1 Introduction

Imagine that you're a detective working on a murder case. You have collected a large body of evidence, such as fingerprints from the crime scene and witnesses' testimonies. Based on the evidence, you believe that John is the murderer. Suppose that your belief is in fact rational. But then you are reminded that you haven't slept for days, and you know that sleep deprivation undermines people's abilities to assess evidence. Should you thereby have less confidence that John is the murderer?

This question lies at the heart of the debate on so-called higher-order evidence (HOE). Your first-order evidence regarding a belief is information that's directly about the content of your belief – it's evidence that makes the content of your belief more (or less) likely to be true. In contrast, HOE is evidence not directly about the content of your belief, but evidence about your first-order evidence or about your belief-forming process.

The central question surrounding the HOE debate is this: what exactly is the significance of HOE on our first-order beliefs? This Element aims to provide a critical review of the debate and defend a novel answer to the central question. In this introductory section, I first explain the notion of HOE and what exactly the debate on HOE is about (Sections 1.1 and 1.2). Then I explain why the debate is important and why the issue surrounding HOE is puzzling (Sections 1.3 and 1.4).

1.1 What's Higher-Order Evidence?

Roughly speaking, HOE is evidence not directly about the content of your belief, but evidence about your first-order evidence or about your belief-forming process. Before I characterize HOE more precisely, let's see two more classic examples of HOE from the literature:

> Drug
> I am a student in a logic class. I believe that I have just solved the logical puzzle given by the professor. But then a classmate told me that into the coffee I just had was slipped some drug that undetectably harms one's logical reasoning ability. People affected by the drug only have a 50% chance of correctly solving the logic puzzle. (Christensen, 2010, p. 187)

> Hypoxia
> A pilot is considering whether he has enough fuel to make it to Hawaii. Based on his past experience and his calculation of how much fuel is needed, the pilot rationally believes that he can make it to Hawaii. But then he is reminded that he is in a state of hypoxia, a condition that often undetectably harms pilots' reasoning, so that they reach the correct conclusion only 50% of the time. (Elga, n.d.)

Although scholars agree that these are typical cases of HOE, they don't characterize HOE in exactly the same way. Some characterize HOE as evidence about the *rationality* of one's belief. For example, according to Christensen (2010, p. 185) and Lasonen-Aarnio (2014, pp. 315–16), the HOE in these examples is evidence that one's belief is 'rationally sub-par' or evidence of one's 'rational failure.' Others describe HOE not as evidence about the rationality of one's belief but as evidence about one's *reliability*, namely, evidence about the objective probability of one's reaching a true belief on the basis of one's evidence (Christensen, 2016b, p. 397). Other scholars focus not on evidence about the rationality of one's belief or the reliability of the agent, but on evidence about *one's evidence*; for instance, Feldman (2005) and Worsnip (2018) use the term 'HOE' to describe evidence about the *evidential support relation* in question, or evidence about *what evidence one in fact has*.

These characterizations differ in some important respects. First, evidence about one's rationality or reliability is more 'agent-focused' than evidence about one's evidence, in the sense that they are more concerned with the agent's cognitive abilities than with the agent's evidence. For instance, one might get evidence that one's belief is formed through a process like random coin-flipping; this is evidence that one's belief is formed through an irrational and unreliable process, but it doesn't provide any evidence concerning what evidence one has or whether one's evidence supports the believed proposition – if I know that your belief that H is formed through random coin-flipping, I don't thereby get to know that your evidence doesn't support H.[1]

Second, evidence about the rationality of a belief can also come apart from evidence about the reliability of the believer.[2] If rationality is essentially the same thing as reliability, as reliabilists tend to think, then the distinction between the two kinds of evidence collapses. But if rationality is essentially a matter of conforming to what one's evidence supports, as evidentialists tend to think, then evidence about reliability will be broader than evidence about rationality. To the extent that irrational beliefs are often unlikely to be true, evidence of irrationality will often also be evidence of unreliability. But evidence of unreliability is not limited to evidence of irrationality. For example, suppose that you are a doctor, and you learn that you've been slipped some drug that makes you unreliable in noticing crucial symptoms of your patient. That is, the HOE says that you have cognitive deficiencies in *collecting* evidence, not

[1] See Christensen (2010)'s discussion of the 'agent-relativity' feature of HOE.

[2] See Christensen (2016a)'s distinction between the two kinds of evidence in the context of peer disagreement. He makes a distinction between disagreement with a *rationality-peer* and disagreement with an *accuracy-peer*, and argues that they provide evidence of irrationality and evidence of unreliability respectively.

cognitive deficiencies in *analyzing* the evidence you have. Such deficiencies in collecting evidence will also make you unlikely to give a correct diagnosis. So, in this case, the HOE is evidence of unreliability, but not evidence of irrationality (in the evidentialist sense of irrationality).

In sum, the term 'HOE' has been used to refer to the following different kinds of evidence: (a) evidence about the rationality of one's belief, (b) evidence about one's reliability, (c) evidence about what evidence one has, and (d) evidence about what one's evidence supports. In the detective example, your evidence is of kind (a) and (b). Clearly, that your ability in assessing evidence is impaired by sleep deprivation is not first-order evidence: it doesn't bear on whether John is the murderer. It's also not HOE of kind (c) and (d): your cognitive impairment due to sleep deprivation doesn't bear on what evidence you have or whether the evidence actually supports the proposition that John is the murderer. Rather, your cognitive impairment is evidence that, no matter what first-order evidence you have and no matter what it actually supports, it's unlikely that you can correctly assess your evidence, and thus unlikely that you can form a true belief in the circumstance.

In this Element, I will use the term 'HOE' to cover all these kinds of evidence, since one of its aims is to provide a broad review of the literature.[3] That said, some of the claims I make in the first part of this Element (Sections 1 and 2) will look more plausible for some kinds of HOE than for other kinds. I will clarify this matter in due course. And in the second part of this Element (Sections 3 to 6), where I discuss Calibrationism, I will focus on HOE about reliability, since the current discussion of Calibrationism has focused more on this kind of HOE than on other kinds.

1.2 What's the Higher-Order Evidence Debate?

The central questions in the HOE debate are these: what's the normative significance of HOE? In particular, does evidence of cognitive impairment defeat the rationality of one's first-order belief? If it defeats the rationality of one's first-order belief, how exactly does it do so?

Why are these questions important? There are at least two reasons. First, the issue of how to understand the significance of HOE is closely connected to other important debates in epistemology and in ethics. For instance, the issue lies at the heart of the debate on peer disagreement and the debate on irrelevant influences (Schoenfield, 2014; Street, 2006; Vavova, 2018; White, 2010).[4] Second, the issue is

[3] Existing introductions to the HOE debate include Whiting (in press) and Dorst (in press).

[4] The debate on HOE is also closely connected to the debate on moral uncertainty, which concerns what we morally should do when we are uncertain about what we morally should do. See Bykvist (2017) for a review of the debate.

itself puzzling: an intuitive position on the debate, which claims that HOE has a significant impact on one's first-order belief, creates a puzzle, and attempts at solving the puzzle have led scholars to endorse various important and surprising conclusions about epistemic rationality.

In what follows, I first explain how the debate on HOE influences other debates, and then I explain the puzzle surrounding HOE.

1.3 Connections with Other Debates

The debate on HOE originates from the debate on peer disagreement. Peers are defined to be persons with roughly the same evidence and roughly the same abilities in assessing evidence. Regarding the question of how to respond to peer disagreement, a popular view, called 'Conciliationism,' says that you should revise your doxastic attitude in the direction of your peer's (Christensen, 2009). Most importantly, it says that you should revise your doxastic attitude regardless of whether your attitude is a rational response to the original evidence. An influential argument for Conciliationism goes as follows (Christensen, 2009, p. 757). Suppose you believe that *p* and you learn that a peer believes that *not-p*. Then,

P_1: Learning of the peer disagreement requires you to be no more than 50 percent confident that you have correctly evaluated your evidence.
P_2: If you should be no more than 50 percent confident that you have correctly evaluated your evidence, you should withdraw your belief that *p*.
So,
C. Learning the peer disagreement requires withdrawing your belief that *p*.

Premise P_1 is motivated as follows. Given that you are peers, when you disagree, you should think that the probability that you have correctly evaluated your evidence equals the probability that your peer has; assuming that these two possibilities are mutually exclusive (i.e., you and your peer cannot both have correctly evaluated your evidence),[5] you cannot be more than 50 percent confident that you have correctly evaluated your evidence. Premise P_2 is motivated by a so-called level-bridging principle. A common formulation of this principle says that one's attitude on the higher-order proposition about whether one's belief is rational should match one's attitude on the relevant first-order proposition. So, if you should be no more than 50 percent confident that you have correctly evaluated your evidence, you should no longer believe that *p*.

[5] This assumption is controversial. See the debate on the so-called Uniqueness Thesis (Kelly, 2014; White, 2005).

This argument for Conciliationism essentially says that peer disagreement provides a kind of HOE, and that HOE has a significant impact on one's first-order belief. The main alternative to Conciliationism, known as the Steadfast View, denies the impact of HOE on one's first-order belief. It says that what doxastic attitude is rational is primarily determined by what one's first-order evidence supports. So, how the significance of HOE is understood greatly affects the debate between Conciliationism and the Steadfast View.

The HOE debate is also tightly connected to the debate on irrelevant influence (Vavova, 2018; White, 2010) This debate concerns this question: When one learns that one's beliefs are influenced by factors that are irrelevant to the truth (factors such as evolutionary forces or one's upbringing), should one retract one's beliefs? How this question is answered clearly depends on how the significance of HOE is understood, since evidence about the origin of one's belief is a kind of HOE – it's not evidence bearing on the content of one's belief, but on whether one's belief is formed in a reliable or rational way.

1.4 A Puzzle about Higher-Order Evidence

The issue of how to understand the significance of HOE is puzzling. To illustrate, let's consider the following dramatic case of HOE. Suppose that you initially believe that *p* and suppose that this belief is rational because it's supported by your first-order evidence E. Then you receive a piece of HOE saying that you have been slipped a drug that causes people to make judgments about *p* through a process like random coin-flipping, although it seems to those people that their judgments are based on rational assessment of evidence.

In this case, the HOE seems to require you to give up your initial belief. However, it's unclear why the HOE is able to do so. After all, evidence that you form judgments about *p* in a random coin-flipping way is not evidence against *p*; nor is it evidence undermining the support relation that E bears on *p* – that you are a coin-flipper just says nothing about the evidential connection between E and *p*. The evidential connection is some *logical, statistical, or explanatory relation* between E and *p*. Such relations are not affected by matters of how you form beliefs about *p*. For instance, suppose that you originally believe *p* because you think that E is best explained by the truth of *p*; then when you get the HOE saying that you are drugged, you don't get to say, 'E is *not* best explained by the truth of *p*.' Whether E is best explained by *p* seems to have nothing to do with whether you are drugged.

To further explain this point that the HOE doesn't undermine the evidential connection between E and *p*, let's fill in this drug case with more details. Imagine that you are a detective deliberating on whether Jack is the murderer.

Your evidence includes fingerprints collected from the crime scene and witnesses' testimonies. After carefully analyzing the evidence, you come to believe that the evidence is best explained by Jack's being the murderer. Then you receive HOE saying that you are drugged so that you form beliefs about homicide cases in a random coin-flipping way, even though things seem perfectly normal to you. Clearly, you should not now think, 'I guess the fingerprints and the witnesses' testimonies are not best explained by Jack's being the murderer after all.' In contrast, in a classic undercutting-evidence case, the evidential connection is undermined and a change in your view of the evidential connection is reasonable. For instance, if instead of being told that you have been drugged, you are told that the fingerprints are planted and the witnesses are unreliable (say, they have poor eyesight), then this *will* undermine the explanatory relation in question, and you should now think that the original evidence is not best explained by Jack's being the murderer.

In this case, the evidence that you are being drugged so that you are effectively a random coin-flipper is a typical case of HOE about an agent's unreliability. This kind of HOE says that you are unlikely to reach a true belief about the proposition in question due to a drug, sleep deprivation, hallucination, or some other condition that impairs cognitive abilities, without being evidence about whether the proposition is true or whether it's supported by your original evidence. To put it another way, this kind of HOE is evidence about *your* cognitive ability. That *you* are unlikely to reach a true belief about the proposition or about the evidential connection in question tells us little about whether the proposition is true or whether the evidential connection is there.

Let's return to the explanation of why HOE creates a puzzle. I have explained that, in the drug case, your HOE *doesn't* undermine the original evidential connection – that is, your original evidence E *still* supports p. But if E still supports p, and if the HOE doesn't provide any evidence against p, then it seems that your new total evidence E&HOE also supports p. Then how come you are required to give up the belief that p? So, to say that the HOE requires you to give up your initial belief seems to conflict with a broadly evidentialist requirement on rational beliefs.

But a plausible case *can* be made for our intuition that you should give up your initial belief when gaining the HOE. It seems that the HOE requires you to adopt a higher-order belief 'it's unlikely that I form a correct belief about p.' But if you hold this belief, it seems that you should not continue to believe that p. The idea is that it seems self-incoherent to continue to believe that p while believing that the cognitive process underlying that belief is essentially a random coin-flipping.

This discussion reveals a conflict between the following two plausible principles of epistemic rationality:[6]

Evidentialism

One's belief that *p* is rational if and only if it's supported by one's total evidence.

Bridging

It's irrational for one to believe that *p* and also believe that the cognitive process underlying the belief is unreliable.

The two principles seem to be in conflict because it seems that your total evidence E&HOE can both support *p* and support a higher-order proposition saying that the cognitive process underlying the belief *p* is unreliable. As we have seen in the drug case, your total evidence E&HOE still supports *p*, since the HOE doesn't undermine the support relation between E and *p* and since it doesn't provide any evidence against *p.* But E&HOE also supports the relevant higher-order proposition: the HOE supports this higher-order proposition, and the first-order evidence E only bears on whether *p* and says nothing about your reliability or any other person's reliability. So, in the drug case, obeying Evidentialism with regard to both the first-order belief *p* and the higher-order belief requires violating Bridging.

Now, as I've explained in Section 1.1, there are various types of HOE discussed in the literature: evidence about one's reliability, evidence about the rationality of one's belief, evidence about what evidence one has, and evidence about the relevant evidential connection. The discussion has explained how HOE about one's reliability can generate a puzzle. While I think that the puzzle is most salient for this type of HOE, some scholars have argued that a similar puzzle – namely, a conflict between Evidentialism and a similar level-bridging principle – can also be generated by other kinds of HOE. For instance, Worsnip (2018) has argued that the puzzle arises for HOE about the evidential connection between E and *p*. According to Worsnip, the puzzle arises because one's total evidence can be misleading about what it supports; that is, one's evidence can both support *p* and support that it doesn't support *p*, so that obeying Evidentialism can lead to violation of the following plausible principle:

Enkrasia

It's irrational for one to believe *p* and also believe that one's evidence doesn't support *p*.

[6] Another interesting puzzle surrounding HOE is 'the Fumerton's Puzzle' for theories of rationality, which says that if HOE affects one's first-order belief, then there can be no sufficient condition for rationality. See Foley (1990), Lasonen-Aarnio (2014), Sepielli (2014), and Ye (2015) for discussion of the puzzle.

You might wonder whether your total evidence can really be misleading about what it supports; that is, can it really support *p* but also support 'my total evidence doesn't support *p*?' The answer here is not as clear as it is in the case of HOE about one's unreliability. While it's intuitive that evidence of one's unreliability (e.g., evidence that one is a coin-flipper) doesn't undermine the evidential support E bears on *p*, evidence that E doesn't support *p* does seem to undermine the support.

However, at least on some prominent understanding of evidence and the evidential support relation (e.g., Williamson's knowledge conception of evidence and the 'high evidential probability' conception of evidential support), your total evidence can indeed be misleading about what it supports. A famous case illustrating this possibility is Horowitz's (2014) 'Dartboard' case, which originates from Williamson's 'Unmarked Clock' case (Williamson, 2000, p. 229):

> Dartboard
> You have a large, blank dartboard. When you throw a dart at the board, it can only land at grid points, which are spaced one inch apart along the horizontal and vertical axes. (It can only land at grid points because the dartboard is magnetic, and it's only magnetized at those points.) Although you are pretty good at picking out where the dart has landed, you are rationally highly confident that your discrimination is not perfect: in particular, you are confident that when you judge where the dart has landed, you might mistake its position for one of the points an inch away (i.e. directly above, below, to the left, or to the right). You are also confident that, wherever the dart lands, you will know that it has *not* landed at any point farther away than one of those four. You throw a dart, and it lands on a point somewhere close to the middle of the board. (Horowitz, 2014, p. 736)

Suppose the dart landed at point <3,3>, and consider the proposition *Ring*: the dart landed on one of <3,2>, <2,3>, <4,3>, or <3,4>. Then your evidence supports a high 0.8 credence in *Ring*, since it's true in four out of five epistemic possibilities, and we assume that each epistemic possibility is equally likely. But your evidence also supports a high 0.8 credence in 'my evidence supports a 0.2 credence in *Ring*,' since this higher-order proposition is also true in four out of five of your epistemic possibilities. So, if we claim that your evidence supports *p* just in case its evidential probability is not lower than 0.8 (which is an arbitrary choice, since we can construct a similar case for other thresholds), this is a case where your total evidence is misleading about whether it supports a proposition *p*. (For further discussion of cases like Dartboard, see Elga [2013], Skipper [2019], and Worsnip [2018].)

So, it seems that, just like Evidentialism can come into conflict with Bridging, it can also come into conflict with Enkrasia, and this conflict is puzzling since

these principles are all intuitively plausible. However, I should note that some authors have expressed doubts about the alleged puzzle. For instance, Skipper (2021) has argued that, although one's HOE can be misleading about one's *first-order evidence* supports, it can never be misleading about what one's *total evidence* supports. Dorst (2020) has argued that if we replace Enkrasia with a more plausible level-connecting principle that he calls 'Trust,' then we no longer have a conflict with Evidentialism. More attempts at solving the puzzle will be discussed in Section 2.

To sum up this section, HOE generates a puzzle: both Evidentialism and Bridging (or Enkrasia) are plausible and yet they seem to come into conflict. What's worse, *the puzzle doesn't merely arise for evidentialists*: it generalizes to a wide range of theories about determinants of rationality. Suppose we say that rationality is determined by some condition C. For a wide range of candidate condition C, it's possible that both your first-order belief p and the relevant higher-order belief satisfy C, just like it's possible that they can both be supported by evidence. For instance, it's possible that both your belief p and your belief 'my belief p is produced by an unreliable process' are produced by a reliable process. So, it's not just the evidentialist who faces the puzzle. However, for simplicity, I will continue to use the evidentialist presentation of the puzzle in my following discussion.

2 Major Positions in the Higher-Order Evidence Debate

In Section 1, we have seen that there is an apparent conflict between Evidentialism and Bridging. We have seen that the conflict arises because:

(a) Given your total evidence, you should believe that your belief that p is unreliably formed (or irrational, or not supported by your evidence, etc.).
(b) Given your total evidence, you should believe that p.
(c) Given plausible level-connecting principles (such as Bridging or Enkrasia), you should not hold both beliefs.

There is an apparent conflict among the three claims. Major positions in the HOE debate can be structured around how they respond to the conflict. I critically review the three main positions in this section.

2.1 Higher-Order Defeat/Calibrationism

The first response to the puzzle is called 'Higher-Order Defeat.' It says that HOE defeats the rationality of one's first-order belief. That is, it keeps (a) and (c) but denies (b) (Christensen, 2010; Feldman, 2005; Neta, 2018; Skipper, 2019). There are several explanations of how the Higher-Order Defeat happens.

The first is an idea of 'bracketing': even though E is evidence about *p*, the HOE of your unreliability or irrationality requires you to *set aside* E in reasoning about *p*. (Christensen [2010] first proposes this idea of bracketing in discussing how to respond to HOE about one's unreliability.) Since we assume that E is all the first-order evidence you have, bracketing E means that you can no longer rationally believe that *p*.

The motivation for the bracketing view can be seen with the following analogy. Consider a knife you use as a tool for cutting things. One situation where you should not use the knife is when you know that it's broken so that it can no longer function as a good tool for cutting. But another situation where you should not use the knife is when you gain evidence saying that you've developed some brain disease so that your hands will shake uncontrollably when you use the knife. Moreover, it seems that you should not use the knife in this situation even if the evidence about your brain disease is in fact misleading. Similarly, we can consider evidence as the tool we use for the purpose of forming true beliefs. One situation where you should not use a piece of evidence in forming beliefs about *p* is when you know that its usual evidential link to *p* is broken (which is what happens when you gain undermining evidence). However, another situation where you should not use the evidence is when you gain evidence saying that you suffer from some cognitive impairment so that your chance of forming true beliefs on the basis of that evidence is very low. Moreover, it seems that you should not use the evidence in this situation even if the evidence about your cognitive impairment is in fact misleading.

The idea behind the analogy is simple: in the knife case, when gaining evidence about the brain disease, you shouldn't trust yourself to properly use the knife for the purpose of cutting; similarly, in the cognitive case, when gaining HOE about your unreliability, you shouldn't trust yourself to properly use the evidence about *p* for the purpose of forming true beliefs about *p*.[7]

The second explanation of how Higher-Order Defeat happens involves 'evidence dispossession': when you have HOE, you no longer *possess* E as available evidence about *p*. (See Gonzalez de Prado [2020] for an explicit defense of this view; Greco [2019] can also be read as a defense of this view.) The motivation for this view is that for a proposition E to be evidence you possess, you must satisfy a certain epistemic condition with regard to E – a proposition doesn't count as evidence you possess if you are utterly unaware of it. Traditionally, the epistemic conditions say that you must *be aware of* E, or

[7] For a detailed explanation of why we should do the bracketing, see Ye (2020), who argues that it's intellectually irresponsible to rely on the evidence that one thinks one is unable to make proper use of.