Reverse-Engineering Understanding: Competence through Predictive Compression

Matthieu Queloz[1] and Pierre Beckmann[2,3]

[1]University of Bern, Department of Philosophy
[2]École Polytechnique Fédérale de Lausanne (EPFL)
[3]Idiap Research Institute
✉ matthieu.queloz@unibe.ch

Abstract: What makes *understanding* an important cognitive state? And what does having the *concept* of understanding do for us? This paper offers a unifying account of understanding by jointly reverse-engineering the function of both the state and the concept. We argue that we care about understanding because it grounds and predicts *robust competence*: the stable ability to succeed across novel scenarios. Our concept of understanding evolved as an efficient *proxy* to track this elusive property, allowing us to identify who to trust and learn from. This highlights the *sociality* of understanding and how it shapes what *kinds* of understanding we are apt to form. We then argue that understanding is the result of convergent pressures on social agents to predict the world using models that are not only accurate, but also compressed enough to be stored, demonstrated, and transmitted. This allows us to integrate a number of ostensibly competing accounts of understanding. Finally, we show how the forces at the root of human understanding elucidate debates over AI understanding.

Keywords: understanding, sociality, trust, transmission, prediction, compression, cognitive models, information theory, structure-sensitivity, explanation, explainability in AI, AI understanding.

1 Introduction

We intuitively grasp the vast difference between a novice chess player who knows a handful of opening traps and a grandmaster who truly *understands* the underlying structure of the position. But why exactly do we care about this difference? And what constitutes this cognitive achievement we call "understanding" the world?

This question has recently moved to the forefront of both philosophy and computer science. In philosophy, epistemologists have widened their focus beyond knowledge to treat understanding as a distinct cognitive achievement.[1] In computer science, rapid advances in machine learning have reignited

[1]See Zagzebski (1996), Elgin (1996), and Kvanvig (2003) in epistemology, and Friedman (1974), Schurz and Lambert (1994), and De Regt and Dieks (2005) in the philosophy of science.

debates over the nature and prospects of AI understanding.[2]

This has given rise to two largely disconnected literatures about understanding which link it with a dizzyingly diverse array of characteristics. Philosophical accounts have taken understanding to consist in "grasping connections" (Belkoniene, 2023; Grimm, 2011; Kvanvig, 2018; Riggs, 2003; Wittgenstein, 1953), especially *explanatory* connections (Khalifa, 2017; Strevens, 2013), or in the possession of certain *abilities*, such as the ability to handle novel and counterfactual cases (Grimm, 2011; Hills, 2015), make reliable predictions (De Regt, 2017), and give explanations (De Regt, 2009; Hills, 2015).[3] By contrast, computer science accounts tend to equate understanding with "compression" (Chaitin, 2002, 2006; Delétang et al., 2024; Hutter, 2005; Li et al., 2024; Maguire et al., 2015; Ramstead et al., 2025; Schmidhuber, 2006, 2008, 2010; Wolfram, 2018; Zenil, 2019). There have been surprisingly few attempts to integrate these two literatures.[4] What unifies this heterogeneous collection of characteristics of understanding, and, ultimately, these disconnected debates?

Instead of tackling this question through yet another analysis of what understanding is, this paper hopes to make headway by reframing the entire inquiry around the question of *why* we care about understanding in the first place. Though we ultimately want to address the question of what understanding is, and how to tell whether a cognitive system possesses it, we propose to let our answers to these questions be guided by how they fit into a more comprehensive account of the forces that drove the emergence of understanding and our interest in it. What practical pressures led human agents to be so concerned with understanding at all? The idea is to let the "Why" be a guide to the "What."[5]

But notice that there are really two sides to the issue of why we care about understanding. One is what drove us to *attribute* understanding (to others, in the first instance, but eventually to ourselves as well); the other is what drove us to *acquire* understanding. Significantly, the first question asks primarily after the development and function not of the cognitive *state*, but of the *concept* of understanding. What practical needs led us to form and retain a concept which renders us sensitive to the distinction between "understanders" and "non-understanders"? Only the second question—what drove us to *acquire* understanding—asks after the development and function of the *state* of understanding. What practical pressures encourage, in each generation anew, the formation of the kind of cognitive organization that the concept of understanding picks out and imbues with significance?

Where such questions have been addressed at all, it has tended to be in a one-sided way, focusing either exclusively on the concept or on the state of understanding.[6] A guiding assumption of our inquiry, by contrast, is that these two sides of the issue must be addressed together. This is partly because nothing less will do to explain why we care about understanding. The mere fact that there is plenty of

[2]For an overview, see Mitchell and Krakauer (2023).

[3]For overviews, see Baumberger et al. (2017) and Grimm (2021).

[4]Wilkenfeld (2018) importantly began the work of introducing the conception of understanding as compression into philosophy. Our paper continues that project by connecting the philosophical discussion to the congenial literature in computer science and information theory from which the idea originates.

[5]We draw inspiration from the increasingly popular methodology of reverse-engineering the most basic functions of our concepts and practices (Craig, 1990; Fricker, 2016; Hannon, 2019; Kelley, 2025; Kusch & McKenna, 2020; Lawlor, 2023; Price, 2011; Price et al., 2013; Queloz, 2021; Thomasson, 2025; Williams, 2002).

[6]See, e.g., Woodward (2003), Grimm (2012), Hannon (2019), and Nado (2025).

understanding around does not suffice. Things might have gone with understanding as they went with the subconscious: it was presumably around for centuries, but people lacked any notion of it before Freud. The human concern with understanding, expressed by our longstanding and ubiquitous use of the concept of understanding, calls for explanation in terms of practical pressures on human agents to be conceptually sensitive to the presence of understanding.

Conversely, having the concept of understanding does not suffice to ensure that the cognitive state of understanding actually forms. The equally ubiquitous presence of that state calls for explanation in terms of the practical pressures on human agents to arrive at the kind of cognitive organization that amounts to understanding, and which thereby gives point to the concept of understanding (since it would be pointless if empty).

But the two sides of the question also need to be addressed together because they are importantly linked. The relationship between the concept of understanding and the cognitive state it denotes is not one-sidedly representational; it is a *co-evolutionary feedback loop*.

On the one hand, having the concept of understanding shapes our cognitive life. By giving a name and value to the state we call "understanding," we create an ideal that shapes our cognitive aspirations and our pedagogy. We are encouraged to structure our thoughts in ways that our community recognizes as "understanding." This is a version of what Hacking (1995) calls a "looping effect," where our classifications of human kinds feed back to alter the very people being classified. And how exactly the concept shapes our cognitive life depends on the *functions* the concept serves, because these determine where and how it is used.

On the other hand, for the concept to gain a foothold, it must have a basis in our actual cognitive capacities and the practical challenges we face. We would be unlikely to develop a concept for a cognitive state that was never achieved, or that served no purpose. To understand why we came to care about understanding sufficiently to develop a concept thereof, we also need to understand what makes the state of understanding so important to human life.

This co-evolutionary feedback loop motivates a two-pronged approach. Beginning with a reconstruction of the functional rationale for the *concept* of understanding will help us discern some of the distinctive pressures shaping the *state* of understanding—pressures one would likely miss if one thought only about the function of the state. This is because thinking about the role of the concept points to the *social* dimension of understanding: we need not just to understand the world for ourselves, but to figure out who to trust and who to learn from—we rely on others based on their understanding and achieve understanding notably by learning from them.[7] It is these other-regarding applications of the concept of understanding which provide its initial *raison d'être*. And the social character of the practice of identifying "understanders" affects what *kinds* of understanding are favoured by these practices of trust and transmission. Sociality, we argue, is what gives human understanding its distinctive character.

Using this two-pronged methodology, we first argue that the primary function of the concept

[7]Following recent work on the sociality of belief (Chrisman & Marušić, 2025), one could even conceptualize the sociality of understanding more radically, as the idea that shared understanding is prior to individual understanding, and that even private individual understanding is best understood as a potential contribution to shared understanding. But our argument here does not depend on this strong reading of the sociality of understanding

of understanding is to act as an efficient *proxy* for tracking robust competence—the stable ability to succeed across a wide range of novel and counterfactual tasks.[8] This behavioural trait is vital for social coordination, and we need to track it if we are to know who to trust and who to learn from. But it is also highly labour-intensive to track directly, only partially observable, and easily confused with something more brittle. We therefore employ a concept to track the underlying cognitive state that grounds and reliably predicts robust competence.

Guided by the function of the concept, we then turn to the pressures driving the emergence of the cognitive state of understanding, arguing that it is the product of convergent pressures on bounded, social agents to predict the world using cognitive models that are not only accurate, but also compact enough to be stored, demonstrated, and transmitted.

The result is what we call the *Competence through Predictive Compression* (CPC) framework. It offers a unified account of understanding both at the object level, by integrating the various characteristics associated with understanding, and at the reflective level, by integrating disconnected literatures about understanding across disciplinary divides. By way of conclusion, we show how our account of the practical pressures at the root of human understanding elucidates debates over non-human understanding, particularly in AI.

2 The Functions of the Concept of Understanding

2.1 Identifying Who to Trust and Who to Learn from

What does the concept of understanding fundamentally help us to achieve? We suggest that the concept of understanding fundamentally acts as a way to track *robust competence* in order to know *who to trust* and *who to learn from*.

As members of societies that depend on a division of labour, we frequently rely on others to do things on our behalf. This means that we constantly need to decide who to trust. Can I rely on the pilot to fly the plane? Should I trust the doctor's recommendation? Locating and assessing competence in performing certain tasks is an ineluctable concern in collaborative societies. And one function of the concept of understanding is to help us *identify who to trust*. When a person displays the marks of understanding something, we take this as evidence that they are *epistemically trustworthy*. This aligns with Linda Zagzebsky's observation that "when we want an expert about a problem, we consult a person who has *understanding* of the subject matter, since such a person is likely to be a reliable problem solver" (2001, 245; see also Nado, 2025).

The concept of understanding thus performs a *fiduciary* function (from Latin *fiducia*, trust) by enabling us to track competence—and not just any kind of competence, but *robust* competence. If someone's competence is to be truly reliable, it should not be limited to a narrow range of routine scenarios, but should hold up even when faced with novel, unexpected, or particularly difficult challenges. Accordingly,

[8] What kinds of tasks, or rather, what kinds of competence? As will emerge below, the kinds achievable through predictive compression. We do not want to commit ourselves to the further claim that all kinds of competence are achievable through this way.

the concept of understanding needs to track *robust competence*: the stable, flexible, and reliable ability to succeed across a wide range of actual and counterfactual tasks.

In addition, the concept of understanding also allows us to tap into what grounds the robust competence of others, because it marks out people from whom something can be learned. This ties in with Michael Hannon's (2019) thesis that the point of the concept of understanding is to flag good explainers. When people display the marks of understanding something, we take this as evidence that they make good teachers on the subject. As legions of long-suffering students can attest, that remains a defeasible inference: some of the least comprehensible lectures have been given by the most comprehending minds. Nevertheless, offering principled explanations is not just a way to signal that one can be trusted. It is, above all, a means of *transmitting* understanding from one agent to another; when things go well, the principles *exhibited* in the explanation are also *transferred* from the explainer to the addressee of the explanation.

The concept of understanding thus also performs a *transmission* function by enabling us to track robust competence. Explanation instills understanding, thereby contributing to the spread of robust competence. And the concept of understanding supports this transmission by helping us decide *who to learn from*.

Both the fiduciary and the transmission function involve tracking robust competence. In the case of the fiduciary function, one tracks robust competence with a view to deciding who to trust, and one's ultimate aim is to rely on someone else's robust competence. In the case of the transmission function, one tracks robust competence with a view to deciding who to learn from; here, one's ultimate aim is to acquire robust competence for oneself. In both cases, however, the immediate aim is to *locate* robust competence. We therefore hypothesise that the concept of understanding acts, in the first instance, as a way of tracking robust competence with a view to identifying who to trust and who to learn from.

This tight link between understanding and robust competence is also underscored by the observation that when an agent is revealed to *lack* robust competence, we take this as evidence that they lack understanding. This was rendered salient by debates over whether AI systems understand. As Mitchell and Krakauer write: "the oft-noted *brittleness* of these AI systems—their unpredictable errors and lack of robust generalization abilities—are key indicators of their lack of understanding" (2023, p. 1). This acknowledges by contraposition that the concept of understanding tracks robust competence: if understanding, then robust competence; if no robust competence, then no understanding.

2.2 The Proxy Hypothesis

Given this tight link, one might be tempted to simply *assimilate* understanding to robust competence. Robert Brandom, for instance, argues that understanding just *is* practical mastery. A student seeking to understand mathematics must not wait for some "inner light" to go on, but simply "practice making the moves … until he masters the practical inferential abilities in question" (2009, pp. 172–3).

Similarly, Philipp Koralus contends that "what is worth caring about in the notion of understanding" is "*competence*" (2022, p. 35). He concludes that it is competence we should focus on going forward, both in the context of human reason and in the context of artificial intelligence.

In our view, however, this collapses two things worth distinguishing: (i) the behavioural property we ultimately care about and (ii) the cognitive property that grounds and predicts it. Our account preserves this distinction and shows why it earns its keep.

Our starting point is the observation that there is a fundamental difficulty with tracking robust competence directly: it is only ever partially observable. The best we can do is observe a thin, noisy slice of past performance, often only under routine conditions. Assembling even this track record of *observed*—as opposed to *observable*—past performance is highly labour-intensive, and its value for predicting future success is limited. What looks like a manifestation of robust competence may be something more brittle that breaks down under challengingly different circumstances.

This creates an epistemic challenge: we have a pressing practical need to identify a behavioural property—robust competence—that is (a) labour-intensive to keep track of, (b) only partially observable, and (c) easily confused with something more brittle.

We hypothesise that the concept of understanding developed as our solution to this challenge. The concept serves as a reliable and efficient conceptual *proxy* for tracking robust competence. Call this the *Proxy Hypothesis*. Given the infeasibility of an exhaustive and definitive assessment of robust competence, we largely track it *indirectly*, by tracking a *cognitive* property—understanding.

Understanding can be efficiently tested for while being a highly reliable predictor of robust competence. What the concept of understanding flags is a *grasp of relational structure*—of the *connections* between parts of an object, a situation, or a subject matter. That comes out more clearly in the Latinate twin of the Germanic word "understanding," namely "comprehension" (from *com-* "together" + *prehendere* "to seize"): it suggests a "grasping together," as if unifying parts in a single clasp.

Several philosophical accounts converge on the idea that understanding involves a *grasp of connections or relations* among elements of a subject matter. Ludwig Wittgenstein speaks of "the kind of understanding that consists in 'seeing connections' " (1953, §122). Jonathan Kvanvig argues that "understanding seems to involve conceptual and explanatory connections between various items of information that are seen or grasped by the person in question" (2018, p. 697). To grasp such connections amounts to seeing "how things hang together" (Riggs, 2003) or "how various parts relate to each other" (Grimm, 2011).

Genuine understanding thus goes beyond the possession of disconnected facts. It is not exhausted by knowledge, which can be a mere passive collection of isolated memories, nor by reasoning, which is an active, but abstract and domain-general process. As Figure 1 illustrates, understanding occupies a middle ground between knowledge and reasoning: a person understands a domain when they integrate domain-specific knowledge into a relational structure that enables them to actively generate new insights and handle novel cases.

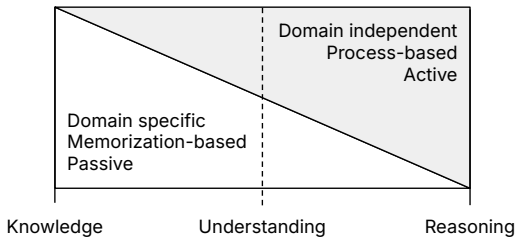

Figure 1: The concept of understanding marks the crossover between knowledge and reasoning.

For example, a novice driver might know a few isolated rules, such as: "If the temperature gauge is in the red, pull over." But a skilled mechanic truly understands the engine because they have grasped the underlying connections. They know the radiator cools the engine fluid, which circulates to prevent the engine from overheating. They know a faulty thermostat can block this circulation, causing the temperature to rise. The mechanic has grasped the causal and functional relationships between the parts, which in turn empowers them to predict what intervention will fix the problem. Similarly, when evaluating a doctor's competence, we are less interested in their ability to recall isolated trivia about diseases and more interested in their grasp of essential connections and dependencies between causes, symptoms, and treatments. Those who grasp how things are connected can apply their understanding to new cases and counterfactuals and make better predictions about them (De Regt, 2017; Hills, 2015).

The connections most crucial for securing this robust competence are *explanatory* ones (Baumberger et al., 2017; Khalifa, 2017). These can be, for instance, *causal-mechanical* connections that explain the physical processes of a phenomenon (Grimm, 2006; Pritchard, 2014) or *law-making* connections that unify disparate events under a general principle (Belkoniene, 2023; Friedman, 1974; Kitcher, 1981). But grasping other types of connections, such as logical, probabilistic, or conceptual ones, can be sufficient for certain forms of competence as well (Kvanvig, 2003, 2018).

What does "grasping" connections consist in? On our account, "grasping" is not best characterized in terms of the feeling of the "Aha!"-moment or the phenomenology of understanding. Nor, as Belkoniene 2023 has recently argued contra Grimm, is the grasping component best understood as a direct insight into modal reality. Rather, we argue that grasping the connections underlying a domain consists in forming a *cognitive model* of the domain—what Kenneth Craik (1943) called "thought models": internal representations, such as schemas, rules, diagrams, and causal models, that exhibit a "similar relation-structure" (1943, p. 51) to the processes they model.[9] These models preserve the relations of the represented domain well enough to support explanation and prediction. Sometimes, the model is explicitly articulated in rules, diagrams, or equations; sometimes, it remains tacit (e.g. a nurse's gestalt of a patient's clinical situation and trajectory).

Our Proxy Hypothesis maintains that the concept of understanding tracks robust competence *by* tracking a grasp of the connections underlying a domain. Grasping such connections in turn consists in the possession of a *cognitive model* whose relational structure is sufficiently isomorphic to the relational structure of its target domain to support robust predictions and explanations in that domain.

Evaluating the robust competence of agents in terms of whether they possess understanding is simultaneously less labour-intensive and more reliable. It is less labour-intensive, because even without a track record of past performance, we can rapidly assess how robustly competent someone is by probing for understanding: we can ask them to summarize or explain the principle behind something, probe their ability to handle a counterfactual, or give them a particularly challenging test case.

This is also more reliable than relying on track record alone. Two agents can have the same track record, yet differ in their cognitive organization in ways that matter for counterfactual and novel cases: only an agent with a genuine grasp of relational structure can be relied upon to continue to draw the right

[9] See Koralus (2022) for an illuminating discussion of Craik's "thought models."

inferences. We know that someone might obtain perfect scores on an exam and still lack understanding. Narrow competence on a test set does not entail *robust* competence. Indeed, one of the challenges of exam design is to put together a set of tasks that test for genuine understanding *as opposed to* the brittle competence achievable through rote memorization. An assessment of understanding is therefore never just about the task at hand; it is always also an assessment of whether the agent's competence will *transfer* to novel and counterfactual tasks. When we judge that a student has not only given the correct answer, but truly understood the principle underlying it, we are implicitly predicting that they would be able to solve a host of different problems as well.

This forward-looking aspect, together with greater efficiency, is the key advantage that appraisals of understanding have over mere appraisals of past performance. Hence the merit of tracking robust competence *via* the concept of understanding (in addition to keeping track of past performance—the two are not exclusive).

In sum, by acting as an efficient and reliable proxy for tracking robust competence, the concept of understanding performs two critical social functions: it helps us identify who to trust (the fiduciary function) and who to learn from (the transmission function).

However, as our co-evolutionary feedback loop hypothesis suggests, these social functions do more than justify the concept's existence. They form the environment that shapes the cognitive state of understanding itself. We now turn to the other side of this loop: the practical pressures that these social demands place on the formation and character of the cognitive state of understanding.

3 From Concept to State: The CPC Framework

3.1 Four Pressures Shaping Understanding

As our introduction outlined, we propose to use the social-epistemic functions of the *concept* of understanding as a guide to the practical pressures driving the formation and perpetuation of the *state* of understanding.

Now if the function of the concept of understanding is to track robust competence, this imposes a fundamental pressure on the underlying state: it must be *predictive*. Robustly competent agents need to be good forecasters: successful action frequently depends on predicting what is likely to happen. And if the state in which an agent has formed a model embodying a domain's underlying connections is a reliable proxy for robust competence, this is primarily because such a model is predictive, empowering the agent to anticipate what will happen next. That predictive power is what grounds robust competence and makes those who possess it valuable to others.

On its own, the predictivity pressure might favour ever more complex models as long as greater complexity translated into greater accuracy. Due to constraints on human cognition, however, there are practical pressures on models to remain *storable* within the bounds of human memory, which in turn helps keep them *manipulable* in one's mind. An overwhelmingly complex model would be useless. Accordingly, the second pressure is one towards *storability-cum-manipulability*.

These two pressures already point to the need for cognitive models that are powerfully predictive,

yet economical. But storability and manipulability require that a model not be overwhelmingly complex; they do not require the model to be particularly elegant or concise. To understand where the even stronger pressures towards simplicity that give human understanding its distinctive character come from, we must factor in the social dynamics of understanding. This reveals two further pressures.

The third pressure is towards *demonstrability*. Since robust competence is not directly observable, we need reliable and efficient signals to know who to trust. This creates a powerful social pressure on understanding to take a particular form. If it is to be quickly and effectively demonstrable to others, understanding cannot afford to take the form of an inscrutable thicket of ineffable connections and intuitions. It should ideally have been boiled down to principles that are discursively stateable, consistent, and coherent. Only such principle-based understanding is easily demonstrated to others. It is what enables concise explanation. And the fewer and more general the principles, the simpler and more elegant the understanding one demonstrates, and the wider the scope of the competence one thereby lays claim to.

The social needs for understanding to be not just tacit but demonstrable thus drive understanding towards principled simplicity. The expert who wants to advertise her competence needs her understanding to be concisely displayable in the form of a few explicit principles. More broadly, those whom others rely on will often be expected to vindicate the trust they receive by demonstrating their understanding.[10] A predictive but incommunicably complex form of understanding would fail to meet these crucial social needs, as it cannot easily be exhibited to earn and justify trust.

The fourth pressure, finally, is towards *transmissibility*. A primary reason we track understanding in others is to identify who to learn from. But this means that understanding must pass through the narrow bottleneck of communication, which requires it to be condensed into principles and systematically structured to facilitate teaching and learning. It is one thing to briefly exhibit some of the rationales guiding one's actions and recommendations, as the doctor earning the trust of a patient might; it is quite another to actually transmit one's full cognitive model, as the professor of medicine might do with her students. The most effective teachers are those who have distilled a domain down to its most basic principles and arranged them in a coherent and memorable system. Thus, the social pressure towards transmissibility is another driver towards principled simplicity. As creatures who do not just acquire understanding individually but share it with others, we are driven towards cognitive models that are elegantly principled and eminently communicable.

As summarised in Figure 2, this gives us four distinct pressures moulding the state of understanding:

Considering the social-epistemic function of the concept of understanding thus highlights additional formative pressures on the cognitive state. The need to identify trustworthy experts and effective teachers favors a particular *kind* of understanding: one that is not only predictive, but also communicable, and therefore such that it can be made discursively explicit in terms of a limited number of principles. The more elegantly principled understanding is, the more efficiently it can be signaled and conveyed. This creates a co-evolutionary feedback loop: just as the practical need to acquire the state of understanding favours the formation of a concept to recognize it in others, the social use of that concept in turn refines

[10] This aligns with Belkoniene's emphasis on the fact that understanding must be based on "reflectively accessible grounds" (Belkoniene, 2022, p. 349).

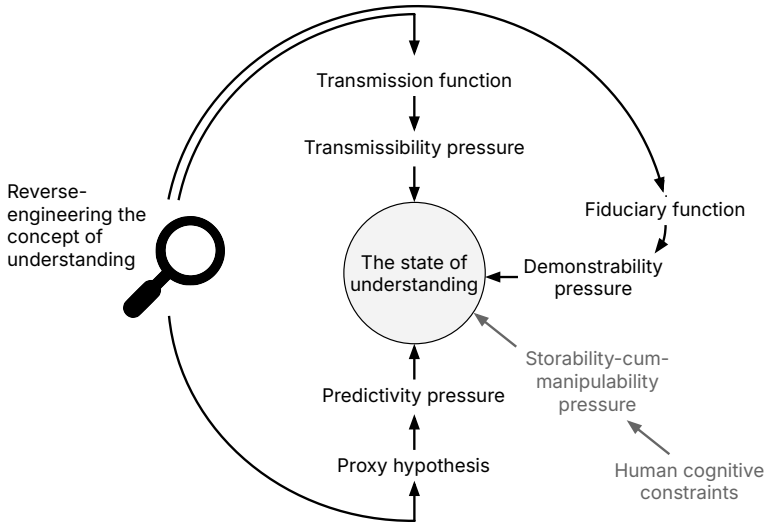

Figure 2: the four pressures shaping the state of understanding.

the state itself. Consequently, the character of human understanding—its tendency towards systematic, communicable models—is explained by the social character of human cognition. Sociality is a key driver of principled simplicity.

3.2 Converging on Predictive Compression

How can a cognitive model simultaneously satisfy the demands for predictivity, storability, demonstrability, and transmissibility? We argue that these pressures converge in favouring the formation of cognitive models that achieve *predictive compression*, empowering one to forecast what happens next while being economical in representation. Our *Competence through Predictive Compression* (CPC) framework treats this as the functional core of the state of understanding: it is a cognitive organization that confers robust competence *through* predictive compression.

Crucially, however, the path to compression is better prediction. The very features that render a model predictive also enable it to compress the data it describes. As Jürgen Schmidhuber puts it, "whatever you can predict you can compress as you don't have to store it extra" (2010). If you can predict something well, it means you discerned a *pattern* in it. It is not random, but obeys some rule. That pattern can then be used to describe the data in question in a shorter, more compressed way. Given a completely random string of letters, for example, you cannot predict what comes next, and are forced to memorize the whole string. But if a string follows a pattern (e.g. "ABABAB"), you can predict the next letter at any point, allowing you to compress the description of the entire string into the rule: "Repeat AB over and over."

The link from better predictions to shorter descriptions is forged by a fundamental insight from information theory: the information content of something can be quantified in terms of how *surprising* it is. As Claude Shannon established, an event that is highly probable is not surprising, and being told it occurred provides little information ("The sun rose this morning" is not news). An improbable event, by contrast, is highly surprising and conveys a great deal of information. The goal of a predictive model

can be framed as an effort to minimize its average surprise at incoming data. A good model is one that assigns high probabilities to events that actually transpire and low probabilities to those that do not.

This provides the bridge to compression. A string of data is truly *random* if and only if the information required to describe (or transmit) the series is *incompressible*: nothing shorter than a verbatim rendition of the entire string will preserve it. Conversely, the string has a *pattern* if and only if it admits of a description that is shorter than the string itself (Chaitin, 1975; Dennett, 1991; Ladyman & Ross, 2007; Suñé & Martínez, 2021). To discern redundancies, repetitions, or even just uniformly coloured areas is already to discern patterns, enabling shorter descriptions by abbreviating the predictably repeated strings or encoding uniform areas by specifying only their boundaries and colour (as in "colour-by-number" books) instead of describing them pixel by pixel. Similarly, once one understands that certain events are more likely than others, one can use shorter descriptions or shorthand symbols for them. Even the realisation that certain letters are more frequent than others enables compression; one gives common letters short encodings and rare letters longer ones (this is done in Huffman coding, for example). The better a model forecasts the data, the shorter the description needed to specify that data *given* the model. The payoff of good prediction is compression.

This holds for understanding the structure of the world more widely. Imagine you must send a message each hour saying whether it is raining ("R") or dry ("D"). A naive baseline treats the probability as being 50/50 each hour and requires 24 symbols per day. Now suppose you discover a pattern: when a cold front arrives, it brings a six-hour band of rain. Because you expect contiguous rainy hours once the front appears, you can encode the day's sequence with fewer symbols—by only giving the starting hour of the band plus its deviation from the rule ("fronts move in six-hour bands").

More generally, encoding stable relations or connections into one's model lowers average surprise across new data—encoding a law, a symmetry constraint, or a conservation constraint into a model, for example, prunes away large regions of the space of possible outcomes, thereby making certain kinds of new data systematically less surprising. This is compression *thanks to sensitivity to structure*. The representation is shorter than the data *because* it exploits the relational structure of the domain. This is why many in computer science view compression as *the* hallmark of understanding: compression is the representational shadow cast by structure-sensitive prediction.

However, structure-sensitivity itself comes in degrees reflecting the *depth* of one's understanding, because one can be sensitive to structures of varying depth. An illustration of this spectrum is the progression in understanding planetary motion realized by Tycho Brahe, Johannes Kepler, and Isaac Newton.[11]

Brahe recorded the motions of the planets with unprecedented precision, thereby acquiring a grasp of their "relational structure" in the most superficial sense: he provided snapshots of how they spatially related to each other at different moments. But he remained unable to *predict* the movements of the planets, because he had not grasped the underlying patterns—the relational structure in the deeper sense of hidden regularities.

Kepler discerned these underlying patterns and articulated them in his three laws of planetary motion:

[11]On Brahe, see Thoren (1991). On Kepler, see Voelkel (2002). On Newton, see Westfall (2015). On the progression in understanding they made possible, see Li et al. (2021).

planets move in elliptical orbits with the Sun at one focus, sweep out equal areas in equal times, and the time a planet takes to complete a full orbit is proportional to its distance from the Sun. This enabled him to predict planetary motion within a small fraction of a degree and facilitated compression: given a few observations about Venus's position, Kepler could work out its position on subsequent days. But he could not explain *why* the planets moved according to the patterns he had discerned.

It was only with Newton, who formulated the laws of motion and universal gravitation underpinning these patterns, that an even deeper understanding of the underlying structure was provided. Newton grasped the relational structure not just in the sense of discerning hidden regularities, but in the sense of capturing the underlying laws. And with Newton's laws unifying our understanding of terrestrial and celestial mechanics, the movements of moons and comets became unsurprising as well.

The progression from Brahe through Kepler to Newton also goes hand in hand with greater compression and easier transmission: instead of handing over a thick book of coordinates, one need only share "orbital elements"—a few key parameters that characterize a planet's orbit—such as its semi-major axis (half its longest diameter) and its eccentricity (its deviation from a perfect circle); the laws can then be used to generate the rest of the path.

A more formal account can be given using information theory and the minimum description length principle (see Grünwald, 2007). Let D be a finite dataset of observations relevant to some task family. Let $L(\cdot)$ be description length in bits. To encode D using a model, one chooses a model M and then encodes the data relative to it. The resulting description length is:

$$L(M) + L(D|M)$$

Think of $L(M)$ as the cost of storing the model (e.g. rules, principles, constraints, and dependencies), and $L(D|M)$ as the cost of storing the data *given* the model, i.e. the cost of storing the *residuals*, such as initial position, deviation from the rule, noise, etc. The description length of the data given the model, $L(D|M)$, is a formal measure of the model's total surprise at that data. Encoding stable connections—a law, a symmetry, or a causal dependency—into one's model has the effect of making certain new data systematically less surprising, thereby reducing $L(D|M)$.

Compression is achieved if the total description length of the model plus the data given the model is smaller than a baseline without a model, e.g. storing D as raw data. We say that M *compresses* D iff

$$L(M) + L(D|M) < L(D)$$

But note that when "data is compressed" in this sense, it is not literally crammed *into* a model; rather, the description of the data is split into two parts: the model M which is the part that encodes reusable *structure*; and the data as encoded under M which only contains residuals, i.e. whatever is left over once one has accounted for the regularities captured by M.

Predictive compression is thus not about discarding data, but about building a model that renders the data less surprising and therefore easier to describe. This measure of a model's average surprise has a formal name: *negative log-likelihood* (or *cross-entropy*). The core idea is to quantify the surprisingness (or "surprisal") of a single outcome as $-log(P(outcome))$. This captures the inverse relationship we

intuitively expect between probability and surprise: as an outcome's probability $P(outcome)$ approaches 1 (certainty), its negative logarithm approaches 0, meaning there is zero surprise—an event that is guaranteed to happen tells us nothing new. Conversely, as an outcome's probability $P(outcome)$ approaches 0 (near impossibility), its negative logarithm grows towards infinity, reflecting immense surprise. A cognitive system that seeks to understand a domain is therefore engaged in a process of adjusting its internal model to minimize this surprise on average across all the data it sees, which is equivalent to maximizing the likelihood it assigns to that data. The total description length of the data given the model, $L(D|M)$, can thus be seen as the sum of all these individual surprisals. It is a formal measure of the model's cumulative predictive error.

Putting these ideas together, we can say that cognitive models must solve a optimization problem: keep representations simple enough to store and share, yet structure-sensitive enough to yield robust predictions. In formal terms, a cognitive model is thus formed in pursuit of the following objective:

$$\min_{M} \; L(D_{future} \mid M) + \lambda \, L(M)$$

Let us call this the *ultimate objective* of cognitive model formation. The first part of this objective function—$minL(D_{future}|M)$—aims to minimize the description length of future data under the model, expressing a striving for models that assign a high probability to future data. Let us call this "data compression." The second part—$minL(M)$—aims to minimize the description length of the model, expressing a striving for models that are easy to store and share. Let us call that "model compression." The parameter λ reflects how heavily one prioritises compact simplicity relative to predictivity. A $\lambda > 1$ encodes a strong preference for simple models, while $\lambda < 1$ allows for more model complexity. Another interpretation of λ would be that the parameter encodes the degree of one's *presumption of systematizability* (that a short, simple, unified model of the domain exists). Different contexts might demand different λ values (e.g., a theoretical physicist searching for a unifying law might have a high λ, prioritizing simplicity; an ethnographer interested in cultural diversity might have a lower λ).

Because we cannot *directly* optimize on as-yet-inaccessible future data, however, agents approximate the above by learning on accessible *past* data. Accordingly, the *proximate* objective (through the pursuit of which the ultimate objective is pursued) is to minimize the total description length of the model plus the *past* data given the model:

$$\min_{M} L(D_{past}|M) + \lambda L(M)$$

Let us call this the *proximate objective* of cognitive model formation. The main risk of this strategy is that one comes to rely on superficial correlations, which can reduce $L(D_{past}|M)$, but tend not to reduce $L(D_{future}|M)$. Hewing too close to the patterns exhibited by past data risks leading to overly complex models that mistake idiosyncrasies for structure. That is another reason why a balance needs to be struck between predictive power and simplicity. Increasing λ helps steer towards structure-sensitive compression by penalizing *overfitting* to past data at the expense of predictivity on future data. And it does this while at the same time steering towards models that are easier to store and share, thereby responding to the social pressures driving the formation of cognitive models.

The trade-off between model simplicity, which corresponds to minimizing $L(M)$, and predictive power, which corresponds to minimizing $L(D|M)$, can be visualized in the two-dimensional landscape of understanding shown in Figure 3.

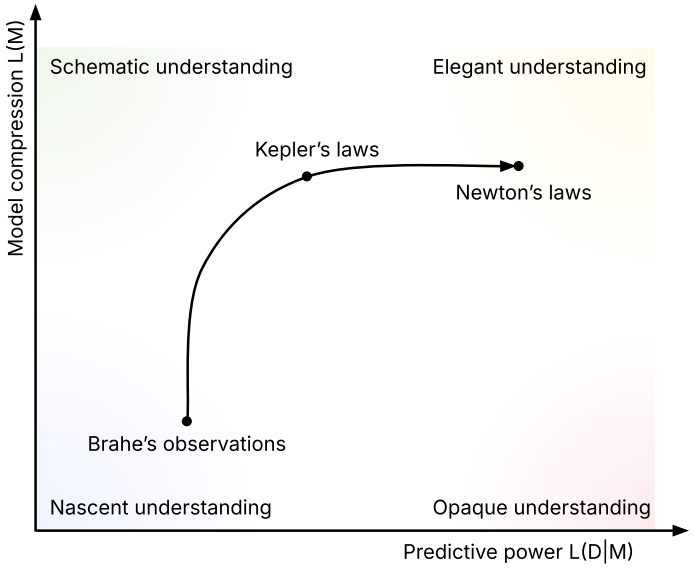

Figure 3: Mapping understanding along two axes: model simplicity (vertical) and predictive power (horizontal). The progression from Brahe through Kepler to Newton illustrates movement from nascent to elegant understanding. (The diagram is not a scientific result; it is intended solely as an illustrative aid.)

On the one hand, there are strong pressures towards *model compression*—minimizing $L(M)$. These pressures, which push cognitive systems towards the upper half of Figure 3, stem directly from our nature as bounded and social agents. When this drive for compression is applied to domains with high contingency and low predictability, such as history, we end up in the top-left quadrant, with *schematic* understanding. Examples include the explanatory schemas invoked by historians to make sense of events such as the outbreak of a war or the fall of an empire. The predictive power and scope of such schemas is modest, but they nonetheless offer a compressed model of a complex world.

The ultimate goal of human understanding, however, is to pair model simplicity with high predictive power, arriving at the top-right quadrant: *elegant* understanding. This is the ideal exemplified by Newton's Laws or General Relativity—highly condensed models that nevertheless yield extremely precise predictions over a vast range of phenomena.

On the other hand, there are pressures towards *data compression*—minimizing $L(D|M)$. These pressures push cognitive models towards the right half of the map. The most fundamental pressure here is towards predictivity. A good model minimizes our surprise at the world by assigning high probability to what actually happens. This predictive success, in turn, allows for data compression (Ramstead et al. 2025, p. 7). Because a simpler model is less likely to mistake noise in past data for a signal, moreover, the pressure to minimize $L(M)$ simultaneously helps avoid overfitting, rendering models more likely to generalize to future data and thereby reducing $L(D|M)$. In Figure 3, this means that moving up also tends to push one to the right. And the pressures towards storability-cum-manipulability, demonstrability,

and transmissibility all favour a shorter $L(D|M)$ as well.

When unchecked by these characteristically human pressures towards compression, however, pushing for predictive accuracy can lead to the bottom-right quadrant: *opaque* understanding. This is where many contemporary deep learning systems and large-scale climate models end up (see Knüsel and Baumberger, 2020). These achieve high predictive accuracy via vast, inscrutable models that are strong on predictivity but weak on storability, demonstrability, and transmissibility.

The bottom-left quadrant, finally, represents *nascent* understanding, where the model consists of little more than raw data—like exploratory fieldwork data. Such a model is complex, offers little predictive leverage, and is hard to transmit.

The progression of scientific insight often charts a path through this landscape. As illustrated in Figure 3, Brahe's observations represent a form of nascent understanding—a complex dataset with little predictive structure. Kepler's laws chart a path upwards and to the right, introducing a dramatically simpler model (higher on the vertical axis) that also unlocks powerful predictive abilities (further on the horizontal axis). Newton continues this path towards the top-right, offering an even more general model that constitutes a deeper, more elegant understanding of the underlying laws.

Table 1 summarises these pressures towards model and data compression using the example of Kepler's laws:

Model compression: **pressures towards minimising** $L(M)$**:**	
1. Pressure towards model *storability* and *manipulability*	Kepler's three laws are easy to remember and apply.
2. Pressure towards model *demonstrability*	One's understanding of planetary motion can be concisely displayed by citing Kepler's laws
3. Pressure towards model *transmissibility*	Kepler's compact laws can be easily taught.
4. Pressure to avoid *overfitting*	Kepler's laws extend to new planets.
Data compression: **pressures towards minimising** $L(D|M)$**:**	
1. Pressure towards *predictivity*	Kepler's laws predict Venus's future positions.
2. Pressure towards data *storability*	Venus's trajectory can be stored by storing only its orbital elements (from which the full trajectory can be retrieved)
3. Pressure towards data *demonstrability*	One's possession of Venus's full trajectory can be concisely displayed by showing only its orbital elements
4. Pressure towards data *transmissibility*	Venus's full trajectory can be transmitted by conveying only its orbital elements

Table 1: The practical pressures driving bounded social agents to form cognitive models.

We can now see how this account unifies the two disconnected literatures that have circled the concept of understanding, and in particular the philosophical literature conceptualising understanding as "grasping connections" and the computer science literature conceptualising understanding as "compression." Far from being at odds, these two research programs describe different aspects of the same

answer to the same set of pressures. Our account explains the philosophical focus on connections—and particularly on explanatory connections—by revealing their functional value: grasping the relational structure of a domain is the most effective strategy for achieving predictive compression. "Grasping the connections" underlying a domain is, at a functional and information-theoretic level, the process of forming a structure-sensitive model of the domain that enables prediction and compression.

At the same time, this perspective makes sense of why many in computer science and AI research (Chaitin, 2002, 2006; Delétang et al., 2024; Hutter, 2005; Li et al., 2024; Maguire et al., 2015; Ramstead et al., 2025; Schmidhuber, 2006, 2008, 2010; Wolfram, 2018; Zenil, 2019) equate understanding with compression or approvingly cite Gregory Chaitin's dictum that "compression is comprehension" (2006, p. 77). On our account, compression is not strictly identical with comprehension or understanding, but the two are closely linked: compression is a direct consequence of a model's predictive power and structure-sensitivity.

Indeed, our account offers an interpretation of common computer science formalisms that reveals them to express deeply human exigencies. For example, when we give equal weight to a model's simplicity and its ability to predict data ($\lambda = 1$), our objective function becomes an instance of the Minimum Description Length (MDL) principle, which computer scientists regard as a computable application of theoretical ideas pioneered by Ray Solomonoff (1964) and Andrey Kolmogorov (1968). What might seem like an abstract rule from information theory is revealed to be a formal expression of concrete practical pressures of human life.

Similarly, the term for model complexity, $\lambda L(M)$, which the machine learning literature regards primarily as a mathematical regularizer designed to prevent overfitting, turns out to have a richer set of rationales in the context of human learning. Our account shows that for human beings, this term also represents something more concrete: the practical pressures of cognitive finitude and social codependence. We are not merely optimizing for predictivity; we are optimizing for predictivity with models that can be stored and shared.

3.3 What Constitutes and What Indicates Understanding

Based on this CPC framework, we can now (a) articulate the constitutive conditions for understanding, (b) derive the indicators we rely on to guide our attributions of understanding, and (c) clarify the relationship between understanding and explaining.

On our account, understanding is neither a particular feeling nor reducible to a purely behavioural property like competence. Instead, we define understanding as follows:

> **Understanding (Def.):** A cognitive system *understands* a domain iff it possesses a structure-sensitive cognitive model that facilitates the prediction and compression of data in that domain. The *depth* of understanding can be graded accordingly: deeper understanding corresponds to a model that is more compressed, less surprised by new data (i.e., more accurate), and applicable to a wider range of data.

This is what it *is* to have understanding, on our account. But since this constitutive condition of understanding is itself not directly observable, we rely on a cluster of observable *indicator properties* to *tell* whether someone should be attributed understanding. Our CPC framework suggests that these key indicative marks of understanding include:

- **Principled Explanation**: The ability to concisely explain by explicitly articulating underlying principles externalizes a compressed model (low $L(M)$). A good explanation strips away incidental details and highlights the core relational structure.
- **Accurate Prediction**: The ability to forecast novel data is the most direct evidence of a cognitive model that minimizes $L(D_{future}|M)$.
- **Success on Hard Cases**: The ability to deal with particularly challenging, previously unseen problems. This tests whether the model is truly structure-sensitive and not overfitted to past data.
- **Handling Counterfactuals**: The ability to accurately predict how a system would change if certain conditions were different showcases a grasp of relational structure.
- **Patterned Error**: Understanding does not imply the absence of error, but that even one's errors exhibit a pattern, often occurring at the predictable boundaries of the model's core assumptions or idealizations. By contrast, an agent without understanding fails erratically.

The importance of principled and concise explanation as an indicator of understanding is a direct reflection of its co-evolution with the *concept* of understanding; we value the ability to explain concisely because our concept was forged in large part to identify those who could effectively signal and transmit their competence.

This accommodates, but strictly speaking *inverts* Michael Hannon's (2019) hypothesis that the point of the concept of understanding is to identify good explainers. Our hypothesis is rather that the concept of understanding fundamentally serves to track robust competence by flagging agents that have achieved predictive compression through cognitive models, and *one way to do this* is to look for good explainers. In humans at least, the ability to explain things well indicates that someone has achieved structure-sensitive compression, which is in turn a good predictor of robust competence. Moreover, once we have identified agents with understanding, the understanding they possess can be *transmitted* through explanation. This social multiplication of the value of understanding further encourages the formation of cognitive models that can more easily be transmitted.

These indicators are not indefeasible, but they are our most reliable guides to the presence of understanding. They work because it is exceptionally difficult, at least for a human agent, to consistently exhibit these behaviours without possessing the cognitive models that constitute understanding.

With this characterization in place, we can sharply distinguish understanding from explaining—a distinction that risks being blurred in the literature focusing on the analytic structure of good explanations (e.g. De Regt, 2009; Khalifa, 2012; Strevens, 2013). De Regt (2009, p. 25), for example, claims that to understand a phenomenon *just is* to have an explanation of that phenomenon. In our framework,

understanding is the state of possessing a cognitive model, while explanation is in the first instance a social, communicative act—specifically, the speech act aiming to transmit a cognitive model from one agent to another.

While other senses of "explanation" exist (e.g., a proposition explaining another without a communicative act), the sense most immediately illuminated by our account is the one in which explaining is an attempt to transfer a cognitive model.[12] In information-theoretic terms, an explanation of a target event T is, in the first instance, a communicative attempt to instill in another agent a model M capable of generating the data D_T relevant to T. An explanation is successful if the receiving agent forms such a model M and can use it to generate D_T, thereby minimizing $L(D|M)$. Explanations are the principal social technology for moving these models between minds.

4 Objections and Replies

4.1 Compression without Understanding?

An objector might point to simple compression algorithms as counterexamples to our account. A program that creates a .zip file achieves compression, but no one would claim that the zip program "understands" the text it compresses. This suggests that compression, on its own, is insufficient for understanding.

Reply: This objection highlights a crucial distinction between two kinds of compression. There is compression which just operates on the data's surface representation, identifying and exploiting redundancies or repetitions without regard for how the data was generated, and hence without any ability to predict future data. This is what creating .zip file does. And then there is compression which involves constructing a model capable of predicting or generating the data. This form of compression is *structure-sensitive*: it hinges on capturing aspects of the underlying generative structure. But even structure-sensitive compression is not *identical* with understanding; compression is the *representational consequence* of forming a predictive model.

4.2 Understanding without Compression?

Conversely, one might imagine a highly predictive model that is vast and unwieldy. Consider an expert with a lifetime's worth of case-based knowledge. Their cognitive model might seem far from compressed—yet it could still ground the ability to make accurate predictions and navigate novel scenarios. Does this not represent a case of understanding without compression?

Reply: This objection rests on a misunderstanding of what compression entails in the CPC framework. If a model, no matter how baroque, yields predictive leverage, then it *ipso facto* offers a path to shorter descriptions of the data. But if the understanding in question involves a truly unwieldy model—meaning it has a high $L(M)$—our framework predicts two deficits. First, it is more likely to result in fragile rather than robust competence, as a model laden with excessive detail is prone to overfitting. Second, it runs counter to the practical pressures towards demonstrability, and transmissibility. It will be difficult to

[12] This aligns with Turri's (Turri, 2015) thesis that the norm of (the speech act of) explanation is to *express* understanding.

advertise, demonstrate, and teach. Our framework therefore correctly classifies such a model as indicative of a lesser degree of understanding, because its complexity undermines its functionality.

4.3 Domains with Low Predictability

The CPC framework, with its emphasis on prediction, seems best suited to domains governed by strict laws, like physics or engineering. What about domains like history, where events are often unique and highly contingent, rendering laws elusive and prediction difficult?

Reply: Our framework is graded and adapts to the structure available in a given domain. In domains with low predictability, the achievable degree of predictive compression is more modest, but it is not zero. A historian who understands a period may not possess a predictive law, but they do possess compressed models in the form of narrative templates, schemas of political and economic pressures, and models of human motivation. This corresponds to the top-left quadrant of our framework (Figure 3): *schematic* understanding.

Schematic understanding manifests as the ability to provide concise explanations and coherent narratives—both forms of compression. And while it cannot precisely predict unique events, it can anticipate typical patterns of behaviour, and also supports counterfactual reasoning (e.g., "What would have happened if Archduke Ferdinand had not been assassinated?"), demonstrating a grasp of causal dependencies. In these domains, understanding thus still consists in capturing available structure—even if that structure is schematic rather than nomic.

5 Conclusion: Beyond Human Understanding

It is often said that past performance is the best predictor of future performance. Our argument suggests that we have evolved a better strategy—probing for understanding. We began by recasting the question "What is understanding?" as "Why do we care about understanding?" This reorientation led to a two-pronged inquiry into the co-evolution of the concept of understanding and the cognitive state it denotes. We argued that the concept functions as a particularly efficient proxy for tracking robust competence, enabling us to identify who to trust and who to learn from. We then explored how these functions generate pressures that shape the cognitive state itself. The result was the CPC framework, which defines understanding as the possession of a cognitive model that facilitates predictive compression. If the concept of understanding is a reliable proxy for robust competence, it is because agents capable of predictive compression are exactly those who exhibit robust competence. This, on our account, is the most basic reason why we came to care–and should continue to care–about understanding. It helps us track the kind of competence that comes with predictive compression. In mnemonic form: *comprehension tracks competence through compression*.

The CPC framework achieves two significant unifications. At the object level, it explains why a heterogeneous cluster of characteristics—explaining, predicting, compressing, handling counterfactuals—are all manifestations of a single cognitive achievement: they are indicators of cognitive models enabling predictive compression. At the reflective level, it connects previously separate conversations around

understanding in philosophy and computer science.

Finally, grasping the forces that drive the emergence of understanding and shape what kind of understanding forms also puts us in a better position to think about AI understanding. It invites us to see the field's preoccupation with whether models "generalize" to new data as the latest instantiation of the longstanding human preoccupation with robust competence.

Yet our framework also suggests that the critical question is not *whether* a system can generalize beyond its training data, but *how* it does so. In particular, does its success spring from a compressed, generative model that has captured the domain's underlying structure? To the extent that the rationale for tracking robust competence via attributions of "understanding" carries over from humans to machines, this gives AI research a reason to move beyond benchmarking black-boxes and probe for the presence of cognitive models.

But our framework also prompts us to think about what *kind* of understanding AI might develop and under what conditions. A key insight of the framework is that the kind of understanding an agent develops is shaped both by cognitive limitations and by the social pressures it is subjected to. Human understanding is not solely a product of the pressure towards predictivity. It is forged in the crucible of social interaction, under pressures for demonstrability and transmissibility, and these pressures drive it towards the kind of principled simplicity that renders it highly communicable.

By contrast, large-scale machine learning typically optimizes almost exclusively for a single objective: minimizing predictive error, which corresponds to minimizing $L(D|M)$ in our framework. Consequently, our framework predicts that in the absence of other pressures, whatever "understanding" AI develops will gravitate towards what we termed "opaque" understanding: high predictive power coupled with inscrutably immense model complexity (a high $L(M)$). This outcome reflects the comparative absence of demonstrability and transmissibility pressures. AI models are not primarily trained to advertise and display their understanding to other models; and while some *are* used to teach other models, they do not teach them *through explanation*. This lack of a communicative bottleneck allows model complexity to grow unchecked.[13] There is far less incentive to form systematically principled and communicable kinds of understanding.

Our framework thus not only makes sense of the inscrutability of AI models, but also points to potential remedies.[14] The forces that historically forged human understanding hold lessons for the future of artificial understanding.

References

Baumberger, C., Beisbart, C., & Brun, G. (2017). What is understanding? an overview of recent debates in epistemology and philosophy of science. In *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 1–34).

[13] On how this complexity affects AI understanding, see Beckmann (forthcoming) for deep learning models generally and Beckmann and Queloz (2025) for large language models specifically.

[14] The power of explanation to instill a better grasp of relational structure, for example, has already been observed in machine learning (Lampinen et al., 2022).

Beckmann, P. (forthcoming). Explanatory and objectual understanding in video generation models. *Synthese*.

Beckmann, P., & Queloz, M. (2025). Mechanistic indicators of understanding in large language models. https://arxiv.org/abs/2507.08017

Belkoniene, M. (2022). The rational dimension of understanding. *Synthese*, *200*(5), 349. https://doi.org/10.1007/s11229-022-03839-z

Belkoniene, M. (2023). Grasping in understanding [doi: 10.1086/714816]. *The British Journal for the Philosophy of Science*, *74*(3), 603–617. https://doi.org/10.1086/714816

Brandom, R. B. (2009). *Reason in philosophy: Animating ideas*. Belknap Press of Harvard University Press.

Chaitin, G. J. (1975). Randomness and mathematical proof. *Scientific American*, *232*(5), pp. 47–52. Retrieved October 16, 2025, from https://www.jstor.org/stable/24949798

Chaitin, G. J. (2002). On the intelligibility of the universe and the notions of simplicity, complexity and irreducibility. https://arxiv.org/abs/math/0210035

Chaitin, G. J. (2006). The limits of reason. *Scientific American*. https://www.cs.virginia.edu/~robins/The_Limits_of_Reason_Chaitin_2006.pdf

Chrisman, M., & Marušić, B. (2025). Transparency, self-knowledge, and the sociality of belief. *Journal of Philosophy*.

Craig, E. (1990). *Knowledge and the state of nature: An essay in conceptual synthesis*. Clarendon Press.

Craik, K. J. W. (1943). *The nature of explanation*. Cambridge University Press.

De Regt, H. W. (2009). The epistemic value of understanding. *Philosophy of Science*, *76*(5), 585–597. https://doi.org/10.1086/605795

De Regt, H. W. (2017). *Understanding scientific understanding*. Oxford University Press.

De Regt, H. W., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, *144*(1), 137–170.

Delétang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L. K., Aitchison, M., Orseau, L., Hutter, M., & Veness, J. (2024). Language modeling is compression. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=jznbgiynus

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy*, *88*(1), 27–51.

Elgin, C. (1996). *Considered judgment*. Princeton University Press.

Fricker, M. (2016). What's the point of blame? a paradigm based explanation. *Noûs*, *50*(1), 165–183.

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, *71*(1), 5–19.

Grimm, S. (2006). Is understanding a species of knowledge? *The British Journal for the Philosophy of Science*.

Grimm, S. (2011). Understanding. In *The routledge companion to epistemology* (pp. 84–93). Routledge.

Grimm, S. (2012). "the value of understanding". *Philosophy Compass*, *7*(2), 103–117.

Grimm, S. (2021). Understanding. https://plato.stanford.edu/archives/sum2021/entries/understanding/

Grünwald, P. D. (2007, March). *The minimum description length principle*. The MIT Press. https://doi.org/10.7551/mitpress/4643.001.0001

Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 351–383). Oxford University Press UK.

Hannon, M. (2019). What's the point of understanding? In M. Hannon (Ed.), *What's the point of knowledge? a function-first epistemology*. Oxford University Press.

Hills, A. (2015). Understanding why. *Nous*, *50*(4), 661–688.

Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer.

Kelley, M. (2025). The normative function of intentional action. *Philosophers' Imprint*, *25*(1), 1–35.

Khalifa, K. (2012). Inaugurating understanding or repackaging explanation? *Philosophy of Science*, *79*(1), 15–37. https://doi.org/10.1086/663235

Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge University Press.

Kitcher, P. (1981). Explanatory unification. *Philosophy of science*, *48*(4), 507–531.

Knüsel, B., & Baumberger, C. (2020). Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science Part A*, *84*(C), 46–56. https://doi.org/10.1016/j.shpsa.2020.08.003

Kolmogorov, A. N. (1968). Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, *2*(1-4), 157–168. https://doi.org/10.1080/00207166808803030

Koralus, P. (2022, December). *Reason and inquiry: The erotetic theory*. Oxford University Press.

Kusch, M., & McKenna, R. (2020). The genealogical method in epistemology. *Synthese*, *197*(3), 1057–1076.

Kvanvig, J. L. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge University Press.

Kvanvig, J. L. (2018). Knowledge, understanding, and reasons for belief. In D. Starr (Ed.), *The oxford handbook of reasons and normativity* (pp. 685–705). Oxford University Press.

Ladyman, J., & Ross, D. (2007). *Every thing must go: Metaphysics naturalized*. Oxford University Press.

Lampinen, A. K., Roy, N., Dasgupta, I., Chan, S. C., Tam, A., Mcclelland, J., Yan, C., Santoro, A., Rabinowitz, N. C., & Wang, J. (2022). Tell me why! explanations support learning relational and causal structure. *International Conference on Machine Learning*, 11868–11890.

Lawlor, K. (2023). A genealogy of reasonableness. *Mind*, *132*(525), 113–135. https://doi.org/10.1093/mind/fzac036

Li, Z., Ji, J., & Zhang, Y. (2021). From kepler to newton: Explainable ai for science. *arXiv preprint arXiv:2111.12210*.

Li, Z., Huang, C., Wang, X., Hu, H., Wyeth, C., Bu, D., Yu, Q., Gao, W., Liu, X., & Li, M. (2024). Understanding is compression. *ArXiv*, *abs/2407.07723*. https://api.semanticscholar.org/CorpusID:271088447

Maguire, P., Mulhall, O., Maguire, R., & Taylor, J. (2015). Compressionism: A theory of mind based on data compression. *EuroAsianPacific Joint Conference on Cognitive Science*. https://api.semanticscholar.org/CorpusID:14116148

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), 1–5.

Nado, J. (2025). Conceptual engineering, the value of knowledge, and the value of understanding. In M. G. Isaac, S. Koch, & K. Scharp (Eds.), *New perspectives on conceptual engineering - volume 2: Across philosophy* (pp. 15–35). Springer Nature Switzerland.

Price, H. (2011). *Naturalism without mirrors*. Oxford University Press.

Price, H., Blackburn, S., Brandom, R., Horwich, P., & Williams, M. (2013). *Expressivism, pragmatism and representationalism*. Cambridge University Press.

Pritchard, D. (2014). Knowledge and understanding. In A. Fairweather (Ed.), *Virtue epistemology naturalized*. Springer.

Queloz, M. (2021). *The practical origins of ideas: Genealogy as conceptual reverse-engineering*. Oxford University Press. https://doi.org/10.1093/oso/9780198868705.001.0001

Ramstead, M. J. D., Pattisapu, C., Fox, J., & Beck, J. (2025). Noumenal labs white paper: How to build a brain. https://arxiv.org/abs/2502.13161

Riggs, W. D. (2003). Balancing our epistemic goals. *Noûs*, *37*(2), 342–352.

Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, *18*(2), 173–187.

Schmidhuber, J. (2008). Driven by compression progress: A simple principle for discovering interestingness [also available as arXiv:0812.4360 / LNCS chapter (see links)]. In *Proceedings / lncs (selected venues) or as arxiv preprint*. https://arxiv.org/abs/0812.4360

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (2010). *IEEE Trans. on Auton. Ment. Dev.*, *2*(3), 230–247. https://doi.org/10.1109/TAMD.2010.2056368

Schurz, G., & Lambert, K. (1994). Outline of a theory of scientific understanding. *Synthese*, *101*(1), 65–120. https://doi.org/10.1007/bf01063969

Solomonoff, R. J. (1964). A formal theory of inductive inference. part i. *Information and Control*, *7*(1), 1–22.

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, *44*(3), 510–515. https://doi.org/10.1016/j.shpsa.2012.12.005

Suñé, A., & Martínez, M. (2021). Real patterns and indispensability. *Synthese*, *198*(5), 4315–4330. https://doi.org/10.1007/s11229-019-02343-1

Thomasson, A. (2025). *Rethinking metaphysics*. Oxford University Press.

Thoren, V. E. (1991). *The lord of uraniborg: A biography of tycho brahe*. Cambridge University Press.

Turri, J. (2015). Understanding and the norm of explanation. *Philosophia*, *43*(4), 1171–1175. https://doi.org/10.1007/s11406-015-9655-x

Voelkel, J. R. (2002). *The composition of kepler's astronomia nova*. Princeton University Press.

Westfall, R. S. (2015). *The life of isaac newton*. Cambridge University Press.

Wilkenfeld, D. A. (2018). Understanding as compression. *Philosophical Studies*, *176*(10), 2807–2833.

Williams, B. (2002). *Truth and truthfulness: An essay in genealogy*. Princeton University Press.

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Ed.). Wiley-Blackwell.

Wolfram, S. (2018). Logic, explainability and the future of understanding. https://writings.stephenwolfram.com/2018/11/logic-explainability-and-the-future-of-understanding/

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Zagzebski, L. (1996). *Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge University Press.

Zagzebski, L. (2001). Recovering understanding. *Knowledge, truth, and duty: Essays on epistemic justification, responsibility, and virtue*, *2001*, 235–252.

Zenil, H. (2019). Compression is comprehension and the unreasonable effectiveness of digital computation in the natural world. In *Unravelling complexity* (pp. 201–238). https://doi.org/10.1142/9789811200076_0011