

Ontology of Data

(Boyan Dafov, University of Sofia)

Abstract: This study offers a systematic ontological investigation of the concept of data. Drawing from the triadic distinction between data, information, and knowledge, it proposes a definition of data as abstract units for the structured retention of availability. Central to the analysis is the role of interpretation, which mediates the transition from the objective being of data to the subjective being of information. The concept of the non-being of data is developed through the dialectic of absolute randomness and absolute determinacy, analyzed in light of Shannon entropy as a measure of interpretive potential. In the final section, the study introduces the novel category of modal data, defined as data that retain not only what *is*, but also what *could be*, *must be*, or *might not be*. This opens new horizons for logical, epistemological, and computational analysis. The study synthesizes insights from ontology, logic, hermeneutics, and information theory.

Keywords: ontology, information theory, data, interpretation, modality

Introduction

The question of how we ought to understand—and consequently, what kind of value we ought to ascribe to—what is today (especially in recent discourse) widely referred to as “data” is of critical importance. This is not only due to its centrality in contemporary conversations surrounding so-called “artificial intelligence,” but also in a much broader context—for instance, in cases where a decision is justified with the claim that “the data indicates so”.

The problem explored in this study is of exceptional significance, particularly given the growing tendency to treat “data” as the highest form of legitimacy in numerous debates, serving as an almost indispensable foundation for any argument that lays claim to being “rational”. A proper understanding of the nature of what data mean can help us make high-quality judgments about when, how, and under what conditions it is appropriate—or inappropriate—to employ or appeal to them.

This paper attempts to address the ontological structure of what

ought to be called “data,” to delineate what is “non-data” and to identify the moment at which data cease to be merely data—that is, when they move beyond their mode of simple availability and become something else: namely, the transition from data to “information”.

The clarification of these two transitions will aid in illuminating the essence of data by revealing, first, the shift from merely determinate being (*Dasein*) to essential being (*Wesen*), along with the disclosure of the non-being (*Nichtsein*) of the object under consideration; and second, the transition from the objective being of data to the subjective being of information—a transition that could be topic of a more extensive inquiry beyond the scope of the present study, which remains focused on a form of object-oriented ontology.

Clarifying the Essence and Defining the Concept

While it is important to emphasize that information depends on data, it is more customary to understand the former in terms of the latter, rather than the reverse.¹ Thus, the question of how data are to be understood *in themselves* remains an open one. At the same time, not all data serve as a source of information in the usual sense, where information is equivalent to some content about a given referent. A music compact disc, for example, may contain gigabytes of data, yet no information *about* anything.²

Returning to the practical motivation for exploring this problem, it is precisely in clarifying these transitions—and in systematically exposing the relations between “non-data” and data, as well as between data and information—that the true horizon of validity of the otherwise too-broadly and often uncritically applied concept of data will be revealed. Why is it, for instance, that in some cases—such as the diagnostic analysis of a phenomenon in the scientific, social, political, or economic domain—we can confidently rely on data, while in others it would be, to say the least, absurd to grant them any relevance for decision-making or judgment? Such examples would include the choice of a romantic partner, aesthetic evaluation of art or other forms of creativity, or the resolu-

¹ Floridi, Luciano. “Information.” In *The Blackwell Guide to the Philosophy of Computing and Information*, edited by Luciano Floridi, 40–61. Oxford: Blackwell, 2003.

² Floridi, Luciano. “Is Information Meaningful Data?” *Philosophy and Phenomenological Research* 70, no. 2 (2005): 351–370.

tion of moral and ethical dilemmas. The answer to such questions helps us grasp the intrinsic limits contained within the objective being (*objektives Sein*) of data, along with the essential characteristics and constraints this being entails.

Let us begin with the etymology of the word. The Latin *data* is the plural of *datum*—a “given”—which in turn is the past participle of *dare*, “to give.” Already at this level, we can provisionally infer that data represent a kind of “retention” (in the sense of the distinction between “retention” and “apprehension,” which parallels the difference between *determinate being* (*Dasein*) and *essential being* (*Wesen*)) of some availability or presence. More concretely, let us formulate the following provisional working definition of data: *Data are abstract units that serve to “retain” raw, unprocessed facts of availability, which can be collected, stored, processed, and analyzed. They can take various forms, structuring this act of retention.*

Through data, we retain elements of availability, granting them a certain form or structure—but not *essence*. We merely hold them as they are. Therefore, through data, we are not meant to do anything beyond preserving certain facts of availability.

Another interesting aspect of the being of data becomes visible in the basic division widely adopted in Data Science, Experimental Psychology, Sociology, and other disciplines that work intensively with data—namely, the fundamental distinction between *qualitative* and *quantitative* data. From both a practical and a theoretical point of view, it seems that data can only be either qualitative or quantitative, but not both at once, nor something entirely outside these two categories.

By delving into this strict and irreducible division through the lens of the first part of Hegel’s *Science of Logic*—the “Doctrine of Being” (*Seinslogik*)—we can further substantiate the interpretation of data as *determinate being* (*Dasein*). This becomes clear through, on the one hand, the qualitative determination of the being of data—the determinacy that makes being what it is³—and, on the other, the quantitative determination of being—the aspect that admits of change without essential alteration.⁴ The impossibility of synthesizing these two dimensions into a unified concept—of grasping them in their speculative unity as

³ Georg Wilhelm Friedrich Hegel, *The Science of Logic*, trans. and ed. George di Giovanni (Cambridge: Cambridge University Press, 2010), 59–83.

⁴ Hegel, *Science of Logic*, 154–167.

*measure*⁵—reveals the fundamental limitation of data: their incapacity to rise to the level of *essential being* (*Wesen*). In this sense, data remain confined to the ontological register of pure availability.

The Systematic Place of the Concept and the Transition from Objective to Subjective Being

Let us return, then, to the notion of data as the retention of availabilities. This immediately raises the question: why would we need to *retain* facts of availability at all—facts which, in certain colloquial contexts, might be referred to as “raw reality”? The answer to this question will help us uncover the *telos* of the being of data precisely as the retention of availabilities. What do we intend to do with something that we first need to retain?

One possible answer—within the conceptual horizon of this investigation—is that what we ultimately aim for is the possibility of *interpreting* that which has been retained, at some future moment. In other words, this act of retention enables us to extract a fragment of availability, potentially preserve and “safeguard” it, so that it may later be interpreted and used in the generation of information—and, eventually, perhaps even in the acquisition of knowledge. Yet within the framework of data, we are strictly limited to this retention alone. The possibilities of interpretation and of knowledge acquisition are not yet realized at this level.

The problem of *interpretation* as such demands a deeper philosophical inquiry, but since it introduces us to the domain of *subjectivity*, and that lies beyond the scope of the present study, I will employ the term in its broad, common-sense usage. Nevertheless, the subjective character of interpretation—as a potentiality *immanent* in data as structured retention of availability—allows us to locate systematically the ontological status of data as object.

It is precisely at this point that we may articulate the *telos* of data: their transformation into information. This constitutes the transition from *objectivity* (*Objektivität*) to *subjectivity* (*Subjektiver Begriff*), realized through the actualization of data as information.

Up to this point, we have arrived at the following systematic understanding of data:

1. First, data are to be conceived as *abstract units* through which we

⁵ Hegel, *Science of Logic*, 302-326.

retain certain availabilities within a given structure, whose form is determined by the specific type of data in question. In this sense, data constitute an *immediate unity of form and content*⁶, which, by structuring a given availability, enables its retention, preservation, and “displacement.”

2. Second, the transition from *objectivity* to *subjectivity*, mediated by the telos of the being of data—namely, their transformation into *information*, a process that is mediated by interpretation.⁷

However, we still lack an answer to the question of the non-being (*Nichtsein*) of data. Before we can approach that issue, we must first demonstrate how interpretation—though not yet actualized within data—is nevertheless *contained within them as a latent potentiality*. This possibility of interpretation embedded in data forms the structural horizon that makes the transition from data to information ontologically thinkable.

The Non-Being of Data

What is that which is *on the way* to becoming data, but is not yet data? To address this question, let us delve more deeply into the essence of data, whose *telos* lies in the possibility of interpretation and thus in their transformation into information. The transition from information to knowledge is also philosophically compelling, yet it exceeds the scope of the present inquiry and could become the focus of a future investigation.

As we have already demonstrated, our guiding conceptualization of data is that of *retained availability*, structured in a particular form. The possibility of employing diverse structures reveals to us the multiplicity of data types—qualitative, quantitative, and others. One essential characteristic that may aid in further clarifying the nature of data is a property stemming from their *teleological orientation*: their *potential for informativeness*.

In everyday discourse, we often encounter questions such as: “*What information is in this data (on a given topic)?*” From a strictly philosophical and systematic standpoint, such a formulation is conceptually inaccurate, as *information is not something immediately contained in data*; rather, it exceeds them conceptually. But what do we mean, then, when we employ such phrases? Clearly, such statements, while

⁶ Hegel, *Science of Logic*, 396-399.

⁷ Ibid.

imprecise, do not strike us as absurd. This suggests that they gesture toward an essential trait of data. Given the distinctions articulated so far, a more accurate reformulation of such a question would be: “*Which part of the data could I interpret so as to obtain information (on the issue that currently concerns me)?*”

Hence, data contains *something* that is *amenable to interpretation*. Going further than that, the possibility of interpretation is both the *telos* and an *essential characteristic* of data. For convenience, and in accordance with colloquial usage, we will continue to refer to this as “*the information contained in the data*”, by which we mean more precisely: “*that content of the data which is interpretable with respect to a given question in a specific context*”.

Once again, we will set aside the possibility of a *hermeneutic approach* to data—particularly the one grounded in Hans-Georg Gadamer’s dialectic of *question and answer* (*Frage und Antwort*) and the living interplay between subject and object in interpretation, as developed in *Truth and Method*⁸—since this would fall within a subjective ontology of information, which lies beyond the scope of this study.

Let us return, then, to the notion of the *interpretability* of data—what we have provisionally termed “*information contained in the data*.” Here, the concept of *entropy*—originally a thermodynamic measure of disorder—becomes helpful. In a more specialized context, this concept appears in Claude Shannon’s *A Mathematical Theory of Communication*⁹, where entropy quantifies the *expected information content* of a probabilistic system. It is important to clarify that in mathematics and physics, “information” is conceived quite differently than in our current philosophical analysis. Within the systematic triad of data, information, and knowledge employed here, the scientific and mathematical notion of “*information*” more closely corresponds to *data* in our sense. In the mathematical sciences, “*information*” refers to availability (*Bestand*) without regard for the subjective character of meaning. This is precisely why we may appropriate the concept of *entropy* from information theory and apply it not to information (in the philosophical sense), but to *data*.

Now, returning to our provisional notion of “information contained

⁸ Hans-Georg Gadamer, *Truth and Method*, 2nd rev. ed., trans. Joel Weinsheimer and Donald G. Marshall (New York: Continuum, 2004), 356–384.

⁹ Claude E. Shannon, “A Mathematical Theory of Communication,” *Bell System Technical Journal* 27, no. 3 (1948): 392–401.

in data,” Shannon entropy (H) can be understood as the *measure of informativeness* of a given data set. It is a probabilistic function that reflects the *predictability* of the data. In this framework, entropy quantifies the degree of *disorder* within the probability distribution defined by the data. If the entropy of a data set is low (approaching 0), the data are highly predictable and thus carry minimal information. Conversely, high entropy indicates unpredictability and thus high informativeness. Shannon entropy values range from 0 to 1 for normalized binary systems. A value of 1 indicates a maximally random distribution—each possible outcome is equally probable. A value of 0 implies complete determinacy—any given selection from the data set yields the same result every time.

A classic illustration of these extremes is the case of a coin toss. A fair coin (50% heads, 50% tails) yields a Shannon entropy of 1: maximal unpredictability. If, however, the coin is biased to always land on tails, the resulting data become entirely predictable, and entropy drops to 0. This analysis reveals an important ontological insight: the interpretability of data is always present as a potential, and this potential may vary qualitatively, depending on the degree of determinacy or randomness. Interpretation presupposes neither total randomness nor absolute certainty. But how does this help us understand the *non-being* (*Nichtsein*) of data? We have now seen how both *absolute randomness* and *complete determinacy* negate the possibility of interpretation: the former because it offers no pattern, the latter because it offers no ambiguity. This raises a subtle question: Is *absolute determinacy* truly a form of interpretability, or is it, paradoxically, a negation of interpretation?

Consider the contrast between a literary text and mathematical proof. The literary work is neither a random sequence of signs nor a completely determinate object. Its meaning is neither fixed nor arbitrary; it exists *in between*. By contrast, a mathematical proof permits no interpretive ambiguity. One does not engage in “interpretive analysis” of a proof; its meaning is fixed in its formal structure. The reader’s task is merely to *understand*, not to *contribute* to meaning. This example illustrates a crucial point: just as *absolute randomness* forecloses interpretation, so too does *absolute determinacy*. From this we may conclude: If the *telos* of data is to become *information*, then they must neither be entirely random nor entirely fixed. For data to fulfill their function, they must point to and retain availabilities that are *neither purely arbitrary nor exhaustively determined*.

In this structured retention and pointing, there lies the potential for essence, the potential for becoming *being-for-another*¹⁰, and thus for transitioning to the *subjective being of information* through interpretation. Therefore, the non-being of data can be understood as that which *lacks the potential for interpretation*: either because it is entirely chaotic, or because it is entirely fixed. Just as in Hegel's dialectic of *pure being* (*Reines Sein*) and *pure nothing* (*Nichts*)—both of which are indistinguishable due to their shared indeterminacy¹¹—so too, in the ontology of data, the two extremes of absolute determinacy and absolute indeterminacy amount to the same ontological invalidity in relation to the *telos* of data. Both are equally devoid of content with respect to interpretability and, by extension, to the transformation of data into information. Hence, we are justified in calling both extremes forms of the *non-being of data*.

Expanding the Ontological Field through the Introduction of “Modal Data”

Let us now turn to the practical dimension of the reflections developed so far—namely, the classification of different types of data and the conceptual articulation of a novel type of data. As we have established, data are a mode of *objective determinacy*, understood as *abstract units* that serve to retain raw, unprocessed facts of presence, which may be collected, stored, processed, and analyzed. These data can take various forms—numbers, text, images, audio recordings, etc.—each structuring availability in its own way. What is common to all these forms is their role as *pure retention*, without inherent interpretation or subjective transformation. Once they are subjected to analysis or use in inference, they move beyond their purely objective character and enter the domain of *subjective being*, i.e., information.

The proposal advanced here is the conceptual introduction of a new kind of data that we shall call *modal data*. Unlike standard data,

¹⁰ Hegel, *Science of Logic*, 128-129.

¹¹ “Pure Being and pure Nothing are, therefore, the same. What is the truth is neither Being nor Nothing, but that Being—does not pass over but has passed over—into Nothing, and Nothing into Being. But it is equally true that they are not undistinguished, that they are not the same, that they are absolutely distinct yet equally unseparated and inseparable, and that each immediately vanishes in its opposite.” Hegel, *Science of Logic*, 59-60.

which retain what *is*, modal data would retain what *could be*, what *must be*, what *might not be*, or what is *impossible*—that is, they would encode presence not as actuality, but as *modality*. This form of data would integrate the classical modal categories: *possibility*, *necessity*, *impossibility*, and *contingency*. In doing so, it would open entirely new domains of analysis, processing, and epistemic extraction. Modal data could serve in the training of artificial systems (so-called “artificial intelligence”) capable of modal inferences, e.g., identifying potentialities or necessities in various contexts. They could also be deployed in simulation environments and the construction of virtual or counterfactual realities.

The conceptual foundation for this proposal draws on developments in *modal logic*, particularly the systems of C. I. Lewis¹² and the later semantics of *possible worlds* elaborated by Saul Kripke.¹³ In Lewis’s system, modal categories are treated as logical operators governed by necessity and possibility; in Kripke’s possible world semantics, these operators are given model-theoretic interpretation across accessible worlds, providing a rigorous framework for counterfactual reasoning. This semantic apparatus allows us to speak not only of what is, but of what might be the case under varying conditions. From a metaphysical perspective, this is deeply resonant with the work of Leibniz, who viewed *possibility* as grounded in the principle of *compossibility*, i.e., the logical and metaphysical compatibility of possible states of affairs.¹⁴ In this Leibnizian sense, modal data would not simply speculate, but systematize the virtual, treating potentialities as ordered elements within a conceptual universe.

In terms of ontological grounding, such data may also be thought of in connection with Hegel’s category of the Possible (*das Mögliche*) in the *Science of Logic*, where possibility is not mere abstraction but a determinate moment within actuality—“*the possible is only insofar as it belongs to the actual*”.¹⁵ Modal data would therefore not exist outside

¹² Clarence Irving Lewis, *A Survey of Symbolic Logic* (Berkeley: University of California Press, 1918), chap. 11, “Strict Implication and Modal Logic,” 366–408.

¹³ Saul A. Kripke, *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980), lecture 2, “Naming and Necessity,” 63–87.

¹⁴ Gottfried Wilhelm Leibniz, *Theodicy: Essays on the Goodness of God, the Freedom of Man and the Origin of Evil*, trans. E. M. Huggard (La Salle, IL: Open Court, 1989), part I, §§8–21, 42–60.

¹⁵ Hegel, *Science of Logic*, §§146–148, 399–415

actual data but would supplement actuality with structured virtuality.

A simple example may illustrate the core idea. Suppose we are conducting a standard survey regarding voter turnout. A typical data collection might ask: “*Will you vote in the upcoming election?*” (Yes/No).

Modal data, by contrast, would introduce structured hypothetical conditions: “*Would you vote in the upcoming election if a political party with a clear commitment to social policy reform were on the ballot?*”

In this way, modal structures are superimposed on standard data to yield *conditional and counterfactual datasets*, enabling modal inferences. Instead of stating merely “*X% of eligible voters will vote*”, one could say “*X% of eligible voters will vote, but if a party with a clear social platform were present, Y% would vote*”. This extension allows not only for descriptive statistics, but for *modal projections*—mapping the space of social possibility. Similar methods could be developed to structure the remaining modal categories into data architectures: necessity (*must vote if...*), impossibility (*would never vote even if...*), and contingency (*might vote if...*).

Such developments could transform not only data science, but the epistemology of prediction and the logic of future-oriented knowledge. In philosophical terms, they bring the *modal horizon* into the ontological fabric of data itself, redefining the boundary between what *is* and what *might be*.

Conclusion

In the present work, we have undertaken an ontologization of data. Beginning with the distinction drawn in the introduction between the concepts of *data* and *information*, we were able to identify interpretation as the central mediating function through which the transition from the objective being of data to the subjective being of information becomes possible.

The practical impetus for this inquiry lies in the increasing centrality of “*data-driven decision-making*” within contemporary political, social, and economic life. A successful investigation into the essence of data—and a clarification of their mode of being—may assist us not only in understanding them more fully, but also in discerning those contexts in which we are justified in *relying on data* as the ground of judgment or action.

Through an analysis of the triadic relation data–information–knowledge, we have seen that data represent the most abstract and ontologically empty of the three categories. Within this framework, we in-

troduced and defended the following working definition:

Data are abstract units that serve to “retain” raw, unprocessed facts of availability, which may be collected, stored, processed, and analyzed. They may take various forms, thereby structuring the retention they perform.

Building on this definition—and on our earlier explication of interpretation as a latent potential within data—we advanced toward a determination of the non-being of data. Here, we introduced the concept of entropy, initially understood (in thermodynamics) as a measure of disorder within a physical system, and subsequently reformulated in information theory as Shannon entropy, which quantifies the degree of uncertainty or unpredictability within a probability distribution.

Through this concept, we articulated the dual character of the non-being of data: on the one hand, absolute randomness, and on the other, absolute determinacy. We argued that both extremes are, with respect to the potential for interpretation, equally ontologically empty. Just as Hegel shows that the distinction between pure being and pure nothing is merely nominal—since both are equally indeterminate—so too is the distinction between total randomness and total order nominal with respect to the capacity of data to become information through interpretation.

Drawing on these ontological reflections—framed by the *telos* of data as becoming—we concluded by proposing a novel category: modal data. These would not merely retain *what is*, but also retain that which *could be*, *must be*, *might not be*, or is *impossible*—thereby expanding the ontological scope of data to include modalities of being. Such data could allow for the retention of potentialities, thus opening new teleological pathways toward the richer and more determinate categories of information and knowledge, by amplifying the interpretive potential that is intrinsic to the essence of data.