Review of Philipp Koralus' *Reason and Inquiry: The Erotetic Theory*

DANIEL.HOEK@VT.EDU, VIRGINIA TECH, OCTOBER 2023.[1]

Over the past decade, an inspiring and potent new idea has taken root in philosophy, whose implications for epistemology and for our understanding of the mind we are only just beginning to appreciate. I'm talking about the notion that the mind is not a passive data churner but an active and searching *inquirer*, driven by curiosity and wonder. The thoughts and views populating this inquiring mind are shaped as much by the questions that give rise to them as by the information that they carry. Information is no longer the sole currency of thought: the mind is abuzz with questions.

This new picture of mind and thought has potentially profound implications, because it suggests that the classical project of understanding cognition in purely informational or propositional terms was fundamentally impaired. From this new perspective, trying to understand the answers our mind settles on in abstraction of the questions that animate it seems a bit like trying to understand an overheard phone conversation without realising there is a person at the other end of the line.

Philipp Koralus was one of the pioneers of this new insight, using question-directed mental content to shed light on the psychology of human reasoning. When he and his co-author Salvador Mascarenhas first published their questioning or *erotetic* theory of reasoning back in 2013, they were amongst the first to harness the power of question-directed cognitive content, and also among the first to co-opt ideas from formal semantics to articulate that conception. Koralus' new book, *Reason & Inquiry*, marks the capstone of ten years of successive refinements, improvements and expansions of the erotetic theory. Where the original erotetic model was limited to propositional deductive reasoning, the book now covers quantificational, statistical and practical reasoning as well. (These chapters are all co-authored with computer scientist Sean Moss; Vincent Wang and Beau Mount also collaborated on a chapter.)

The most notable and impressive feature of the erotetic theory is its ability to capture not only correct reasoning, but also to explain the attractiveness of a large number of common reasoning fallacies. Essentially, the framework allows such slip-ups to be understood as artefacts arising from the application of reasoning steps that normally promote accuracy and cognitive efficiency. Moreover, the theory distinguishes itself from alternative accounts of fallacy with a story about how we are able to avoid fallacious reasoning, and reason our way back out.

With this book's expanded horizon, the erotetic theory now seeks to account for statistical and practical fallacies as well as deductive ones. Thus the ambitious task the book sets itself is to assemble a disparate set of cognitive phenomena under a single umbrella: from affirming the consequent to base rate neglect, from the endowment effect to the conjunction fallacy, and from

[1] *Mind*, online first, November 2023: 1-11. Thanks to Dan Harris, Jordan MacKenzie and Rohan Sud for helpful discussion.

the Wason selection task to framing effects. They are all to be explained within the erotetic framework. The grand vision is to establish the erotetic theory as a single unified theory of all human reasoning, accounting for both its successes and its failures.

While I am not persuaded the book realises this sweeping ambition, a lot is achieved in the course of trying. For one, I *am* persuaded that some of the parallels Koralus and his collaborators identify between deductive, statistical and practical fallacies are real. These observations yield a compelling argument for a question-directed view of cognitive content. They also demonstrate the fruitfulness of studying all these different domains of reasoning in conjunction, rather than siloing them off as is commonly done in philosophy and psychology. More broadly, the book presents an inspiring image of what a rigorous and comprehensive psychological theory of reasoning should look like (and clarifies what distinguishes such a theory from, say, a *logic*).

Besides being rich in original insights, the book is also highly informative. The exposition of the erotetic theory is everywhere accompanied by a wealth of empirical and theoretical knowledge stemming from psychology, linguistics, epistemology, psychology, neuroscience, computer science and behavioural economics. With this interdisciplinary footing, there should be something new to learn for everyone. A major boon for me were the bounteous, example-filled surveys of the relevant empirical psychological literature in each chapter. Anybody who can read this book will stand to learn a great deal from it.

It must be said, however, that this book does not part with its treasures lightly. While it is ostensibly aimed at a broad audience spanning psychology, philosophy and computer science, background in all three areas tends to be assumed. The hardest parts are the expositions of the erotetic model, which are for the most part dense, formula-heavy and confusing. These expository sections make up the bulk of all the key chapters. Due to the rigidly cumulative structure of the book, there is no good way of side-stepping the dicey parts: if you skip to a chapter or section of interest, you probably won't understand what is going on.

The book thus allows some of its best ideas to be obscured behind a near impenetrable thicket of formulas. This is made worse by inadequate signposting and misleading, idiosyncratic notation. (Why do $[\cdot]^{AP}$, $[\cdot]^{N}$, $[\cdot]^{M}$, $[\cdot]^{Q}$ and $[\cdot]^{C}$ all look the same, when they each denote a completely different type of operation, mathematically and conceptually?) Besides good ideas, the thicket hides some muddles too, and when you go through the painstaking process of deciphering the derivations and definitions, you often run into hidden assumptions, missing formal details, or inconsistencies. (In §5.4-5, *W*s appear on every line, when that operator has not yet been defined and is not supposed to occur in the relevant inference procedure.) This all makes it hard to verify many of the book's claims, and almost impossible to get enough of a grasp on the final model to play around with it, or make an independent assessment of it. Formalism really is no substitute for clarity.

Moreover, it is hard not to wonder if a simpler theory might not have been more explanatory. As you progress deeper into the book, further epicycles are added with each successive expansion

of the theory. There is never an illuminating unification or elegant resolution. By Chapters 4 and 5, the definitions are baroque, strangely disjunctive structures spanning half a page. Thus the model's complexity grows exponentially as its coverage increases linearly.

But although *Reason & Inquiry* is a frustrating book in some ways, it remains a remarkable work in its scope and vision, with a powerful idea at its core. So there are substantial rewards for the persistent reader, and in this review I will focus on those. My aim will be to distil some key takeaways from the book, and to put those in conversation with other recent work in philosophy. Section 1 provides a simplified outline of the erotetic theory, explaining the core ideas behind the model. Section 2 gives a chapter-by-chapter synopsis of the book. Sections 3 describes the book's most consequential empirical generalisation, and brings out its implicit argument in favour of question-directed cognitive content.

1. Outline of the Erotetic Theory

Koralus' stated aim for the erotetic theory is to describe human reasoning at the computational level. That is to say, we are not looking to describe the exact physical brain processes whereby reasoning is executed, but rather to identify the abstract mappings computed by those processes (cf. Marr 1977). In line with that conception of the project, Koralus proposes to understand reasoning in terms of the manipulation of the *contents* of thought, rather than the manipulation of specific inner representations of that content.

To illustrate the basic structure of the erotetic model, the book's introductory chapter sports a helpful diagram of a game controller attached to a screen (Figure 1 below, taken from p. 51 of *Reason & Inquiry*). In this metaphor, the screen shows the active mental content or *view* that is presently the object of your attention, and of your reasoning. The buttons on the game controller represent the ten or so basic mental operations you have at your disposal for manipulating that view. Each press of a button corresponds to a reasoning step. In addition to the buttons, the controller has an input source for feeding in a further piece of content which can then be brought to bear in some way on the active view on the screen. For example, you can feed in a piece of information and hit the *Update* button to *add* that information to the active view, and to rule out open alternatives where possible. If you press the *Inquire* button instead, the input proposition will be used to distinguish alternative states within the active view.

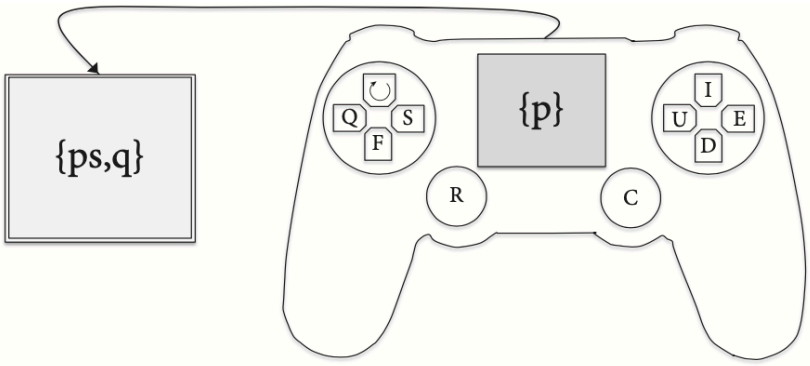

Figure 1. Koralus' game controller model of reasoning

Once you are satisfied with the results of your reasoning, you may hit the *Commit* button to save the view on the display (your conclusion) to a background stack of "cognitively available" commitments (p. 68). There is also a *Reorient* button that uploads views from the stack, erasing whatever was previously on display. This commitment stack is assumed to include any further premises that may be used in your reasoning. Later in the book, it is through the stack that reasoning interfaces with memory and background beliefs (§5.6; cf. Hoek fc, where question-guided reasoning operates directly on the belief state). On the erotetic model, reasoning is by default hypothetical, and makes only indirect contact with belief.

What makes this a model of an *erotetic* or questioning reasoner is the way that the mental contents on the screen and console are understood. While the details of how this is spelt out change from chapter to chapter, the core idea is that a view specifies a number of alternative states the reasoner is choosing between (or *wondering* between). It is in virtue of those distinguished alternatives that we can think of views as question-directed. The reasoning process is (ordinarily) aimed at whittling the set of alternatives down to one, which is just to say it aims at answering the question the agent is wondering about. The erotetic account thus fits neatly within the broader picture sketched at the start of this review, of an inquiring mind that seeks to resolve the questions that animate it (as in Friedman 2013, 2017).

Picking up on this intuitive picture, Koralus also stresses a concurrent *psychological* pressure to reduce alternatives, stemming from the cognitive burden of entertaining multiple scenarios. Moreover, the book's final chapter brings to the fore the *practical* need to reduce alternatives, rooted in the imperative to act. This practical need arises because decision problems raise questions that want answers, much in line with Van Rooij 2003 and Hoek 2022.

The final ingredient needed to get verifiable predictions out of the model are hypotheses about the particular algorithm or *inference procedure* reasoners will employ in various concrete reasoning tasks. For example, one kind of task is this: a subject is given a set of premises and asked whether a certain conclusion follows. If we want to use the erotetic model to make a prediction about how subjects will respond to this task, we need to make a specific hypothesis about which buttons they will press, which views they will input, and in what order. When the premises are given as sentences, we also need an *interpretation rule* that maps sentences to views. Here the erotetic theory makes contact with the semantic theories that inspired it, as Koralus employs the clauses familiar from inquisitive semantics and truthmaker semantics (Groenendijk and Roelofsen 2009, Mascarenhas 2009, Fine 2012).

To get a feel for the erotetic model, let's look at an example. Consider the following premises, and ask yourself what if anything follows:

> P_1. Either Jane is kneeling by the fire and watching TV or else Mark is standing at the window and peering into the garden.
>
> P_2. Jane is kneeling by the fire.

Did you say *Jane is watching TV*? Then you have fallen prey to the *illusory inference from disjunction*, as did 90% of participants in a study by Clare Walsh and Philip Johnson-Laird (2004). This inference is psychologically very attractive but invalid. Here it is in schematic form:

$$\begin{array}{l}(A \wedge B) \vee C \\ \underline{\quad A \quad} \\ \therefore B\end{array}$$

To see the invalidity, consider the case where A and C are true but B is false. This illusory inference is in many ways the flagship fallacy of the erotetic theory, and most fallacies in the book are explained by a direct or indirect analogy to this one (as we'll see below).

Roughly, here is how the erotetic model explains this illusory inference. We start out with an empty screen. Then we update with the first disjunctive premise, yielding a view with two alternative states: AB and C. Because we always want to reduce the number of alternatives, any further update will be used to try to decide between these two. So when we update with the second premise A, which confirms AB but not C, the latter alternative is discarded. Thus we end up with the view AB, which has the illusory conclusion B as a part. (The notion of confirmation in play here is not probabilistic, but is spelt out in terms of matches between atomic states.)

Besides explaining why we go in for fallacies like the illusory inference, the erotetic theory also seeks to explain how we are able to avoid them. This is an Achilles heel for many theories of fallacy. Once we have given an explanation for why an agent is prone to accept a certain invalid inference, this tends to make it difficult to explain why the agent should ever come to regard that reasoning as flawed: once employed, an inference rule like *affirming the consequent* can be used to prove its own soundness. Koralus calls this *Plato's problem*, and it can be viewed as the fallacious cousin to Kripke's Adoption Problem, which has sparked much discussion recently (e.g. Birman 2015, Devitt and Roberts 2023, Williamson fc).

The authors propose to overcome this problem by noting that the fallacious inferences no longer go through if we ask enough questions. If, prior to updating with the second premise, we had asked whether A is true supposing C is, this question would have split the scenario C into two alternatives, AC and $\bar{A}C$. The illusory inference would then be blocked, since the update with A will eliminate $\bar{A}C$ but not AC. According to Koralus, this explains why we are not irretrievably lost to the illusory inference from disjunction, though we find it attractive: a friend can convince you of your mistake just by raising the right question to salience.

The book generalises this observation, establishing a sequence of soundness results which say, in effect, that reasoners who ask sufficiently many questions will avoid any missteps. Or in the book's terminology: conclusions reached under *erotetic equilibrium* are always classically correct. Koralus nicely encapsulates these results by noting that "the picture … that emerges is quite Socratic. We can be sure that our conclusions are valid if we could still reason our way to them, regardless of how much hostile questioning we might have to endure." (p. 94)

(Note for prospective readers: there is a slight technical slip-up in Chapter 2's Soundness theorem, which as stated is not quite true. However, the error can be fixed by adding a proviso to the definition of the *Update* operation $\cup$ in this chapter. It should have been specified that an update $[D]^{\cup}$ with a view D can only proceed as defined when D is in the background commitment stack C; else $[D]^{\cup}$ is to leave the active view unchanged. Thus $[D]^{\cup}$ is not, strictly speaking, an operation on views, but on stack-view pairs. The revised definition of *Update* in Chapter 4 is restricted in this way. There too, it was meant to be understood that $[D]^{\cup}$ reduces to identity when $D \notin C$. Thanks to Philipp Koralus for providing clarification on this issue.)

2. Synopsis of the Book

After a general introduction (Chapter 1), the business of each chapter in the book is to flesh out a model along the lines just described to capture our reasoning in a given domain. Chapter 2 concerns reasoning with Boolean connectives, with special attention to the illusory inference from disjunction I just described. This chapter sets out the first version of the erotetic model, specifying a notion of mental content suitable for propositional reasoning, and defining a set of applicable operations and inference procedures. In each successive chapter, this model will be expanded and modified. Chapter 2 also establishes the first Soundness result, complemented by roughly analogous theorems in Chapters 4, 5 and 6.

Chapter 3 concerns reasoning with indicative conditionals, featuring a treatment of *affirming the consequent* and other problematic inferences involving conditionals. To model these inferences, an interpretation rule for conditionals is needed. Koralus ends up offering a couple of different rules, arguing different interpretations are required to capture different inference patterns. The protagonist of this chapter is the (in)famous Wason selection task, for which a dedicated inference procedure is introduced called *information source selection*. The chapter concludes with a brief discussion of possibility modals.

The second half of the book is co-authored with Sean Moss, beginning with Chapter 4 on reasoning with predicates and quantifiers. Vincent Wang and Beau Mount collaborated on this chapter as well. The fallacies treated in this chapter are mostly just direct analogs of fallacies from the previous chapters, but there is also a section on reasoning with generics. To handle quantification, the authors invoke Kit Fine's notion of arbitrary objects. Correct handling of these objects is a subtle matter, and ends up requiring a complete overhaul of the formalism set up in Chapter 2, including a newly tailored notion of mental content, as well as more complex definitions for all the basic operations.

Chapter 4's section on generics is worth highlighting, as it gives a nice illustration of the book's interesting perspective on the rational status of fallacies. Sarah-Jane Leslie noted that bare plural generics like "dogs are lazy" have strikingly asymmetric inferential properties: such claims are accepted on slim evidence, and yet seem to license strong inferences (Leslie 2008). The authors sketch an original and intriguing explanation of this asymmetry. On their account, all bare plural generics express necessary truths, which is why they are so easily accepted. The twist is

that while the content of "dogs are lazy" is vacuous, the erotetic model allows it to serve as a springboard for invalid inferences to substantive conclusions like "John's dog is lazy". Such inferences *ex nihilo* are clearly fallacious in the sense of being invalid. The authors argue that in the right context they are nonetheless both useful and reasonable (part perhaps of our ordinary ability to generalise from particulars).

Chapter 5 is about statistical and probabilistic reasoning. Among the phenomena targeted there are base-rate neglect and the conjunction fallacy. The chapter also casts light on another very interesting category of inferences, namely those that draw probabilistic conclusions on the basis of non-probabilistic premises. For instance, when told the card is either yellow or brown, people will typically conclude that the card is 50% likely to be yellow. To expand the erotetic theory's reach into this new arena, another wholesale overhaul of the model is required. The core innovation this time is to supply the alternatives in a view with weights.

Finally, Chapter 6 is about practical reasoning or decision making. This chapter has a distinctly Gibbardian flavour, in that making decisions is understood in the first instance as answering the question *What to do* (cf. Gibbard 1990). Decision makers' selection of a course of action is amalgamated with the broader project of choosing between cognitive alternatives, which allows Koralus and Moss to bring their erotetic model to bear. The fallacies targeted in this chapter derive chiefly from the behavioural economics literature, including discussions of a number of different framing effects and the endowment effect.

As a digital companion to the book, the authors are currently developing a Python package that can be used to calculate inferences in the erotetic theory.

3. The Mother of All Fallacies?

Probably the most significant empirical generalisation in the book is this: a seemingly disparate range of fallacies have a structural similarity to the illusory inference from disjunction, in that they can be quite naturally explained in terms of the discarding of unconfirmed disjuncts/ alternatives. In this section, I will flesh out that pattern, and explore how it relates to some recent developments in epistemology.

First, recall the illusory inference from disjunction. From $(A \wedge B) \vee C$ and A, people are strongly inclined to conclude B. The erotetic explanation runs as follows: the disjunction presents us with alternatives AB and C, and since A confirms one but not the other, we discard the latter and are left with AB.

We can explain *Affirming the consequent* analogously. From "If A then C" and C, people often conclude A. The major premise presents us with alternatives $\bar{A}$ and AC. Because the second premise C confirms only the second alternative, we are inclined to discard the first and are thus left with AC. An essential component of this explanation is the interpretation of the conditional as presenting us with these particular alternatives. But that interpretation is natural within the parameters of the erotetic framework, and independently motivated.

Moving on to probabilities, consider the celebrated conjunction fallacy (Tversky and Kahneman 1983). We are asked to rank a number of alternative propositions about Linda in order of probability, based on the following evidence: "Linda is outspoken and bright. As a student, she majored in philosophy and was deeply concerned with discrimination and social justice." Most people rank the alternative *Linda is a bank teller* as less probable than *Linda is a bank teller and is active in the feminist movement*. But since the former is a conjunct of the latter, it cannot really be less probable. As first noted by Sablé-Meyer and Mascarenhas (2022, p. 586), there is a clear analogy to the illusory inference. We are presented with alternatives *B* (bank teller) and *FB* (feminist bank teller). Our evidence confirms the latter but not the former, whence we are inclined to discard or at least disprefer the former.

For a decision-theoretic example, consider the decoy effect. When given a choice between an online magazine subscription for $59, a print subscription for $125, and a combo print+online subscription also for $125, people are much more likely to choose the combo option than when the print-only option (the "decoy") is omitted (Ariely 2008, p. 4-6). This is widely thought to be linked to the fact that in the three-way choice, the combo option has clear dominance over one of the alternatives, while in the two-way choice neither the online option nor the combo unambiguously dominates the alternative.

Here is how Koralus and Moss propose to capture this with the erotetic theory. The decision problem in the decoy case confronts us with a question: *Which subscription is the best deal?* Thus we start off our deliberation with three alternatives up on the screen: online is best, print is best, or combo is best. We quickly observe that *combo is better than print*, as you get more for the same price. So we update with that information. Since this confirms only the third disjunct, the other two are discarded, and we end up with the conclusion that the *combo is best*. Koralus and Moss formalise it as follows (p. 279):

$$\frac{(O>C \wedge O>P) \vee (P>C \wedge P>O) \vee (C>O \wedge C>P) \qquad (C>P)}{\therefore (C>O \wedge C>P)}$$

This construal renders the analogy to the illusory inference immediate: we update with a conjunct of one of the disjuncts, resulting in the other disjuncts being discarded.

As these brief sketches make clear, particular assumptions are needed to fit each example to the general pattern. But to me at least, it seems like the authors are onto something here. If they are, and if an explanation along these lines does account for this broad pattern of fallacious reasoning, two things follow independently of whether we want to go in for a wholesale adoption of the erotetic model. First, there must be some distinguished set of alternatives that subjects consider at any given stage of the reasoning process. Second, there is pressure to *choose* between these alternatives, or in other words to settle the question those alternatives represent.

So assuming the pattern that Koralus and collaborators identify holds up, it seems to me that this yields an important piece of psychological evidence for the idea that cognitive contents are

question-directed, as argued on other grounds by e.g. Yalcin 2018, Drucker 2020, Hoek 2022 and Holguín 2022.

The parallels with Ben Holguín's account of believing and guessing are especially suggestive. It is natural to sum up the common core of the fallacies discussed in this section as follows: in all these cases, reasoners make a *guess* about the right alternative based on inconclusive evidence. This is grist to the mill of Holguín's contention that guesses play a more pervasive role in belief formation than has generally been acknowledged. Much as Holguín suggests that guessing is a normal and frequently reasonable way to acquire beliefs (cf. also Dorst and Mandelkern 2023), so *Reason & Inquiry* maintains that jumping to conclusions is not always a cardinal sin, but a normal component of healthy cognition.

References

Ariely, Dan, 2008, *Predictably Irrational*. New York: Harper Perennial.

Birman (formerly Padro), Romina, 2015, *What the Tortoise Said to Kripke: The Adoption Problem and the Epistemology of Logic*. PhD Dissertation, CUNY.

Devitt, Michael, and Jillian Rose Roberts, 2023, "Changing Our Logic: A Quinean Perspective." *Mind*, online first.

Dorst, Kevin and Matthew Mandelkern, 2023, "Good Guesses." *Philosophy and Phenomenological Research* 105 (3): 581-618.

Drucker, Daniel, 2020, "The Attitudes We Can Have." *The Philosophical Review* 129(4): 591-642.

Fine, Kit, 2012, "Counterfactuals without Possible Worlds." *Journal of Philosophy* 109(3): 221-46

Friedman, Jane, 2013, "Question-Directed Attitudes." *Philosophical Perspectives* 27(1): 145-74.
2017, "Why Suspend Judging?" *Noûs* 51(2):302-326.

Gibbard, Allan 1990, *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

Groenendijk, Jeroen and Floris Roelofsen 2009, "Inquisitive Semantics and Pragmatics." *Proc. of the ILCLI International Workshop on Semantics, Pragmatics and Rhetoric*: 41-72.

Hoek, Daniel 2022, "Questions in Action." *Journal of Philosophy* 119(3): 113-143.
fc, "Minimal Rationality and the Web of Questions." In: *Unstructured Content*, eds. Kindermann, Van Elswyk and Egan. Oxford: Oxford University Press.

Holguín, Ben, 2022, "Thinking, Guessing and Believing." *Philosophers' Imprint* 22(6).

Koralus, Philipp, and Salvador Mascarenhas, 2013, "The Erotetic Theory of Reasoning" *Philosophical Perspectives* 27(1): 312-365.

Leslie, Sarah-Jane, 2008, "Generics: Cognition and Acquisition." *The Philosophical Review* 117(1): 1-47.

Marr, David, 1977. "Artificial intelligence—a personal view." *Artificial Intelligence*, 9(1): 37–48.

Mascarenhas, Salvador, 2009, *Inquisitive semantics and logic*. Master's Thesis, University of Amsterdam.

Rooij, Robert van 2003, "Questioning to Resolve Decision Problems." *Linguistics and Philosophy* 26: 727-763.

Sablé-Meyer, Mathias and Salvador Mascarenhas, 2022, "Indirect illusory inferences from disjunction: a new bridge between deductive inference and representativeness." *Review of Philosophy and Psychology* 13: 567–92.

Tversky, Amos and Daniel Kahneman, 1983, "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment." *Psychological Review* 90(4): 293-315.

Walsh, C.R. and P.N. Johnson-Laird, 2004, "Co-reference and Reasoning." *Memory & Cognition* 32: 96–106.

Williamson, Timothy, fc, "Accepting a Logic, Accepting a Theory." In: *Saul Kripke on Modal Logic*, eds. Romina Birman and Yale Weiss, New York: Springer.

Yalcin, Seth, 2018, "Belief as Question-Sensitive." *Philosophy and Phenomenological Research*, 97(1): 23-47.