated synthetic faces as more trustworthy than real ones. (Nightingale and Farid 2022). A 2023 study further supported these findings by also testing participants'

[5] Another type of foundation model known as diffusion model is also used to create images, videos and audio. The most famous example of such type of diffusion model is DALL-E (Sheldon and Kerner 2024).

ability to identify synthetic faces from real ones across multiple experiments. The authors concluded that "people cannot reliably detect that AI-generated faces are synthetic," and warned that as generative technology continues to improve, the public's ability to critically evaluate visual evidence may decline further (Miller et al. 2023). Similarly, in the audio domain, this type of generative AI can create highly realistic speech by replicating the voice, tone, rhythm, and emotional inflection of real individuals. A study by Barrington, Cooper, and Farid (2024) found that participants correctly identified AI-generated voices only 60% of the time (Barrington, Cooper, and Farid 2025). This difficulty in voice verification is not just theoretical: in March 2019, the chief executive of a UK subsidiary of a German energy firm was reportedly tricked into transferring nearly £200,000 to a Hungarian bank account after receiving a phone call from someone mimicking the voice of the German parent company's CEO (Sample 2023). In the video domain, this type of generative AI can produce entirely synthetic footage or manipulate existing videos in ways that make them appear completely authentic. This technology can, for instance, fabricate a video of a person saying or doing something they never actually said or did, yet make it look entirely realistic and believable. A striking example occurred in March 2022, when a fabricated video circulated online showing Ukrainian President Volodymyr Zelenskyy seemingly urging his troops to surrender (Wakefield 2022). The video spread across social media, creating confusion and panic before eventually being debunked. Cases like this are becoming increasingly common, drawing widespread public and media attention, and highlighting the growing difficulty of discerning truth from fabrication. Collectively, AI-generated synthetic images, audio, and video that convincingly mimic real people or events are often referred to under the term "deepfakes". The term was originally coined in the context of manipulated videos, but it has since been broadened to include hyper-realistic AI-generated content across modalities (Yasar, Barney, and Wigmore 2025).

It is important to note that some people might argue that, with careful observation, it is still possible to distinguish real content from AI-generated fabrications. However, given the evidence discussed above, and considering that we are only at the early stages of what is likely to be an exponentially advancing technology, especially with the emergence of quantum computing[6], it is fair to assume that the gap between the real and the synthetic will continue to narrow. In this context, it is reasonable to assert that, for the average person without specialized tools or expertise, AI-generated content is, or soon will be, truly indistinguishable from authentic material. Recent breakthroughs like Google Veo 3, which can generate detailed and photorealistic videos from simple text, support this argument and suggest that this indistinguishability will only deepen (Saltiel 2025).

[6] For further discussion on how quantum computing may accelerate the development of generative AI, see Shu et al. (2024) and Campbell (2024).

IV. EPISTEMIC JUSTIFICATION IN THE AGE OF GENERATIVE AI

In the previous sections, I examined the concept of epistemic justification and its central role in both the formation and evaluation of beliefs. I also presented a threefold framework of *justifier, justification, and belief*, which clarifies how beliefs are formed and supported. Building on this, I discussed how major theories in epistemology (foundationalism, coherentism, evidentialism, and reliabilism) define appropriate justification and specify the conditions under which a belief is considered justified. I then explained how generative artificial intelligence is increasingly blurring the boundary between authentic and synthetic content by creating multimodal content that is indistinguishable from real content. In this section, I bring these two threads together by arguing that generative AI disrupts the very conditions under which epistemic justification can be said to function appropriately, both on a subjective level and on an intersubjective level. Furthermore, I will show how generative AI undermines the traditional theories of justification.

The starting point of my analysis is the fact that generative AI can produce justifiers that are entirely fabricated yet perceptually indistinguishable from real ones. These synthetic justifiers (e.g. deepfake videos, AI-generated texts, AI-generated images, fabricated audio recordings) mimic the features of authentic evidence so convincingly that agents have no reliable way of telling them apart from appropriate justifiers. As a result, individuals may come to accept fundamentally inappropriate justifiers as valid sources of justification. This indistinguishability, therefore, introduces a critical problem: if a belief is justified on the basis of an inappropriate justifier that cannot be differentiated from an appropriate one, then the entire justificatory structure becomes compromised. This core issue sets the stage for the two key consequences that follow at the subjective and at the intersubjective level.

One of the most immediate impacts of generative AI lies in how it alters the subjective experience of justification. A justification's strength, as discussed earlier, is understood as how compelling the agent takes it to be. In other words, an individual may view one justification as stronger than another simply because they find it more convincing, even if it is ultimately based on inappropriate justifiers. This subjective dimension of strength has always been vulnerable to error: people can be misled by hallucinations, lies, illusions, or manipulated evidence and yet remain unaware of the inappropriateness of their justifiers. What generative AI changes is that an individual's confidence in inappropriate justifications may now reach unprecedented levels. Because generated AI content can seamlessly align with an individual's sensory and cognitive expectations, they can provide a false sense of certainty in having inappropriate justifiers. This means that an agent might judge a justification for a belief to be extremely strong (even though it is inappropriate), simply because they are highly confident in the justifiers that support it, even if those justifiers are themselves inappropriate. If the justifiers align with their direct sensory experience, the agent has no reason to doubt their veracity, and will therefore not question the justification for the belief, nor the belief itself. Moreover, unlike traditional justifiers such as faulty memory or second-hand testimony, which might be reassessed and discredited upon further reflection or verification, AI-generated justifiers retain their deceptive plausibility even upon re-examination. This means that even after the belief has been formed, going back to inspect the justification, say, rewatching the video or rereading the fabricated article, will not raise concerns for the agent. The justifiers continue to appear appropriate, reinforcing the illusion of epistemic legitimacy and making it much harder for the agent to revise or abandon the belief. For example, whereas in the past an individual might have expressed doubt about their belief or its justificatory basis, saying something like, "Maybe I misheard that" or "Maybe I did not look properly", they are now far more likely to insist, "I saw it with my own eyes, here is the video, so I know I am right and justified in believing this."

In such cases, the belief is held with increased confidence because the justification appears exceptionally strong to the subject. And the justification, in turn, appears strong precisely because the justifier, the sensory experience of seeing the video, perfectly aligns with the agent's perceptual expectations. Yet, if the video is in fact AI-generated, the justifier is inappropriate, the justification is flawed, and the belief may well be false, despite the agent's unwavering confidence. The vividness and realism of the AI-generated justifier strengthen the subjective evaluation of the justification, even though the epistemic appropriateness of the justifier is deceptive. What makes this even more concerning, as explained above, is that if the agent revisits the justifier, by rewatching the video, for instance, it will still appear just as credible and appropriate as it did initially. In contrast to traditional cases of inappropriate justifiers where reflection or reevaluation might reveal errors or inconsistencies, AI-generated content maintains its deceptive realism over time. This reinforces the belief retroactively, making it even more resilient to revision or doubt. In short, generative AI does not merely increase the likelihood of error by rendering inappropriate justifiers indistinguishable from appropriate ones; it also undermines the reliability of retrospective critical evaluation, further increasing the conviction with which individuals hold unjustified or falsely justified beliefs, making inappropriate justifications appear stronger than ever before. This intensifies the epistemic risk because falsehoods are now more confidently believed, and therefore more resilient to correction.

The second, and arguably more significant, impact of generative AI on epistemic justification concerns the intersubjective evaluation of justification, our ability to assess whether someone else's belief is justified. Normally, as explained in the first section, justification functions as a filter that helps us distinguish well-supported beliefs from mere opinions. Without reliable ways to assess whether a justification is appropriate, all beliefs risk appearing equally credible, even when some are clearly false or poorly grounded. To illustrate

this, imagine two individuals holding contradictory beliefs: "Smith killed Jones" vs. "Smith did not kill Jones". One of these beliefs must be false, as both cannot be true at the same time. How can we determine which one? In traditional contexts, we start by evaluating the appropriateness of the justifiers. But, with generative AI, it is not clear whether we are still able to conduct such an assessment. Suppose we have access to the justifiers the individuals used to form their justification, for example, one person presents a video allegedly showing Smith committing the murder, while the other presents a different video showing Smith in a different location at the same time. If one of these videos is a deepfake and the other is authentic, we (as non-expert) would not be able to tell which one is the real one, hence we would not be able tell who is relying on appropriate justifier and who is not. Second, if both beliefs are backed by justifications that satisfy the criteria of existing epistemological theories, for example, both are the result of seemingly reliable processes like visual perception (as required by reliabilism), or both cohere with the rest of the agent's beliefs (as in coherentism), then, even on a theoretical level, both beliefs appear appropriately justified, despite being contradictory. Unable to rely on an assessment of the justifier appropriateness nor the appropriateness of the justification itself, we would regard both individual's justification for their belief as being appropriate. This situation reveals a fundamental breakdown in the intersubjective evaluation of beliefs, where even contradictory beliefs may be treated as appropriately justified. As a result, justification begins to lose its normative power in helping us assess which beliefs are more likely to be true.

In what follows, I will analyze in a detailed way how the emergence of generative AI undermines the traditional theories of appropriate justification.

Evidentialism is perhaps the most directly threatened. According to evidentialism, a belief is justified if and only if it is supported by the believer's evidence. However, since AI-generated content is indistinguishable from real one, this means that a person may hold a belief that appears to be well-supported by strong evidence, when in fact the evidence is entirely

fabricated. Consider, for example, someone believing that a public figure has committed fraud based on what seems to be overwhelming evidence: a leaked email exchange, a recorded phone call, and a video confession, all entirely plausible, and generated by AI. From an evidentialist standpoint, the belief appears appropriately justified: the person has multiple, independent, and apparently credible pieces of supporting evidence. Yet if all these justifiers are in fact fabricated, then the belief is inappropriately justified, even though it might seem perfectly legitimate to the subject. This highlights a deep vulnerability in evidentialism: if fabricated evidence can be just as convincing as real evidence, then evidentialism loses its ability to distinguish between genuinely justified and merely seemingly justified beliefs, thereby undermining it as a framework for epistemic evaluation.

Foundationalism is equally challenged by generative AI. Traditionally, foundationalism holds that some beliefs, especially those grounded in direct sensory experience, are basic and justified independently of other beliefs. These basic beliefs then support the justification of further, non-basic beliefs. However, generative AI fundamentally challenges this theory. If our sensory experiences can be deceived by AI-generated content, then a belief that appears to be grounded in direct perception, and would therefore qualify as basic, may in fact rest on inappropriate justifiers. In other words, a basic belief can be entirely false without the subject realizing it. For instance, suppose a person watches what seems to be a livestream video of a natural disaster occurring in a specific region. They form the basic belief: "A tsunami is hitting Jakarta right now," based on what appears to be direct visual perception. Because this belief seems to be grounded in immediate sensory experience, it qualifies as basic. Yet if the video is a deepfake, the justifier is in fact inappropriate, and the belief is false. But generative AI threatens foundationalism in a second, more subtle way. Even when a person's basic beliefs are accurate, such as "I am watching a video," "I hear someone speaking," or "this appears visually real", "I can trust my vision", AI can exploit those reliable perceptions to justify false non-basic beliefs. For example, suppose someone watches a

professionally produced AI-generated documentary that contains realistic footage, expert interviews (entirely fabricated), and persuasive narration claiming that a major pharmaceutical company knowingly distributed a harmful vaccine. The viewer's basic beliefs "I am watching a video," "I hear a narrator explaining a sequence of events," "these people appear to be medical professionals," "this looks like credible documentary footage", are all true. There is no deception at the level of their direct sensory experience: they are indeed seeing and hearing what they report. However, the non-basic belief they infer, "the pharmaceutical company deliberately harmed people and covered it up" is false. It is built upon true basic beliefs, but the coherence and realism of the AI-generated content make the higher-level inference appear justified. The problem lies not in the basic beliefs, but in the non-basic belief constructed from them. In this case, the non-basic belief is false but appears appropriately justified because it aligns with true basic beliefs. Thus, even when the foundational inputs remain intact, the system is still vulnerable to manipulation. Together, these challenges reveal a fundamental weakness in foundationalism: generative AI can both create false basic beliefs through deceptive sensory content and also produce false non-basic beliefs that appear justified by true basic beliefs. This dual vulnerability undermines foundationalism's claim to offer a secure and reliable structure for epistemic justification.

Coherentism might respond to the challenge faced by foundationalism, where basic beliefs can be impacted by inappropriate justifiers, by emphasizing the interconnected nature of beliefs. For coherentists, a belief is justified not by its inferential relation to a basic belief, but by how well it fits within a larger, mutually supportive system of beliefs. Each belief is justified in virtue of its coherence with the broader network of beliefs that support it. However, generative AI undermines this framework as well: it can introduce false beliefs based on (AI-generated) inappropriate justifiers that nonetheless cohere with an individual's existing network of beliefs. Consider, for example, a person who, after reading multiple AI-generated articles presenting fabricated scientific studies, comes to believe that a certain dietary

supplement drastically improves cognitive function. These articles reference fake expert opinions and seemingly legitimate experimental results, all crafted to fit common beliefs about health and wellness. The person's existing web of beliefs includes trust in scientific research, a general acceptance that dietary supplements can improve health, and a personal interest in cognitive enhancement. Despite being entirely false, the belief that the supplement is effective is justified because it coheres well with the person's existing beliefs about health trends, trust in scientific research, and personal interest in cognitive enhancement. This example illustrates how AI-generated falsehoods can seamlessly integrate into an existing belief network, making it difficult for coherentism to distinguish whether a belief is appropriately justified (and true). This problem worsens once such a false belief is integrated into the existing network of beliefs: it can go on to support further false beliefs, increasing the likelihood that other AI-fabricated claims will also appear coherent and thus justified. In this way, generative AI not only inserts falsehoods into belief systems but gradually "infects" the system itself, making it easier for future false beliefs to gain coherence-based justification. The very structure that coherentism relies on (mutual support among beliefs) is exploited, revealing that coherence, like foundational support, no longer reliably tracks truth when the inputs are manipulated by AI. In short, generative AI can implant false beliefs that fit in a coherent way within an existing web of beliefs and then propagate further falsehoods, thereby undermining coherentism's ability to serve as a reliable guide to epistemic justification.

Reliabilism, too, is seriously undermined by generative AI. According to reliabilism, a belief is justified if it is produced by a cognitive process that is generally reliable, one that tends to produce mostly true beliefs. Processes like visual perception, memory, or testimony are often cited as examples of such reliable mechanisms. However, the problem introduced by generative AI is that these once-reliable processes can now be systematically deceived by inappropriate justifiers (created by generative AI) that are indistinguishable from appropriate ones. This means, for example, that an individual can form a belief based on a deepfake video, and that

belief would be the result of what would typically count as a reliable process, visual perception in this case, but the belief would ultimately be false and inappropriately justified because of the inappropriateness of the justifier. To illustrate this more saliently, suppose a person watches a deepfake video of a political leader making an inflammatory statement and forms the belief that the event occurred. The belief is formed through visual perception, a process that is normally reliable. Yet because the video (the justifier) was artificial and deceptive, the resulting belief is false and unjustified, despite appearing to meet reliabilism's criteria. The person relied on a generally truth-conducive process, but in this case, that process was compromised by AI-generated content. Since reliabilism depends on the overall reliability of the belief-forming process, the introduction of highly realistic synthetic content fundamentally compromises its reliability. Since generative AI can create multimodal synthetic content, it seems reasonable to argue that most processes which were once regarded as reliable according to this theory are no longer so. Visual perception, hearing, and testimonial trust, among other processes, are no longer consistently truth-conducive, as AI-generated content can mimic reality so closely that these processes frequently produce false beliefs. It is important to note that one might argue that if the belief-forming process in question is narrowly specified (e.g. forming beliefs based on videos encountered on social media without independent verification), then perhaps one could maintain that such beliefs are not formed via a generally reliable process and therefore would not be justified under reliabilism. However, this response ultimately circles back to the generality problem while also placing unrealistic demands on epistemic agents. The disruption introduced by generative AI is not confined to the digital realm; it also extends into offline environments through the testimony of others. If, for instance, you regard someone as trustworthy and they themselves have been misled by AI-generated content (e.g. synthetic audio, a fabricated article), you may come to adopt a belief based on their testimony. In this case, even though the immediate process appears reliable (trusting a credible person), the underlying justifier remains inappropriate. As a result, the reliability of the entire belief-

forming process is compromised, even in seemingly non-digital contexts. This illustrates how pervasive and insidious the epistemic disruption caused by generative AI truly is, and, how, like the other theories, reliabilism fails to provide a solid framework for determining whether a belief is appropriately justified.

What emerges from this analysis is a deeply troubling epistemic picture. Generative AI does not merely increase the risk of false beliefs, it disrupts the very mechanisms by which we distinguish between justified and unjustified ones. At the subjective level, individuals become increasingly confident in inappropriate justifications, because AI-generated justifiers perfectly align with their sensory experience, and even retrospectively, these justifications no longer appear questionable. At the intersubjective level, our ability to evaluate others' beliefs is equally compromised: when justifiers become indistinguishable and justifications satisfy formal epistemic standards, even false beliefs can appear to be fully appropriately justified. As it has been demonstrated, this breakdown renders our traditional theories of justification (foundationalism, coherentism, evidentialism, and reliabilism) largely ineffective in practice. If we can no longer determine which beliefs are appropriately justified, then justification loses its normative function as a filter guiding us on the path towards truth. In this sense, the distinction between true and false beliefs collapses not in theory, but in application. The problem is no longer just that people can hold false beliefs, as this has always been the case. The real danger now is that the very tool we usually rely on to identify and correct those false beliefs, justification, is being weakened. This weakening of justification as a reliable guide to truth has consequences that reach beyond isolated epistemic failures. What is ultimately at stake is our capacity to function as epistemic agents. When individuals and communities can no longer reliably distinguish between genuinely justified beliefs and unjustified ones, because inappropriate justifications appear entirely appropriate, our ability to responsibly form and assess beliefs begins to erode. This collapse not only weakens individual judgment but

threatens to undermine our collective epistemic practices. In the following section, I explore the broader implications of this shift, focusing on the collapse of epistemic agency and the rise of a new kind of skepticism brought about by generative AI.

Before turning to these broader epistemic consequences, it is important to recognize that the epistemic threat posed by generative AI is not merely passive or incidental: the indistinguishability of justifiers can be used actively by individuals who seek to deceive others. If one desires to spread a lie, generative AI now provides the unprecedented ability to construct not only a superficially appropriate justification, but also a perceptually convincing justifier to support that lie, increasing the persuasive power of the lie and making it more resistant to scrutiny. A particularly concerning manifestation of this is the phenomenon of disinformation[7]. Generative AI significantly amplifies the scope and effectiveness of disinformation campaigns by allowing actors to fabricate entire ecosystems of supporting content: realistic videos, expert-like commentary, official-looking documents, and corroborating testimonies, all of which can be generated on demand. As the ACM report describes, the danger lies in the emergence of "doppelgänger disinformation", an operational structure in which multiple AI-generated elements mutually reinforce each other, creating an "ecosystem of credibility" that mimics the appearance of authentic justification (Jaidka et al. 2025). This means that falsehoods are no longer simply asserted: they are embedded within a web of fabricated justifiers that closely resemble those used in legitimate epistemic practices. An example of how generative AI can be used in disinformation campaigns happened early May when France accused Russian operatives of creating and disseminating an AI-generated video falsely alleging that Brigitte Macron, the wife of President Emmanuel Macron, sexually assaulted a 12-year-old pupil in the 1980s. The video, part of Russia's alleged disinformation operation called Storm-1516, gained nearly 10 million views. French authorities confirmed the video featured a deepfake of a former

[7] false information deliberately created or spread to deceive, manipulate, or influence people's beliefs, often for profit or political gain (Epstein 2023).

pupil who denied the accusations, but this clearly highlights the disruptive power of AI-generated synthetic media in spreading false narratives (Chazan 2025). In this context, disinformation becomes not just a political or ethical concern, but an active epistemic weapon, one that exploits the indistinguishability of justifiers to erode the very evaluative structures upon which both individual and collective epistemic agency depend.

V. COLLAPSE OF EPISTEMIC AGENCY AND RISE OF SKEPTICISM

Epistemic agency can be defined as an individual's capacity to exercise control over their epistemic life, that is, their ability to actively form, evaluate, revise, and take responsibility for their beliefs in accordance with epistemic norms such as truth, justification, and evidence. This concept presupposes a subject who is not merely a passive recipient of information, but an agent capable of assessing reasons and exercising epistemic responsibility (Engel 2013; Nieminen and Ketonen 2024). Yet, as the previous section has demonstrated, the epistemic disruptions introduced by generative AI increasingly undermine this agency. When individuals are unable to reliably determine whether the justifiers of their beliefs, such as videos, images, or documents, are appropriate or fabricated, they lose the ability to assess the epistemic status of their beliefs in any meaningful way. Likewise, if one cannot evaluate whether another person's justification is appropriate, nor rely on existing theories of justification to guide this assessment, the evaluative frameworks that typically support epistemic agency break down. As a result, individuals find themselves epistemically challenged, unable to critically assess their own beliefs or engage with those of others. Even if we strive to be responsible epistemic agents, our ability to act as such collapses when we lack the means to distinguish appropriate justifiers from inappropriate ones, because generative AI outperforms our cognitive and perceptual capacities. As Coeckelbergh insightfully argues, "In an environment where it is no longer clear what is true or not, real or not, I cannot exercise my capacities for epistemic agency" and "if AI fakes the news (or might fake the news: the point is that I never know if AI was used or not), then I have effectively less epistemic agency: I have less control over the formation of my knowledge", and further "My epistemic agency in this respect is minimal compared to what AI can do (Coeckelbergh 2023, 1344-1345). This collapse is, therefore, not a matter of epistemic laziness or apathy, but of structural incapacity: the

conditions that once enabled responsible belief formation, trust in perception, the ability to evaluate sources, and shared justificatory norms, are eroded by AI's ability to simulate convincing and appropriate justifiers at scale. In short, the collapse of epistemic agency is not due to a failure of will, but of epistemic infrastructure: it occurs because the informational environment created by generative AI exceeds the capacities of human agents to responsibly navigate it. It is important to note, however, that this informational environment was already deeply challenging for epistemic agents, even before the full force of generative AI was felt. As Buzzell and Rini argue, the modern digital epistemic landscape places "excessive epistemic demands" on individuals, demands that are often structurally impossible to meet (Buzzell and Rini 2023, 5). The sheer volume of information, the speed at which it circulates, the decline of traditional gatekeepers, and the fragmentation of epistemic communities make it unrealistic to expect individuals to verify and assess knowledge claims independently. They write that this environment "requires of individuals a kind of epistemic superheroism that few, if any, can reliably perform" (Buzzell and Rini 2023, 5). In this context, the arrival of generative AI does not initiate but drastically intensifies these pressures. It transforms an already overloaded and unstable environment into one where core epistemic functions, like evaluating justification, begin to break down entirely, thereby accelerating the collapse of epistemic agency.

Some might argue that the epistemic disruptions outlined in this thesis are confined to the digital realm, and thus not as serious or far-reaching as they may appear. However, this objection underestimates the central role that the digital environment now plays in our epistemic lives. As Neuberger et al. argue, *"*digital media have become central to knowledge acquisition, verification, and dissemination*"* (Neuberger et al. 2023, 3) and increasingly shape how individuals access, evaluate, and circulate information. Indeed, today, the vast majority of people rely on digital platforms as their primary source for learning about the world, whether it is reading the news, researching topics, or forming beliefs about politics, science, and society.

The 2024 Reuters Institute Digital report, for example, highlights that a majority of people worldwide now access news primarily through online platforms, overtaking traditional media such as television and print in many regions (Reuters Institute 2024). In this sense, the digital sphere is not a marginal domain of knowledge, it is the primary infrastructure through which most knowledge practices are now realized, and through which we understand the world. Consequently, the challenges posed by generative AI within this environment, particularly around justification, trust, and epistemic agency, carry deep and widespread consequences. The collapse of epistemic agency in the digital realm thus entails a broader epistemic crisis, not because it originates offline, but because online environments have become inseparable from contemporary knowledge practices.

The collapse of epistemic agency does not merely mark a failure to meet certain epistemic standards; it gives rise to a deeper epistemological condition: skepticism. This skepticism is not rooted in abstract philosophical doubt, but in an increasingly common and concrete experience: the recognition that we may no longer be able to reliably determine what is true, who to trust, or whether we ourselves are still capable of navigating the epistemic terrain responsibly (Harris 2024, 7). Importantly, skepticism does not follow automatically from the collapse of epistemic agency; it emerges when individuals become aware of its collapse, as well as the capabilities of generative AI. Many people may continue to operate under the assumption that their epistemic environment is stable, that their senses are trustworthy, and that their beliefs are well-grounded. However, once one becomes aware that generative AI systems can fabricate justifiers indistinguishable from real ones, that these systems now outperform human perceptual and cognitive capacities, and that this makes traditional processes of justification unreliable, then a form of epistemic skepticism becomes almost inevitable. I argue that this recognition gives rise to a triadic structure of skepticism: skepticism about reality, as in doubt about our own sensory experience; skepticism about others, as in doubt about the

reliability of testimony and interpersonal epistemic trust; and skepticism about the self, as in doubt about our own capacity to function as competent epistemic agents.

The first form of skepticism introduced by generative AI echoes classical philosophical skepticism, the ancient doubt about whether we can trust our sensory experience to reveal the external world as it truly is. While generative AI does not invoke metaphysical skepticism, it produces a similar epistemic effect. In a digital world where synthetic videos, images, voices, and texts are perceptually indistinguishable from real ones, the line between reality and fabrication blurs. It is not, like in classical skepticism, the fear that nothing is real, but the growing uncertainty about what can be trusted as real. This form of skepticism is grounded in the failure of justifiers, a problem already outlined in the previous section. As generative AI systems increasingly produce content that mimics real-world sensory inputs, the default trust in perception and apparent evidence begins to erode. People begin to question not only isolated pieces of information but the general reliability of perceptual and testimonial inputs, even seeing something "with one's own eyes" no longer guarantees its reality. In other words, this skepticism extends to the trustworthiness of perception itself as a source of knowledge in digital contexts, undermining sensory experience as a reliable foundation. As AI-generated content becomes seamlessly integrated into the digital landscape, this leads to a generalized epistemic insecurity, which fosters distrust toward all digital content (Rini 2023, 7). Indeed, people may begin to regard any content they encounter online, whether it's a news clip, a scientific study, or a personal video, as potentially artificial. The everyday manifestation of this skeptical shift can be seen in the proliferation of questions like *"Is this real?"*, *"Is this true?"*, or *"Was this made by AI?"* These reactions reflect a deeper uncertainty not about all knowledge, but about the reliability of justifiers in a digital context.

The second form of skepticism introduced by generative AI targets our ability to trust others as reliable epistemic agents. This skepticism arises in two stages. First, once we become aware that other people may confidently hold false beliefs based on inappropriate justifiers, not because they are negligent or irrational, but because, as previously explained, generative AI can produce justifiers that align so perfectly with their perceptual and cognitive expectations that they seem entirely appropriate, a fundamental doubt about the legitimacy of their belief emerges: we begin to wonder: "*Can I really trust what this person says, or were they just misled by something that only looked real?*". Indeed, if people can be entirely confident in their beliefs while unknowingly relying on fabricated content, then their testimony, judgments, or reported evidence lose much of their epistemic weight, and the broader norm of epistemic trust, which underpins testimonial exchange, begins to erode. This skepticism deepens when we realize that, regardless of whether a person formed a belief based on appropriate justifiers, we ourselves may be unable to tell whether the belief is appropriately justified. As established in the previous section, generative AI has rendered it increasingly difficult to distinguish between appropriate and inappropriate justifications. Even when someone appears to offer a well-structured, seemingly appropriate justification, we cannot determine whether it is genuinely appropriate because of the indistinguishability of AI-generated justifiers. The traditional role of justification, to serve as a normative filter between beliefs that are justified and those that are not, begins to break down. As a result, we are left in a position of epistemic skepticism: there seems to be no reliable way to tell whether other people's belief are justified. In this context, skepticism toward others is not only reasonable, but often becomes unavoidable.

The third and final form of skepticism introduced by generative AI concerns the self, specifically, our capacity to function as competent epistemic agents. This form of skepticism emerges when individuals become aware that they may no longer be able to reliably distinguish between appropriate and inappropriate justifiers, a realization that calls into question their own

competence as knowers. Indeed, if we can no longer trust our sensory experience (as in the first form of skepticism), and we can no longer trust others' beliefs or justifications (as in the second), we are inevitably led to question our own epistemic competence. This form of skepticism is not about isolated errors in judgment or occasional lapses in reasoning. It is the more unsettling realization that, even when we do our best to think critically, verify sources, and reflect carefully, we may still be systematically deceived, not because we are careless, but because our cognitive and perceptual capacities are simply no longer adequate to the informational environment we inhabit. This skeptical stance solidifies when we recognize that generative AI not only deceives us, but does so in ways that surpass our epistemic abilities. AI-generated content does not merely compete with human cognition, it often outperforms it, creating justifiers so convincing that no amount of scrutiny seems sufficient to reveal their inauthenticity. As this realization sets in, we do not merely doubt specific beliefs, but begin to lose confidence in our own epistemic competence. In other words, we become skeptics of our own capacities as epistemic agents, recognizing that deception is possible not despite our best efforts, but precisely because those efforts are no longer sufficient. In Coeckelbergh's words, "The problem is also that citizens can no longer believe their own eyes and hence start doubting and mistrust not only others but also their own epistemic capacities. If AI fakes increasingly more "believable", then I start questioning my own capacities as an epistemic agent to distinguish truth from falsehood" (Coeckelbergh 2023, 1344). The result is a form of skepticism that is deeply self-reflective: not just "*Is this real*?" or "*Can I trust others?*" but "*Can I still trust myself to know what is true?*" or "*Am I still capable of knowing reliably?*". Some people might argue, however, that this loss of confidence in our own epistemic competence can be resolved. After all, if generative AI is the source of epistemic instability, perhaps AI itself can be used to counteract it, specifically, through AI detection tools (ADDs) designed to identify fabricated content, which would then enable us to tell whether a justifier is appropriate or AI-generated. Yet this apparent solution introduces a deeper epistemic cost.

First, it is far from clear that such detection systems will be able to keep pace with the rapid evolution of generative models. But more importantly, even if they could, relying on them would mean delegating a fundamental aspect of epistemic evaluation. As Harris notes, "to rely on an ADD instead of one's own faculties is to accept the vulnerability of the process by which many of one's most confidently-held beliefs have been formed." This reliance fosters "a general kind of self-doubt that is inimical to epistemic autonomy" (Harris 2024, 15). That is, when we come to believe that we cannot even trust our own senses without the aid of external verification tools, we do not merely adopt a cautious epistemic stance, we begin to mistrust ourselves as knowers. This reinforces the very form of skepticism introduced in this section: doubt not just about what we perceive or whom we trust, but about whether we ourselves remain capable of navigating the world as autonomous epistemic agents. As Harris explains, "when it comes to the formation of beliefs through perception of video footage, we will regard the evidence of one's senses as an insufficient basis on which to form beliefs," and this shift "amounts to ceding some epistemic autonomy" (Harris 2024, 15). What is presented as a solution to the epistemic threat of generative AI may in fact exacerbate the very problem it aims to fix: reinforcing this skepticism about our own ability to know reliably without artificial assistance.

VI. CONCLUSION

Entering the "Post-Knowledge Era"

In this thesis, I have argued that generative AI seriously undermines the epistemic role of justification by disrupting both the subjective and intersubjective dimensions through which justification typically operates. At the core of this disruption lies the indistinguishability of AI-generated justifiers from authentic ones: AI-generated multimodal content can mimic the perceptual and structural features of genuine justifiers so convincingly that agents often cannot reliably tell them apart. As a result, inappropriate justifiers are increasingly treated as appropriate, compromising the very basis upon which beliefs are formed and evaluated. At the subjective level, individuals may come to regard inappropriate justifications as unusually strong, precisely because the justifiers supporting them align so well with their cognitive and perceptual expectations. At the intersubjective level, the traditional function of justification as a normative filter collapses: even when presented with seemingly credible justifications, we can no longer reliably assess whether those justifications are appropriate. Consequently, the ability to distinguish justified from unjustified belief is severely weakened, making it increasingly difficult to separate falsehoods from truth in our epistemic practices. Moreover, I have showed that the indistinguishability of justifiers renders traditional epistemological theories of justification such as foundationalism, coherentism, evidentialism, and reliabilism ineffective as frameworks for evaluating justification and beliefs in practice. I then explored how these disruptions lead to a collapse of our epistemic agency. This collapse arises because generative AI exceeds our cognitive and perceptual capacities: even when we try to reason responsibly, we may still be systematically misled by justifiers that are indistinguishable from authentic ones. This collapse, when combined with the growing realization of both its implications and the capabilities of generative AI, gives rise to a triadic structure of skepticism. First, skepticism about reality, where the reliability of sensory perception is called into

question. Second, skepticism about others, as we can no longer assess whether someone else's justification is appropriate, weakening the role of testimony. And third, skepticism about ourselves, as we begin to doubt our own capacity to form justified beliefs in a digital environment we can no longer navigate with confidence.

These epistemic disruptions resulting from the development and spread of generative AI culminate in a broader transformation, one that calls for a new conceptual vocabulary. I propose the term "*post-knowledge"* to name this emerging condition. Drawing inspiration from the notion of "post-truth", which describes a political climate where facts matter less than emotions and beliefs (Suiter 2016), *post-knowledge* goes deeper. It does not merely reflect an indifference to truth, it marks a structural collapse in our ability to access, evaluate, and claim knowledge itself: the mechanisms that traditionally allowed us to separate mere belief from genuine knowledge, namely, the reliability of justification and the viability of epistemic agency, are eroding. In other words, post-knowledge does not mean that knowledge vanishes altogether, but that our ability to confidently know has been profoundly destabilized and undermined in a world where almost anything can be artificially constructed and convincingly justified. In short, while post-truth reflects a disregard for truth, post-knowledge argues that even those who seek to know may no longer be able to. It is not a crisis of attitude, but a collapse of epistemic conditions: in the *post-knowledge era*, individuals can no longer reliably determine whether they truly know something, or whether their belief merely appears to be knowledge. More than a theoretical concept, *post-knowledge* is a life condition, deeply embedded in the digital texture of our social, civic, and informational lives. In a world where people routinely ask themselves "Is this real?", "Was this made by AI?", and "Can I still trust myself to know?", societies become vulnerable to strategic doubt and epistemic manipulation; skepticism can be weaponized. Ill-intentioned actors can exploit this environment to spread disinformation, fracture shared understanding, and undermine trust in institutions, media, and

each other. In this sense, the *post-knowledge* condition is not just a philosophical problem, it is a societal vulnerability. Future research should investigate its connection to growing polarization, radicalization, and the erosion of democratic discourse.

This brings us back to the fundamental question: "*Can we still know?*". In the face of generative AI's disruptive epistemic consequences, the answer grows increasingly fragile. We may have already crossed the threshold into the post-knowledge era, but the full consequences are not yet inevitable. If we do not begin to address the epistemic challenges introduced by generative AI, rethinking justification, strengthening epistemic norms, and rebuilding trust, we risk slipping beyond recovery, into a condition where the very idea of knowledge becomes obsolete. And one day, future generations may look back on this moment, on our brilliant innovations and our blind and untamed curiosity, and say of us what was once said of a myth: *they opened the box, and the world was never the same.*

VII. BIBLIOGRAPHY

- Ayer, A.J. 1936. *Language, truth and logic*. London,: V. Gollancz.

- Barrington, S., Cooper, E.A. & Farid, H. 2025. "People are poorly equipped to detect AI-powered voice clones". *Sci Rep* 15, 11004. https://doi.org/10.1038/s41598-025-94170-3

- Brown, Sara. 2021. "Machine Learning, Explained". *MIT Sloan*. Accessed April 19, 2025. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

- Buzzell, A., & Rini, R. 2022. "Doing your own research and other impossible acts of epistemic superheroism". *Philosophical Psychology*, 36(5), 906–930. https://doi.org/10.1080/09515089.2022.2138019

- Campbell, Charlie. 2024. "Gen AI Has Already Taken the World by Storm. Just Wait Until It Gets a Quantum Boost." *TIME*, Accessed May 13, 2024. https://time.com/6977355/generative-ai-quantum-computing-us-china-technology/

- Cao, Yihan, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 2023. "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT". *arXiv*. https://doi.org/10.48550/arXiv.2303.04226

- Chazan, David. 2025. "France Accuses Russia Over AI Video of Brigitte Macron Sex-abuse Claims." *The Times*, Accessed May 20, 2025. https://www.thetimes.com/world/europe/article/brigitte-macron-teacher-russia-fake-video-q85t36sf6

- Coeckelbergh, M. 2023. "Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence". *AI Ethics* 3, 1341–1350. https://doi.org/10.1007/s43681-022-00239-4

- Copeland B.J. 2025. "Artificial Intelligence (AI) | Definition, Examples, Types, Applications, Companies, & Facts." *Encyclopedia Britannica*. Accessed April 12, 2025. https://www.britannica.com/technology/artificial-intelligence/Methods-and-goals-in-AI

- Engel, P. 2013. "IS EPISTEMIC AGENCY POSSIBLE?" *Philosophical Issues*, *23*, 158–178. http://www.jstor.org/stable/26382647

- Epstein, B. 2020 "Why It Is So Difficult to Regulate Disinformation Online". In: Bennett, W.L. and Livingston, S., Eds., *The Disinformation Age*, Cambridge University Press, 190-210. https://doi.org/10.1017/9781108914628.008

- Feldman, R., and Conee, E. 1985. "Evidentialism". *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 48(1), 15–34. http://www.jstor.org/stable/4319765

- Poston, Ted. n.d. "Foundationalism". *Internet Encyclopedia of Philosophy*. Accessed April 26, 2025 https://iep.utm.edu/foundationalism-in-epistemology/

- Gettier. Edmund L. 1963. "Is Justified True Belief Knowledge?", *Analysis*, Volume 23, Issue 6, Pages 121–123, https://doi.org/10.1093/analys/23.6.121

- Goldman, A.I. 1979. "What is Justified Belief?". In: Pappas, G.S. (eds) Justification and Knowledge. *Philosophical Studies Series in Philosophy*, vol 17. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-9493-5_1

- Harris, K.R. 2024. "AI or Your Lying Eyes: Some Shortcomings of Artificially Intelligent Deepfake Detectors". *Philos. Technol.* 37, 7. https://doi.org/10.1007/s13347-024-00700-8

- Ichikawa, Jonathan Jenkins and Steup, Matthias. 2024 "The Analysis of Knowledge", *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), Accessed April 18, 2025 https://plato.stanford.edu/archives/fall2024/entries/knowledge-analysis

- Independent High-Level Expert Group on Artificial Intelligence. 2019. "A Definition of AI: Main Capabilities and Scientific Disciplines". Brussels: *European Commission*. https://digital-strategy.ec.europa.eu/fr/node/2226

- Ivanova, Yordanka and European Commission. 2021. "Generative AI and the AI Act." https://www.europarl.europa.eu/cmsdata/277774/02.YordankaIvanova.pdf

- Jaidka, Kokil, Tsuhan Chen, Simon Chesterman, Wynne Hsu, Min-Yen Kan, Mohan Kankanhalli, Mong Li Lee, Gyula Seres, Terence Sim, Araz Taeihagh, Anthony Tung, Xiaokui Xiao, and Audrey Yue. 2025. "Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy." *Digital Government: Research and Practice*6 (1): Article 11. https://doi.org/10.1145/3689372

- Landymore, Frank. 2025. "An AI Model Has Officially Passed the Turing Test." *Futurism*, Accessed April 29, 2025. https://futurism.com/ai-model-turing-test

- Microsoft. n.d. "How Does Generative AI Work?" *AI 101*, accessed April 8, 2025, https://www.microsoft.com/en-us/ai/ai-101/how-does-generative-ai-work

- Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A. M., Krumhuber, E. G., & Dawel, A. 2023. "AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones". *Psychological Science*, *34*(12), 1390-1403. https://doi.org/10.1177/09567976231207095

- Mucci, Tim. 2024. "History of artificial intelligence." *IBM*. Accessed April 14, 2025. https://www.ibm.com/think/topics/history-of-artificial-intelligence

- Neuberger, C., Bartsch, A., Fröhlich, R., Hanitzsch, T., Reinemann, C., & Schindler, J. 2023. "The digital transformation of knowledge order: a model for the analysis of the epistemic crisis". *Annals of the International Communication Association*, 47(2), 180–201. https://doi.org/10.1080/23808985.2023.2169950

- Nieminen, J.H., Ketonen, L. 2024. "Epistemic agency: a link between assessment, knowledge and society". *High Educ* 88, 777–794. https://doi.org/10.1007/s10734-023-01142-5

- Nightingale S.J. & Farid. 2022. H. “AI-synthesized faces are indistinguishable from real faces and more trustworthy”, *Proc. Natl. Acad. Sci. U.S.A.* 119 (8) e2120481119, https://doi.org/10.1073/pnas.2120481119

- Pappas, George, "Internalist vs. Externalist Conceptions of Epistemic Justification", The Stanford Encyclopedia of Philosophy (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), Accessed June 2, 2025. https://plato.stanford.edu/archives/spr2023/entries/justep-intext/

- Reuters Institute. 2024. *Digital News Report*. Accessed May 22, 2025. https://www.digitalnewsreport.org

- Rini, Regina 2020. “Deepfakes and the Epistemic Backstop”. *Philosophers' Imprint* 20 (24):1-16. http://hdl.handle.net/2027/spo.3521354.0020.024

- Sample, Ian. 2020. “What Are Deepfakes – and How Can You Spot Them?” *The Guardian*, Accessed April 15, 2025. https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them

- Saltiel, François. 2025. “Google Veo 3 Ou La Menace Du Chaos Informationnel.” *France Culture*, Accessed May 28, 2025. https://www.radiofrance.fr/franceculture/podcasts/un-monde-connecte/google-veo-3-ou-la-menace-du-chaos-informationnel-5531960

- Sheldon, Robert, and Kerner, Sean Michael. 2024. “What Is Dall-E and How Does It Work?” *Search Enterprise AI.* Accessed April 18, 2025. https://www.techtarget.com/searchenterpriseai/definition/Dall-E

- Shu, Ruiyang, et al. 2024. *“*Variational Quantum Circuits Enhanced Generative Adversarial Network*.” arXiv* preprint arXiv:2402.01791. https://arxiv.org/abs/2402.01791

- Stryker, Cole, and Mark Scapicchio. 2025. “Generative AI.” *IBM*. Accessed April 11, 2025. https://www.ibm.com/think/topics/generative-ai

- Suiter, J. 2016. “Post-truth Politics”. *Political Insight*, 7(3), 25-27. https://doi.org/10.1177/2041905816680417

- Swinburne, R. 2001. Epistemic justification. Clarendon Press.

- Wakefield, By Jane. 2022. “Deepfake Presidents Used in Russia-Ukraine War.” *BBC.* Accessed May 14, 2025. https://www.bbc.com/news/technology-60780142

- Watson, Jamie Carlin. n.d. "Epistemic Justification." In Internet Encyclopedia of Philosophy. Accessed April 24, 2025. https://iep.utm.edu/epi-just/

- Yasar, Kinza, Nick Barney, and Ivy Wigmore. 2025. “What Is Deepfake Technology?” *WhatIs*. Accessed May 22, 2025. https://www.techtarget.com/whatis/definition/deepfake

- Zahavi, Dan. 2005 *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Cambridge, MA: *MIT Press.*

• Zewe, Adam. 2023. “Explained: Generative AI.” MIT News | Massachusetts Institute of Technology. November 9, 2023. Accessed April 11, 2025. https://news.mit.edu/2023/explained-generative-ai-1109