

Tomasz Stępień

@ t.stepien@uksw.edu.pl

ID <https://orcid.org/0000-0003-2422-1812>

Cardinal Stefan Wyszyński University in Warsaw, Poland

ORCID <https://ror.org/05sdyjv16>

***Confusa Imago* and Artificial Intelligence: Peter Abelard's Theory of Abstraction as a Framework for Understanding the Cognitive Status of AI Systems**

DOI <https://doi.org/10.15633/pch.16104>

Abstract

This article addresses a philosophical problem arising from recent Vatican documents on artificial intelligence: how can AI systems demonstrate functional generalization while lacking the capacity for abstraction proper to human cognition? Drawing on the medieval philosophy of Peter Abelard (1079–1142), particularly his theory of the *confusa imago* (confused image) developed in the context of the twelfth-century dispute over universals, this article proposes a novel framework for understanding the cognitive status of contemporary AI systems. Abelard's distinction between the representational level (*confusa imago*—an indeterminate mental image common to all instances of a kind) and the intellectual level (*intellectus*—the conscious act that transforms representation into genuine understanding) maps with striking precision onto the functional architecture of modern artificial neural networks. The article argues that AI systems achieve something structurally analogous to the generation of confused images through statistical learning, without performing the intellectual acts that would transform these representations into authentic concepts. This framework resolves the apparent paradox of generalization without abstraction, provides a principled basis for the claim that human cognitive involvement remains structurally necessary (not merely contingently useful) in human–AI collaboration, and offers concrete implications for the ethical design and deployment of AI systems. The analysis demonstrates that medieval philosophical categories retain unexpected explanatory power for contemporary problems in the philosophy of artificial intelligence.

Keywords

abstraction, artificial intelligence, cognitive science, *confusa imago*, human, Middle Ages, mind, Peter Abelard, philosophy, universals

On January 28, 2025, the Dicastery for the Doctrine of the Faith and the Dicastery for Culture and Education of the Holy See published a joint document entitled *Antiqua et Nova: Note on the Relationship between Artificial Intelligence and Human Intelligence* (Antiqua et Nova 2025). This document, the first comprehensive statement of the Catholic Magisterium on the philosophical and theological implications of artificial intelligence, advances a claim that poses a significant problem for the philosophy of AI. The authors assert that AI “does not account for the full breadth of human experience, which includes abstraction, emotions, creativity, and the aesthetic, moral, and religious sensibilities” (Antiqua et Nova 2025, §11). The document elaborates: “AI’s advanced features give it sophisticated abilities to *perform tasks*, but not the ability to *think*” (Antiqua et Nova 2025, §12).

The puzzle emerges from the apparent tension between this claim and the observable capacities of contemporary AI systems. If AI lacks abstraction — understood in the traditional philosophical sense as the cognitive operation by which the mind extracts universal concepts from particular sensory data — then how do AI systems achieve the functional generalization they manifestly exhibit? A language model can respond appropriately to sentences it has never encountered; a vision system can classify images absent from its training data; a game-playing agent can discover strategies not present in its learning corpus. These capacities appear to require some form of generalization beyond particular instances.

The present article proposes that one of the possible resolutions of this puzzle lies in a distinction developed by Peter Abelard in his treatment of the medieval problem of universals. Abelard distinguished between two cognitive levels: the *confusa imago* (confused or indeterminate image) — a representation that is “proper to none and common to all” instances of a kind — and the *intellectus* (intellectual act) — the fully conscious apprehension that transforms mere representation into genuine understanding. This article argues that contemporary AI systems realize something structurally analogous to the generation of *confusae imagines* through statistical learning processes, while failing to perform the *intellectus* that would constitute authentic cognition.

This framework makes it possible to obtain several significant results. First, it can explain how functional generalization is possible without abstraction in the full philosophical sense. Second, it can provide a philosophical basis for the claim that human cognitive involvement in AI systems is not a contingent

limitation to be overcome by technological progress, but a structural necessity rooted in the nature of cognition itself and the very structure of how AI models function. Third, it can offer concrete guidance for the ethical design and deployment of AI systems. Finally, it demonstrates that medieval philosophical categories, developed in response to apparently distant problems, retain unexpected explanatory power for contemporary challenges.

1. The Medieval Problem of Universals: Historical and Philosophical Context

The problem of universals — one of the most persistent questions in medieval philosophy — concerns the ontological and epistemological status of general terms and the concepts they express (de Libera 1996, 39–49). When we predicate “human” like Socrates and Plato, to what does this common term refer? Is there a universal “humanity” that both individuals instantiate? If so, what is its ontological status? If not, what grounds the common predication? The medieval debate took its specific form from three questions posed by the ancient philosopher Porphyry in his *Isagoge*, an introduction to Aristotle’s *Categories*: (1) Do genera and species exist in reality or only in thought? (2) If they exist, are they corporeal or incorporeal? (3) Do they exist separately from sensible particulars, or in them? (*Isagoge*, Ia, 9–13; trans. 1975, 27–28). Porphyry himself declined to answer these questions, judging them too difficult for an introductory work. For medieval philosophers, however, this was a living issue since Porphyry’s text was required reading for students of philosophy at the beginning of their education.

The dominant position in the early twelfth century was realism — the view that universals possess real existence in the particulars that instantiate them, independent of the minds that cognize them. William of Champeaux, Abelard’s teacher at the cathedral school of Notre-Dame in Paris, initially defended a strong version of this position: individuals belonging to the same species share a numerically identical substantial essence. On this view, Socrates and Plato possess exactly the same humanity, but they differ only in their accidents (King 1982, 137–173).¹

¹ William of Champeaux presented this theory in his lectures at Notre-Dame around 1108–1109. The primary source for his views is Abelard’s critique of that theory.

Abelard's critique of this position, preserved in his *Logica Ingredientibus*, was devastating (*Logica Ingredientibus*, 10–11).² The central objection runs as follows: if the universal “animal” is wholly and numerically identically present in Socrates (who is rational) and in a donkey (which is irrational), then the same thing would possess contradictory properties: it would be simultaneously rational and irrational. Furthermore, the theory generates a vicious regress: the accidents supposed to individuate the common essence are themselves universals requiring further individuation (*Logica Ingredientibus*, 11–12).

Under the pressure of Abelard's critique, William modified his position, adopting what is known as the “indifference theory.” It holds that individuals are not identical in essence but are called “the same” because they “do not differ” (*non differunt*) in certain respects (Marenbon 1997, 108–114). Abelard found this reformulation equally unsatisfactory. As he observed, Plato and Socrates also “do not differ in being a stone” — neither is a stone, so they are identical in that respect. Mere non-difference cannot ground positive similarity or justify common predication (*Logica Ingredientibus*, 12).

After defeating his teacher, Abelard was able to present his own solution. Significantly for the present argument, he, like his defeated opponent before him, had to do so without the crucial source that explained the meaning of abstraction, since Aristotle's *De anima* had not yet been translated into Latin and was unknown to Western philosophical circles. He therefore had to develop an account of abstraction from scratch (Gilson 1950, 9–10).³

His central thesis is that universality is a property of certain linguistic expressions — specifically, common nouns and the concepts associated with them — not something that exists in reality. Only words can be universals in the strict sense, because only words are predicable of many things (King 1982, 352–354). However, this linguistic thesis requires supplementation. What grounds the common application of universal terms? Why is it correct to call

² The critical edition is: Peter Abelard, *Logica Ingredientibus (Glossae super Porphyrium)*, ed. B. Geyer, in *Peter Abaelards Philosophische Schriften*, Beiträge zur Geschichte der Philosophie des Mittelalters 21 (Münster: Aschendorff, 1919–1933). An English translation of key passages is provided in Spade (1994, 32–33).

³ See Gilson's account (1950, 9–10) of the whole debate between William of Champeaux and Peter Abelard, including his commentaries on Abelard's limitations. In Gilson's view, the absence of Aristotle's *De anima* paradoxically precipitated Abelard's originality: “In short, Abailard was in that blessed state of ignorance which makes it so easy for a clever man to be original.”

both Socrates and Plato “human”? Here, Abelard introduces his distinctive concept of *status* — the state or condition of being a certain kind of thing. The *status* of being human is not itself a thing (*res*) or essence (*essentia*); Abelard emphatically denies that *status* possesses substantial existence. Rather, *status* functions as “a cause that is not a thing” — it explains the commonality without postulating a common metaphysical element (King 1982, 354–374). In *Logica Ingredientibus*, he claims that when we say that Socrates and Plato agree in the status of man (*in statu hominis*), we refer to their *status* although we are not appealing to anything (*Logica Ingredientibus*, 20, 7–9).⁴

The concept of *status* has an unusual ontological profile: it is sufficiently robust to ground linguistic imposition and common predication, yet it is not an additional entity in the world. Abelard thinks that statuses can be understood as causes, just as not wanting to go to the forum can be the cause of being flogged. As Abelard states, *not wanting to go* is “no essence.” *Status thus denotes* a cause that is not itself a thing, so it functions causally without being a substance (*Logica Ingredientibus*, 20, 10–17).⁵

2. The Theory of *Confusa Imago*: Abelard’s Cognitive Psychology

The psychological complement to the semantic notion of *status* is Abelard’s account of the mental representations associated with universal terms. When we hear a common noun like “human,” what arises in the mind? Since Abelard denies that there are universal things for such terms to represent, the mental content cannot be a representation of a universal entity. Yet it must somehow support the universal application of the term. Abelard’s answer is the *communis et confusa imago* — the common and confused image. In *Glossae super Porphyrium* he distinguishes understanding of universal and particular objects:

⁴ “Sed nihil aliud sentimus, nisi eos homines esse, et secundum hoc nullatemus differre, secundum hoc, inquam quod homines sunt, licet ad nullam vocemus essentiam. Statum autem hominis ipsum esse hominem, quod non est res, vocamus, quod etiam diximus communem causam impositionis nominis ad singulos, secundum quod ipsi ad invicem conveniunt.”

⁵ “Saepe autem cause nominae ea quoque quae res aliqua non sunt, appellamus, ut cum dicitur: Verberatus est, quia non vult ad forum. Non vult ad forum, quod ut causa ponitur, nulla est essentia.” See also Gilson (1950, 23).

...the understanding that goes with a universal name conceives a common and confused image of many things. But the understanding a singular word generates comprises the proper and so to speak “singular” form of one thing — that is, a form related to one “person” only. Thus, when I hear ‘man,’ a kind of model rises in my mind that is related to a single man in such a way that it is common to all of them and proper to none. But when I hear ‘Socrates,’ a certain form rises up in my mind that expresses the likeness of a certain person (*Logica In-gredientibus*, 21, 28–33; trans. Spade 1994, 44).⁶

The term *confusa* requires careful interpretation. It does not mean “unclear” or “muddled” in a pejorative sense. Rather, it carries a technical meaning: the image is indeterminate with respect to individuating features. The term “human” captures what humans have in common — rationality, mortality, animal nature — while lacking the specific determinations (this particular height, this hair color, this spatial location) that would make it an image of Socrates rather than Plato (Marenbon 1997, 145–160).

We can illuminate this concept by a contemporary analogy. Consider an anatomical illustration of the human body in a medical atlas. Such an illustration does not depict any particular person — it lacks a specific face, a determinate skin color, individual proportions. Yet it represents all human beings; it displays the anatomical structure common to the species. In Abelard’s terms, it is “proper to none and common to all.” The *confusa imago* functions analogously: it is a mental representation that captures what is common while remaining indeterminate with respect to what it individuates.

In *Tractatus de Intellectibus*, Abelard further characterizes the concepts associated with universal terms using three technical descriptors (*Tractatus de Intellectibus*, 3.19–3.24; King 2007, 169–187). The *imago* is *sola* (alone), detached from current sense perception of any particular individual. It operates independently of the presence of any specific instance. It is *nuda* (bare) abstracted from some or all accidental forms. It retains only what is essential or common, stripped

⁶ “...quod ille qui universalis nominis est, commune et confusam imaginem multorum concipit, ille vero quem vox singularis generat, propriam unius et quasi singularem formam tenet, hoc est ad unam tantum personam se habentem. Unde cum audio ‘homo,’ quoddam instar in animo surgit, quod ad singulos homines sic se habet, ut omnium sit commune et nullius proprium. Cum autem audio ‘Socrates,’ forma quaedam in animo surgit, quae certae personae similitudinem exprimit.”

of individualizing accidents. And, finally, it is *pura* (pure) — unadulterated by reference to any single individual. It maintains its indeterminacy without collapsing onto a particular instance. These characteristics jointly describe a representation that enables universal predication precisely because of its indeterminacy. Because the *confusa imago* does not single out any particular individual, it can serve as the cognitive basis for applying a term to all individuals of the relevant kind.

What is significant in the context of this article is that Abelard insists that all cognition starts with sensation, which is the only source of data. As Kevin Guilfooy explains the mind has an initial confused conception of a particular discrete item thanks to the powers of sensation and imagination. This image is confused because the mind is merely conscious of the item, and the whole process of cognition is a kind of perfecting this confused image of sensation and imagination (Guilfooy 2004, 203). It is worth noticing that the first part of this process — sensory perception — is for Abelard purely mechanical (Marenbon 1997, 167).

The second part, which is imagination, is more perfect than the first, but it is still considered to be an imperfect type of cognition. Its primary function is to produce images, which are “mere preparation for thought” (Marenbon 1997, 168). The confused image is the *content* upon which cognition operates; the *intellectus* is the *act* of conscious apprehension that transforms mere representation into genuine understanding (King 1982, ch. 4). As J. Marenbon underlines: “thoughts are sharply distinct from images of any kind” and therefore thought “necessarily involves abstraction” (Marenbon 1997, 168).

This distinction has several important implications. First, Abelard explicitly rejects the view that mental images by themselves explain intentionality, the property of being *about* something. Images are merely “intermediary signs of things” (*intersigna rerum*) (*Logica Ingredientibus*, 23; King 2007, 180–183). The same image can be used to think about different objects depending on which features receive attention. Intentionality derives from the act of attention (*attentio*), not from intrinsic properties of the image. Second, the *intellectus* involves normative orientation.

It is not merely a causal response to stimuli, but an act directed toward grasping the nature of things, toward understanding what something is. This normative dimension distinguishes genuine cognition from mere information processing. Third, the *intellectus* is phenomenally conscious. It may be described as the proper and most perfect cognitive act, while sensation is mechanical and imagination is imperfect and blurred. There is something it is like to understand,

a qualitative character to the act of intellectual apprehension, and this dimension is essential to cognition itself. Thus, the confused image is a necessary but not sufficient condition for understanding. It provides the representational raw material, but without the *intellectus*, there is no genuine cognition.

3. AI as Generator of *Confusae Imagines*

AI systems realize something structurally analogous to the generation of *confusae imagines* through statistical learning processes, while failing to perform the *intellectus* that would transform these representations into authentic concepts. Accordingly, the Vatican document correctly asserts that AI lacks abstraction in the full philosophical sense — the sense that involves the *intellectus* as a conscious, normatively oriented act of apprehension. Yet AI manifestly generalizes. The resolution lies in recognizing that generalization at the representational level does not require intellectual apprehension. AI achieves the former without the latter.

Consider the training process of a contemporary deep artificial neural network. The system is exposed to millions of particular instances — specific images of cats, specific sentences in a natural language, and specific game states and outcomes. Through iterative optimization (typically via gradient descent), the network develops internal representations — patterns of activation across hidden layers — that capture statistical regularities in the training data. As LeCun, Bengio, and Hinton observe in their foundational review: “Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction” (LeCun, Bengio, and Hinton 2015, 436).⁷

The concept of *distributed representation*, introduced by Hinton, McClelland, and Rumelhart in their seminal 1986 work, provides the theoretical foundation for understanding how artificial neural networks develop these abstract patterns (Hinton et. al. 1986, 77–109). In a distributed representation, concepts are not represented by single dedicated units but by patterns of activity across many

⁷ This landmark review provides a comprehensive overview of how deep artificial neural networks learn hierarchical representations of data. It is worth noting that the term “abstraction” as used here describes generalization, and future development of AI is predicted to involve “complex reasoning” (LeCun, Bengio, and Hinton 2015, 442).

units. Each unit participates in representing many different concepts, and each concept is represented by the combined activity of many units. This architecture enables what Hinton termed “automatic generalization”: when the weights in the network change to incorporate new knowledge about one concept, the changes affect knowledge associated with other concepts represented by similar activation patterns (Hinton 1986, 1–12).

These learned representations exhibit a characteristic property: they are indeterminate with respect to particular training instances. The representation of “cat” that emerges in a trained vision model does not correspond to any specific cat from the training set. It is an abstract pattern that captures what cats statistically have in common — typical shapes, textures, proportions — while remaining indeterminate with respect to the individualizing features of particular cats. Zeiler and Fergus’s influential visualization studies of convolutional neural networks demonstrate this phenomenon empirically. Higher layers of trained networks develop feature detectors that respond to increasingly abstract and general patterns, moving from edges and textures in early layers to object parts and whole objects in deeper layers (Zeiler and Fergus 2014, 818–833).⁸

We can observe a parallel between the representations developed by such systems and what Abelard calls images that are “proper to none and common to all.” The parallel extends to the functional role of indeterminacy. In Abelard’s theory, the “confusion” of the image — its lack of individuating specificity — is precisely what enables its universal application. Because it singles out no particular, it applies to all. Similarly, the general character of learned representations in artificial neural networks (their distance from any specific training instance) is what enables generalization. The network succeeds on novel inputs precisely because its internal representations have abstracted away from particular training examples.

The parallel, however, stops at the representational level. Contemporary AI systems do not perform anything analogous to the *intellectus* — the conscious, normatively oriented act of intellectual apprehension. To support this claim, it is worth returning to John Searle’s classical Chinese Room argument (Searle 1980, 417–457). He argued that a system manipulating symbols according to formal rules — however sophisticated — performs merely syntactic operations without semantic understanding. As Searle put it: “Formal symbol manipulations

⁸ This study introduced deconvolutional network techniques to visualize the features learned at each layer of a CNN, demonstrating the increasingly abstract and general nature of representations in deeper layers.

by themselves don't have any intentionality; they are quite meaningless; they aren't even symbol manipulations, since the symbols don't symbolize anything. In the linguistic jargon, they have only a syntax but no semantics" (Searle 1980, 422).⁹ While debates about the Chinese Room continue, Searle's core insight — that rule-governed symbol manipulation does not by itself constitute understanding — resonates with Abelard's distinction between possessing an image and performing an intellectual act upon it.

More recently, Bender, Gebru, and colleagues have applied similar reasoning to large language models, characterizing them as "stochastic parrots" that "haphazardly stitch together sequences of linguistic forms... according to probabilistic information about how they combine, but without any reference to meaning" (Bender et al. 2021, 616). This characterization aligns precisely with the present analysis: language models generate representations (analogous to *confusae imagines*) that enable statistically appropriate outputs without performing the semantic grasp that constitutes the *intellectus*.

There is no evidence that AI systems possess phenomenal consciousness, that there is "something it is like" to be a language model processing a sentence or a vision system classifying an image. The *intellectus* in Abelard's framework essentially involves phenomenal presence; the act of understanding is not a mere functional state but a conscious event (Chalmers 1996, chaps. 1–2). Moreover, AI systems optimize objective functions defined by their designers. They do not pursue truth as such or aim at grasping the natures of things. The *intellectus* involves a normative orientation toward understanding that is absent from systems defined purely by loss minimization. If this analysis is correct, such systems manipulate representations that track statistical regularities, but they do not *understand* what those representations mean. A language model processes tokens in ways that produce statistically appropriate outputs without grasping the semantic content of the sentences it generates. The *intellectus* essentially involves semantic understanding, not merely syntactic manipulation.

The framework derived from Abelard's cognitive psychology seems useful for explaining several otherwise puzzling features of contemporary AI systems. They frequently produce correct outputs (accurate classifications, appropriate responses, valid solutions) without understanding why these outputs are correct.

⁹ For a comprehensive analysis of the argument, see the Stanford Encyclopedia of Philosophy, s.v. "The Chinese Room Argument," <https://plato.stanford.edu/entries/chinese-room/> (accessed February 19, 2026).

This is explicable on the present account: the confused images and their statistically determined relations can accurately track regularities in the world without the *intellectus* that would constitute insight into the reasons.

AI systems generalize on the basis of statistical similarity rather than essential features. They may classify correctly by detecting accidental correlations (texture rather than shape, spurious features rather than causally relevant ones) (Goodfellow, Shlens, and Szegedy 2014).¹⁰ Similarly, Geirhos and his colleagues demonstrated empirically that ImageNet-trained convolutional neural networks classify images primarily on the basis of texture rather than shape — an instance of generalization based on accidental rather than essential features (Geirhos et al. 2019). This follows from operating at the representational level without intellectual grasp of natures or essences. Finally, when AI systems are prompted for explanations, they produce post hoc rationalizations rather than genuine justifications. They lack the “inner light” of understanding that would ground authentic reasons. This is precisely what the absence of *intellectus* predicts. As Lipton observes in his analysis of model interpretability, there is a fundamental distinction between a system’s capacity to produce outputs and its capacity to explain *why* those outputs are appropriate — a distinction that maps onto Abelard’s image/*intellectus* divide (Lipton 2018, 31–57).¹¹

Antiqua et Nova states (§35) that “AI should not be seen as an artificial form of human intelligence but as a product of it” (*Antiqua et Nova* 2025, § 35). It seems that now we can support this claim with Abelard’s framework and even make this thought more precise: AI is an instrument for generating representations analogous to confused images. These representations are a product of human intelligence and require human intelligence to be transformed into knowledge.

¹⁰ This demonstrates that neural networks can be deceived by imperceptible perturbations—evidence that they rely on features quite different from those humans consider essential.

¹¹ Lipton distinguishes between transparency (understanding how a model works) and post hoc explanation (providing reasons for outputs), arguing that these involve different epistemic relationships.

4. Implications for Human–AI Relationships

The analysis developed above can be seen as having significant implications for understanding the proper relationship between human cognition and artificial intelligence. The first and most fundamental implication concerns the necessity of human cognitive involvement. If AI systems operate at the level of *confusa imago* without achieving the *intellectus*, then the transformation of AI-generated representations into genuine knowledge requires human intellectual activity. This is not a contingent limitation that future technological development might overcome; it is a structural feature of the distinction between representation and understanding. Just as Abelard's *confusa imago* requires the *intellectus* to become a genuine concept, so AI-generated representations require human cognition to become knowledge in the full sense. The human contribution is not merely to oversee, verify, or correct AI outputs; it is to perform the intellectual acts that constitute understanding — grasping meaning, evaluating truth, and comprehending implications.

This analysis enables the rejection of two common frameworks for conceptualizing human–AI relationships. What may be called the “replacement model” holds that AI will progressively substitute for human cognition across an expanding range of domains until it reaches the status of General AI, when replacement will be complete. The present analysis reveals the incoherence of this model since AI and human cognition operate at different levels. The former cannot replace the latter because they accomplish different things. Similarly, one could hold that AI and humans perform the same cognitive tasks with different degrees of efficiency, what may be termed the “competition model.” Our analysis reveals that this model is also inadequate, because the tasks are not the same. Generating a representation that enables correct classification is not the same task as understanding why the classification is correct.

The appropriate model of the human–AI relationship is what may be termed the “instrumental–intellectual model,” where AI serves as a powerful instrument for generating rich representational structures, while humans perform the intellectual acts that confer meaning and cognitive value upon these structures. The relationship is analogous to that between a craftsman and sophisticated tools: the tools extend capabilities, but the exercise of craft (skillful, conscious, purposive activity) remains the domain of the human agent.

The foregoing theoretical analysis yields several practical implications for the design and deployment of AI systems. If human cognition must operate

upon AI-generated representations to transform them into knowledge, then humans require access to those representations in forms amenable to intellectual engagement. “Black box” systems that deliver only final outputs (classifications, recommendations, decisions) deprive humans of the representational material upon which the *intellectus* can operate. As Rudin argues in her influential critique of explainable AI, there is a fundamental difference between explaining opaque models post hoc and designing inherently interpretable models; the former may provide rationalizations without genuine understanding, while the latter enables authentic intellectual engagement (Rudin 2019, 206–215).¹² Interpretability is not merely a desideratum for debugging or accountability; it is a condition for the cognitive partnership that makes AI outputs epistemically valuable. Otherwise, AI outputs cannot be evaluated because truthfulness and coherence can be identified only at a higher level of abstraction.

Another risk emerges from excessive reliance on AI systems as illegible result generators: the atrophy of human intellectual capacities through disuse. If humans habitually accept AI outputs without performing intellectual work upon them (without asking what the outputs mean, how they were produced, whether they are true, and how they describe the nature of things) the very capacities that make those outputs meaningful may deteriorate. This concern has been documented empirically across multiple domains. Studies of automation in aviation, for example, have demonstrated that pilots who rely heavily on automated systems show degraded manual flying skills and reduced situational awareness when automation fails (Casner et al. 2014, 1506–1516).¹³ Similar patterns have been observed in medical diagnosis, where doctors using AI-assisted systems show decreased detection rates when AI support is withdrawn (Dratsch et al. 2023).¹⁴ Vallor has articulated the broader concern as “moral deskilling” — the erosion of human capacities for judgment through delegation to automated systems (Vallor 2015, 107–124). The risk is not that AI will replace human understanding but that humans will cease to exercise, and thereby gradually lose, the capacity to understand.

¹² Rudin argues that post hoc explanations of black box models may be misleading and advocates for inherently interpretable models in high-stakes domains.

¹³ See also Federal Aviation Administration, *Safety Alert for Operators 13002* (2013), which recommends increased manual flight practice due to concerns about skill degradation.

¹⁴ This study found that radiologists’ performance declined when AI assistance was incorrect, with less experienced practitioners showing greater susceptibility to automation bias.

Finally, authentic judgment (epistemic or even ethical) requires intellectual apprehension, not merely reliable differential response. In Abelard's framework, ethical values and principles are not objects that can be placed in the domain of *confused images*. For an act of judgment or an act of conscience, it is necessary to know and understand general rules that can be obtained and processed only at the intellectual level of abstraction. Although it may seem to involve making a simple connection between rules and cases that can be associated with them, understanding general rules is necessary for the very ability to make such connections. For example, in ethical judgment, to claim that drug use is wrong one must have a deeper understanding of "Thou shalt not kill," and such understanding cannot be obtained from probabilistic analysis from even an enormously large number of cases of humans who intentionally damage their health.

Since AI systems do not perform intellectual acts, they cannot bear the epistemic responsibility that attaches to claims of knowledge or the moral responsibility that attaches to practical decisions. Humans must remain the locus of responsibility because they are the locus of understanding (Matthias 2004, 175–183). This has implications for the deployment of AI in high-stakes contexts: human judgment must remain genuinely engaged, not merely nominally present. The phenomenon of "automation bias" — the tendency to defer uncritically to automated recommendations — represents precisely the failure to exercise the *intellectus* that would evaluate and eventually transform AI outputs into genuine knowledge (Goddard, Roudsari, and Wyatt 2014, 368–375; Skitka, Mosier, and Burdick 1999, 991–1006).

Conclusions

This article has proposed that Peter Abelard's twelfth-century theory of the *confusa imago* provides a precise and illuminating framework for understanding the cognitive status of contemporary artificial intelligence systems. He distinguished between two cognitive levels — the *confusa imago* (an indeterminate representation generated by processing particular experiences through sensation and imagination) and the *intellectus* (the conscious act that transforms representation into genuine understanding). Through their statistical learning processes, contemporary AI systems generate representations structurally analogous to *confusae imagines*, which are indeterminate patterns capturing

what instances have in common while remaining uncommitted to particulars. However, AI systems do not perform the *intellectus* — the conscious, normatively oriented act of intellectual apprehension. They achieve the representational level without ascending to the intellectual one.

This framework resolves the central problem posed by *Antiqua et Nova*. That document, by asserting that AI systems cannot perform abstraction, appeared to contradict the evident observation that AI systems are capable of performing tasks at a general level. AI generalizes without abstracting because generalization at the representational level does not require the intellectual act that constitutes abstraction in the full philosophical sense. AI generates representational patterns analogous to *confusae imagines* and processes them without genuine understanding.

If the analysis is correct, then human cognitive involvement in AI systems is not a temporary limitation to be engineered away but a permanent necessity rooted in the structure of cognition itself. The *intellectus* cannot be delegated to systems that operate only at the level of *confusa imago*. This provides philosophical grounding for practices of human oversight, interpretability, and epistemic responsibility in AI deployment.

This analysis also demonstrates that philosophical resources developed in response to apparently distant problems, such as the medieval dispute over universals, retain unexpected explanatory value for contemporary challenges (Marenbon 2013, 146–206).¹⁵ Abelard, grappling with questions about how minds that know only particulars can think universally, developed distinctions that illuminate our contemporary perplexity about machines that generalize without understanding.

References

- Abelard, Peter. 1919. "Glossae super Porphyrium." In *Peter Abaelards philosophische Schriften. I. Die Logica "Ingredientibus"*, edited by Bernhard Geyer, 1–109. Beiträge zur Geschichte der Philosophie des Mittelalters 21.1. Münster: Aschendorff.
- Abelard, Peter. 1994. *Des intellections*. Edited and translated by Patrick Morin. Sic et Non. Paris: Vrin.

¹⁵ The philosophical insights of Peter Abelard have been acknowledged as having explanatory value in many contemporary fields such as the philosophy of language, epistemology, and metaphysics; see especially Marenbon (2013, pt. 3, 146–206).

- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. New York: Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>.
- Casner, Stephen W., Richard W. Geven, Matthias P. Recker, and Jonathan W. Schooler. 2014. "The Retention of Manual Flying Skills in the Automated Cockpit." *Human Factors* 56 (8): 1506–1516. <https://doi.org/10.1177/0018720814535628>.
- Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Philosophy of Mind. New York: Oxford University Press.
- Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education. 2025. *Antiqua et nova: Note on the Relationship between Artificial Intelligence and Human Intelligence*. January 28, 2025. https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_en.html.
- Dratsch, Thomas, Xue Chen, Mohammad Rezazade Mehrizi, Roman Kloeckner, Aline Mähringer-Kunz, Michael Püsken, Bettina Baeßler, Stephanie Sauer, David Maintz, and Daniel Pinto dos Santos. 2023. "Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance." *Radiology* 307 (4): e222176. <https://doi.org/10.1148/radiol.222176>.
- Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. "ImageNet-Trained CNNs Are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness." In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Bygh9j09KX>.
- Gilson, Étienne. 1950. *The Unity of Philosophical Experience*. New York: Charles Scribner's Sons.
- Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. 2014. "Automation Bias: Empirical Results Assessing Influencing Factors." *International Journal of Medical Informatics* 83 (5): 368–375. <https://doi.org/10.1016/j.ijmedinf.2014.01.001>.
- Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2014. "Explaining and Harnessing Adversarial Examples." arXiv:1412.6572. <https://arxiv.org/abs/1412.6572>.
- Guilfooy, Kevin. 2004. "Mind and Cognition." In *The Cambridge Companion to Abelard*, edited by Jeffrey E. Brower and Kevin Guilfooy, 200–222. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CCOL0521772478.007>.
- Hinton, Geoffrey E. 1986. "Learning Distributed Representations of Concepts." In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, 1–12. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hinton, Geoffrey E., James L. McClelland, and David E. Rumelhart. 1986. "Distributed Representations." In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Vol. 1, *Foundations*, edited by David E. Rumelhart, James L. McClelland, and the PDP Research Group, 77–109. Cambridge, MA: MIT Press.
- King, Peter O. 1982. "Peter Abailard and the Problem of Universals in the Twelfth Century." PhD diss., Princeton University.

- King, Peter. 2007. "Abelard on Mental Language." *American Catholic Philosophical Quarterly* 81 (2): 169–187. <https://doi.org/10.5840/acpq200781236>.
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–444. <https://doi.org/10.1038/nature14539>.
- Libera, Alain de. 1996. *La querelle des universaux: De Platon à la fin du Moyen Âge*. Des travaux. Paris: Éditions du Seuil.
- Lipton, Zachary C. 2018. "The Mythos of Model Interpretability." *ACM Queue* 16 (3): 31–57. <https://doi.org/10.1145/3236386.3241340>.
- Marenbon, John. 1997. *The Philosophy of Peter Abelard*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511582714>.
- Marenbon, John. 2013. *Abelard in Four Dimensions: A Twelfth-Century Philosopher in His Context and Ours*. The Conway Lectures in Medieval Studies. Notre Dame, IN: University of Notre Dame Press.
- Matthias, Andreas. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (3): 175–183. <https://doi.org/10.1007/s10676-004-3422-1>.
- Porphyry. 1887. "Isagoge." In *Porphyrii Isagoge et in Aristotelis Categorias Commentarium*, edited by Adolf Busse, 1–22. Commentaria in Aristotelem Graeca 4.1. Berlin: Georg Reimer.
- Porphyry. 1975. *Isagoge*. Translated, with an introduction and notes, by Edward W. Warren. Mediaeval Sources in Translation 16. Toronto: Pontifical Institute of Mediaeval Studies.
- Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–215. <https://doi.org/10.1038/s42256-019-0048-x>.
- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–424. <https://doi.org/10.1017/S0140525X00005756>.
- Skitka, Linda J., Kathleen L. Mosier, and Mark Burdick. 1999. "Does Automation Bias Decision-Making?" *International Journal of Human-Computer Studies* 51 (5): 991–1006. <https://doi.org/10.1006/ijhc.1999.0252>.
- Spade, Paul Vincent, trans. and ed. 1994. *Five Texts on the Mediaeval Problem of Universals*. Indianapolis: Hackett.
- Vallor, Shannon. 2015. "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character." *Philosophy & Technology* 28 (1): 107–124. <https://doi.org/10.1007/s13347-014-0156-9>.
- Zeiler, Matthew D., and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Networks." In *Computer Vision — ECCV 2014*, edited by David Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, 818–833. Lecture Notes in Computer Science 8689. Cham: Springer. https://doi.org/10.1007/978-3-319-10590-1_53.