Externalism and the Value of Information

Nilanjan Das

UNC Chapel Hill/NYU Shanghai

Abstract

It is natural to think that it is always instrumentally rational for an agent to gather and use cost-free information for making decisions. In this essay, I argue that this thesis is in tension with an appealing conception of evidence, namely *evidence externalism*, according to which an agent's evidence can include non-trivial propositions about the external world.

Guide for FEW participants: My presentation will focus on sections 1, 2, 4, and 5.

Our evidence is our best guide to the truth. To be successful in our theoretical and practical projects, we need to believe the truth about the relevant subject-matters. Therefore, we ought to gather more evidence and use it for making decisions about our projects, *unless* gathering evidence and using it is too costly. This supports:

> VALUE OF INFORMATION. If evidence is available to an agent (for gathering and use) at a negligible cost, then it is instrumentally rational for that agent to gather that evidence and use it for making decisions.[1]

Despite its intuitive appeal, VALUE OF INFORMATION is surprisingly hard to defend. In this essay, I argue that VALUE OF INFORMATION is in tension with an appealing conception of evidence, namely EVIDENCE EXTERNALISM.

> EVIDENCE EXTERNALISM. An agent's evidence may include non-trivial propositions about the external world.[2]

To show this, I begin with I. J. Good's [1967] argument for VALUE OF INFORMATION (§1). Good assumes two controversial "access principles" about evidence.

> POSITIVE ACCESS. If an agent's evidence entails a proposition X in a world w, then her evidence in w entails that her evidence entails X.
> NEGATIVE ACCESS. If an agent's evidence doesn't entail a proposition X in a world w, then her evidence in w entails that her evidence doesn't entail X.

Given certain plausible assumptions about evidence, all evidence externalists should reject NEGATIVE ACCESS, and some of them may even reject POSITIVE ACCESS (§2). Therefore, the

[1] For decision-theoretic arguments for VALUE OF INFORMATION, see Peirce [1967], Ramsey [1990], and Good [1967].

[2] John McDowell [1995, 2011], Timothy Williamson [2000], and Goldman [2009] are three prominent defenders of EVIDENCE EXTERNALISM.

challenge for evidence externalists is to preserve VALUE OF INFORMATION without accepting both these access principles.

I show that evidence externalists cannot satisfactorily answer this challenge. First, I observe that when both NEGATIVE ACCESS and POSITIVE ACCESS fail, it isn't always instrumentally rational for an agent to gather and use cost-free information (§3). Then, following a proposal due to John Geanakoplos [1989], I consider an externalist position that attempts to save VALUE OF INFORMATION by replacing NEGATIVE ACCESS with another condition called *nestedness* (§4). But this strategy, I claim, doesn't succeed: nestedness should be rejected for the same reason which makes NEGATIVE ACCESS incompatible with EVIDENCE EXTERNALISM (§5). Finally, I respond to two objections against my argument (§6).

In the last part of the essay, I suggest that we shouldn't try to save VALUE OF INFORMATION by rejecting EVIDENCE EXTERNALISM; for there is no acceptable alternative conception of evidence that preserves VALUE OF INFORMATION (§7). We should just embrace the seemingly implausible claim that gathering and using cost-free information isn't always instrumentally rational.

1 Good's Theorem

Consider the following example.

> *Example 1.* You work in a chemical laboratory. You want to determine the chemical properties of a certain solution: you know that it is either acidic or alkaline, but you currently have neither more nor less reason to think that it is acidic rather than alkaline. You have three options: storing the solution with the acids, storing it with the alkalis, and doing nothing. If it is acidic, you want to store it with the other acids; if it is alkaline, you want to store with the alkalis. In any case, you don't want to misclassify the solution: this will make certain experiments go wrong. You have at your disposal a blue litmus paper and a red litmus paper. If the blue litmus paper turns red when brought in contact with the solution, you will learn that the solution is acidic. If the red litmus paper turns blue when brought in contact with the solution, you will learn that the solution is alkaline. Should you test the solution using these pieces of litmus paper before you decide where to store the solution?

The answer, intuitively, is, "Yes, of course!" Given that you can gather information about the chemical properties of the solution without any cost whatsoever, you ought to do it. VALUE OF INFORMATION vindicates this intuition: it says that, if evidence is available (for gathering and use) at a negligible cost, then it is instrumentally rational for an agent to gather more evidence and use it for making decisions.[3]

[3]Note why the qualification about the evidence is being negligibly costly is necessary. If one were always waiting for more information before making decisions, one would end up making very few decisions. That is suboptimal, so evidence-gathering in such cases turns out to be costly. In order to rule out such scenarios, we need to restrict our attention only to scenarios where the evidence is cost-free or involves negligible costs.

I. J. Good [1967] offered an argument for VALUE OF INFORMATION. According to a widely-accepted picture of instrumental rationality, an agent who is instrumentally rational adopts options that maximize expected value. Good proved that, under certain idealizing assumptions, if evidence is available (for gathering and use) at a negligible cost, then gathering more evidence and using it for making decisions always maximizes expected value.

To state Good's theorem more precisely, we need some formal machinery.

1.1 Frames

To represent an agent's evidence, I shall use Kripke- or Hintikka-style relational structures, called *frames*, commonly used in the semantics for logics of knowledge, belief, and evidence.

A *frame* is a structure $\mathcal{F} = \langle W, P \rangle$. In this structure, W is a finite set of "worlds" or "states." Let a *proposition* be a set of worlds in W. For example, the proposition that Boston is the capital of Massachusetts just is the set of worlds in which Boston is the capital of Massachusetts. The power set of W, $\mathcal{P}(W)$, is the set of all propositions. The function $P : W \to \mathcal{P}(W)$ (equivalent to an *accessibility relation* in Kripke- or Hintikka-style relational structures) maps each world to a body of evidence $P(w)$ the agent possesses in that world, which is also a proposition. For example, an agent's evidence in w entails that Boston is the capital of Massachusetts if and only if $P(w)$ is a subset of the proposition that Boston is the capital of Massachusetts.

In *Example 1*, suppose you are going to run the litmus test anyway. So, your evidence before and after the test in that scenario can be represented using two frames $\langle W, P \rangle$ and $\langle W, Q \rangle$, where W includes just two worlds *ac*—i.e., the world the solution is *ac*idic—and *al*—the world where it is *al*kaline.

The frame $\langle W, P \rangle$ represents your total evidence before the test. Every world is compatible with your evidence in every world at that stage. So, $P(ac) = P(al) = \{al, ac\}$. Figure 1.1 is a graph-theoretic representation of your evidence before the test (where there is a path from a node A to a node B if and only if the world represented by B is compatible with the agent's evidence in the world represented by A).

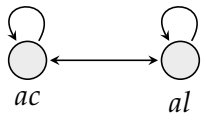

Figure 1.1: Your Evidence Before the Test

Now, consider the frame $\langle W, Q \rangle$, which represents your evidence after the test. In each world, you learn something about the solution. In *ac*, your evidence entails that the solution is acidic, so it no longer includes the world *al*. Similarly, in the world *al* where you learn that the solution is alkaline, your evidence rules out *ac*. Hence, $Q(ac) = \{ac\}$, while $Q(al) = \{al\}$. The frame $\langle w, Q \rangle$ represented in Figure 1.2.

Figure 1.2: Your Evidence Before the Test

A property of frames will be useful for our purposes.

Definition 1. A frame $\mathcal{F} = \langle W, P \rangle$ is *partitional* iff three conditions are satisfied:

- $\langle W, P \rangle$ is *reflexive*, i.e., for any world $w \in W$, w is compatible with the agent's evidence in w; formally,

  $$(\forall w \in W)(w \in P(w)).$$

- $\langle W, P \rangle$ is *transitive*, i.e., for any $w, w', w'' \in W$, if w' is compatible with the agent's evidence in w, and w'' is compatible with the agent's evidence in w', then w'' is compatible with the agent's evidence in w; formally,

  $$(\forall w, w', w'' \in W)((w' \in P(w) \,\&\, w'' \in P(w')) \Rightarrow w'' \in P(w)).$$

- $\langle W, P \rangle$ is *euclidean*, i.e., for any $w, w', w'' \in W$, if w' is compatible with the agent's evidence in w, and w'' is compatible with the agent's evidence in w, then w'' is compatible with the agent's evidence in w'; formally,

  $$(\forall w, w', w'' \in W)((w' \in P(w) \,\&\, w'' \in P(w)) \Rightarrow w'' \in P(w')).$$

The relevant property is called *partitionality*, because in a reflexive, transitive, and euclidean frame $\langle W, P \rangle$, P imposes an partition on W where, for any world w, each $P(w)$ is a cell containing all and only those worlds in which the agent's evidence is $P(w)$. We can see that the frames represented in Figures 1.1 and 1.2 satisfy partitionality: they are reflexive, transitive, euclidean.

Reflexivity, transitivity, and euclideanness correspond to the following three properties of evidence respectively:

> FACTIVITY. If an agent's evidence entails a proposition X in a world w, then X is true in w.
> POSITIVE ACCESS. If an agent's evidence entails a proposition X in a world w, then her evidence in w entails that her evidence entails X.
> NEGATIVE ACCESS. If an agent's evidence doesn't entail a proposition X in a world w, then her evidence in w entails that her evidence doesn't entail X.[4]

[4]It is easy to check this. An agent's evidence entails a proposition X if and only if the proposition that represents her evidence is a subset of X.

Suppose FACTIVITY is false: there is a proposition X entailed by an agent's evidence in a world w, but X is false at w. This would mean that w isn't compatible with the agent's evidence in w, which violates reflexivity. The converse also holds.

Suppose POSITIVE ACCESS is false; so, there is a world w where the agent's evidence entails a claim X, but doesn't entail that it entails X. In that case, there is a world w' that is compatible with the agent's evidence in w, such that the agent's evidence in w' doesn't entail X. This can only be the case if there is a world w'' compatible with the agent's evidence in w', but not compatible with her evidence in w. But this violates transitivity. The converse also holds.

Suppose NEGATIVE ACCESS is false; so, there is a world w where the agent's evidence doesn't entail a claim X,

Since the frames $\langle W, P\rangle$ and $\langle W, Q$, depicted in Figures 1.1 and 1.2, are partitional, your evidence before and after the test in *Example 1* satisfies FACTIVITY, POSITIVE ACCESS, and NEGATIVE ACCESS.

Here is another property of frames that will be useful to us.

Definition 2. For any two frames $\langle W, P\rangle$ and $\langle W, Q\rangle$, P is *coarser* than Q if and only if, for any $w \in W$, $Q(w) \subseteq P(w)$.

In *Example 1*, the frame $\langle W, P\rangle$ represents your evidence before the test, and $\langle W, Q\rangle$ represents your evidence after the test. In this scenario, P is coarser than Q: for any w, $Q(w)$ is a subset of $P(w)$. Intuitively, this means there is no world where you have strictly less evidence after the test than you have before the test. In other words, you don't lose any information at any world in the course of running the test.

1.2 Decision Problems

Next, I am going to represent decision-making scenarios that an agent may face as *decision problems*.

A *decision problem* is a structure $D = \langle W, P, A, \pi, \mu, f\rangle$. Here, $\langle W, P\rangle$ is the frame that the decision problem D is *based* on. A is a countable set of acts. The function $\pi : \mathcal{P}(W) \to \mathbb{R}$ is a regular probability measure defined over propositions, which reflects the agent's initial credences about various propositions prior to receiving any evidence; it is what we shall call an *ur-prior*.[5] The function $\mu : A \times W \to \mathbb{R}$ is a utility function, such that $\mu(a, w)$ reflects the value of performing an action a in a world w. Finally, $f : W \to A$ is the function that picks out the act the agent prefers to perform in each world relative to her evidence.

A few more remarks about the *ur-prior* π. For any $X \subseteq W$, $\pi(X)$ is the agent's initial credence in X prior to receiving any evidence whatsoever. For simplicity, for any world $w \in W$, I will write $\pi(\{w\})$ as $\pi(w)$. For any $X, Y \subseteq W$, $\pi(X|Y) = \dfrac{\pi(X \cap Y)}{\pi(Y)}$ is the conditional credence that the agent initially assigns to X given Y. Since π is regular, $\pi(X|Y)$ will always be defined.

The following property of a decision problem will be helpful to remember.

Definition 3. A decision problem $D = \langle W, P, A, \pi, \mu, f\rangle$ satisfies *Bayesian rationality* if and only if, at any world $w \in W$, two conditions are satisfied.

- CONDITIONALIZATION. Relative to her total body of evidence $P(w)$, the agent's credence function is the conditional credence function $\pi(.|P(w))$.[6]

but doesn't entail that it doesn't entail X. If there is a world w where the agent's evidence doesn't entail a claim X, then there must be a world w' such that w' is compatible with the agent's evidence in w, but X is false at w'. If the agent's evidence doesn't entail that it doesn't entail X, then there is a world w'' that is compatible with the agent's evidence in w, such that the agent's evidence in w'' entails X. But this means that w' isn't compatible with the agent's evidence in w''. This violates euclideanness. The converse also holds.

[5] A *regular* probability function defined over a set of propositions assigns non-zero probability to every proposition in that set.

[6] According to Bayesian orthodoxy, conditionalization is often understood as a diachronic constraint on rational-

- EXPECTED UTILITY MAXIMIZATION. The agent prefers an act that maximizes expected utility relative to her credence function and her utility function. More formally,

$$(\forall w \in W)(f(w) \in \arg\max_{a \in A} \sum_{w' \in W} \pi(w'|P(w))\mu(a, w')).^7$$

Thus, Bayesian rationality encodes two constraints. The first is a constraint on what credences are rational for an agent is adopt at every stage of inquiry: these are just the initial conditional credences that she assigns to various propositions given her total body of evidence in that stage of inquiry. The second is a constraint on what preferences are rational for an agent to adopt at every stage of her inquiry: she prefers an act that maximizes expected utility relative to her current credence function and utility function. The first is a constraint of epistemic rationality; the second is a constraint of instrumental or practical rationality.

To get a sense of how this works, consider *Example 1* once more. Suppose the decision problems based on $\langle W, P\rangle$ and $\langle W, Q\rangle$, represented in Figures 1.1 and 1.2, are $\langle W, P, A, \pi, \mu, f\rangle$ and $\langle W, Q, A, \pi, \mu, g\rangle$ respectively.

First, let us ask what your credences before and after the test are.

> *Current Credences*. Before the test, in both *ac* and *al*, you have neither more nor less reason for thinking that the solution is acidic than for thinking that it is alkaline. So, you assign credence 0.5 to both *al* and *ac*.
>
> *Future Credences*. After the test, in *ac*, you learn in the future that the solution is acidic, while in *al* you learn in the future that the solution is alkaline. Assuming that you satisfy Bayesian rationality, in *ac*, you assign credence 1 to *ac* and credence 0 to *al*, and in *al*, you assign credence 1 to *al* and credence 0 to *ac*.

Table 1 reflects the credences you assign to *ac* and *al* relative to your evidence before and after the test in *ac* and *al*. Here, the first two rows reflect your credences in *ac* and *al* relative to your evidence before the test, while the second two rows reflect your credences after the test.

	Worlds	
Credence Functions	*ac*	*al*
$\pi(.\|P(ac))$	0.5	0.5
$\pi(.\|P(al))$	0.5	0.5
$\pi(.\|Q(ac))$	1	0
$\pi(.\|Q(al))$	0	1

Table 1: Your Credences in *Example 1*

ity, according to which an agent, on learning a piece of E, adopts as her posterior credence towards any proposition X the prior conditional credence that she assigned to X given E. However, what we are calling CONDITIONALIZATION is a synchronic constraint. However, if we build into the picture the fact that the agent doesn't lose any evidence over time, then CONDITIONALIZATION becomes equivalent to the diachronic constraint that is part of the Bayesian picture.

[7] Here, we are using standard decision theory, which implicitly assumes that the states of the world don't depend (epistemically or causally) on which acts we perform. If we reject that assumption, we have to either accept evidential decision theory or causal decision theory. As Skyrms [1990] notes, evidential decision theory doesn't preserve VALUE OF INFORMATION, but causal decision theory does.

Second, let's see what the payoffs for the various acts are. You have three options available to you: storing the solution with acids (*Acid*), storing it with the alkalis (*Alkali*), or doing nothing (*Nothing*). When the solution is acidic, and you store it with the acids, then you get to use it for the right purposes; similarly, when the solution is alkaline, and you store it with the alkalis, then you get to use it for the right purposes. In each case, let's say, the payoff is 10. But if you store the solution at the wrong place, your experiments go wrong; so, the payoff is negative, say, -100. Doing nothing has no negative or positive payoff; so, the payoff is 0.

Accordingly, Table 2 specifies the values of the utility function μ.

	Worlds	
Acts	*ac*	*al*
Acid	10	-100
Alkali	-100	10
Nothing	0	0

Table 2: Payoffs for *Example 1*

Next, we focus on your preferred acts before and after the test.

Current Preferences. By lights of your credences before the test in both *ac* and *al*, the expected value of both storing the solution with the acids and storing it with alkalis is $0.5 \times 10 + 0.5 \times (-100) = -45$. By contrast, the expected value of doing nothing is $0.5 \times 0 + 0.5 \times 0 = 0$. So, if you satisfy Bayesian rationality, your preferred act before the test in both *ac* and *al* is *Nothing*.

Future Preferences. In *ac*, by lights of your credences after the test, the expected value of storing the solution with the acids is $1 \times 10 + 0 \times (-100) = 10$; by contrast, the expected value of doing nothing is 0, and the expected value of storing the solution with the alkalis is $1 \times (-100) + 0 \times 10 = -100$. By similar reasoning, in *al*, by lights of your credences after the test, the option of storing the solution with the alkalis maximizes expected value. If you satisfy Bayesian rationality, after the test, your preferred act in *ac* is storing the solution with acids, i.e., *Acid*, while your preferred act in *al* is storing the solution with alkalis, i.e., *Alkali*.

Accordingly, Table 3 sets out the values of the functions, f and g, which reflect your preferred acts before and after the test respectively.

	Worlds	
Preference Functions	*ac*	*al*
f	*Nothing*	*Nothing*
g	*Acid*	*Alkali*

Table 3: Your Preferred Acts in *Example 1*

We are now in a position to state Good's result.

1.3 Good's Theorem

Good [1967] proves the following theorem.

> **Good's Theorem (Good 1967).** Suppose there are two decision problems $D_1 = \langle W, P, A, \pi, \mu, f\rangle$ and $D_2 = \langle W, Q, A, \pi, \mu, g\rangle$, which are based on partitional frames and satisfy Bayesian rationality, such that P is coarser than Q. Then, for any world $w \in W$,
>
> $$\sum_{w' \in W} \pi(w'|P(w))\mu(f(w'), w') \leq \sum_{w' \in W} \pi(w'|P(w))\mu(g(w'), w'),$$
>
> with strict inequality unless, for all $w' \in P(w)$, $f(w') = g(w')$.

This inequality is sometimes called *Good's inequality*.

If we take $\langle W, P\rangle$ and $\langle W, Q\rangle$ to represent an agent's present and future bodies of evidence, then the theorem says the following. Suppose that, in every world, an agent

> (i) satisfies FACTIVITY, POSITIVE ACCESS, and NEGATIVE ACCESS at present as well as in the future,
> (ii) has the same unique and precise *ur-prior* and the same utility function at present as well as in the future,
> (iii) doesn't lose any evidence in the course of gathering new evidence between the present and the future, and
> (iv) updates her credences by conditionalizing her ur-prior on her total body of evidence and prefers acts that maximize expected utility relative to her current credence function and utility function at present as well as in the future.

Then, the expected value of performing the act that the agent prefers relative to her present evidence is *less than or equal to* the expected value of performing the act that the agent prefers relative to her future evidence, and *strictly less* when there is at least one world where, relative to her future evidence, she prefers to perform a different act from the one that she prefers to perform relative to her present evidence.

Now, suppose an agent is *permitted* by instrumental rationality to perform an act if and only if it is one of the acts that maximize expected value by lights of her current credences and utilities. Also, suppose that she is *required* by instrumental rationality to perform an act if and only if it is the only act that maximize expected value by lights of her current credences and utilities. If this conception of instrumental rationality is correct, then Good's theorem entails the following: when some evidence is available for gathering and use at a negligible cost, and conditions (i)-(iv) are satisfied, then an agent is always *permitted* by instrumental rationality to gather more evidence and use it for making decisions, and *required* to do so when, relative to her future evidence, she possibly prefers an act different from the one she prefers relative to her current evidence. This is how Good's theorem lends support to VALUE OF INFORMATION.

Let us see how this applies to *Example 1*. In that scenario, we have two decision problems $\langle W, P, A, \pi, \mu, f\rangle$ and $\langle W, Q, A, \pi, \mu, g\rangle$, which are based on partitional frames and satisfy

Bayesian rationality, such that P is coarser than Q. Plugging in the values from Tables 1-3, we get, for any $w \in W$,

$$\begin{aligned}
\sum_{w' \in W} \pi(w'|P(w))\mu(f(w'),w') &= \pi(ac|P(w))\mu(f(ac),ac) + \pi(al|P(w))\mu(f(al),al) \\
&= 0.5 \times \mu(Nothing, ac) + 0.5 \times \mu(Nothing, ac) \\
&= 0.5 \times 0 + 0.5 \times 0 \\
&= 0. \\
\sum_{w' \in W} \pi(w'|P(w))\mu(g(w'),w') &= \pi(ac|P(w))S(g(ac),ac) + \mu(al|P(w))\mu(g(al),al) \\
&= 0.5 \times \mu(Acid, ac) + 0.5 \times \mu(Alkali, al) \\
&= 0.5 \times 10 + 0.5 \times 10 \\
&= 10.
\end{aligned}$$

Therefore, Good's inequality holds in *Example 1*. So, you are required by instrumental rationality in this scenario to run the test before you decide where to store the solution.

1.4 Relaxing the Assumptions

Good showed that if certain idealizing assumptions hold, then the expected value of performing an act that one prefers relative to one's future evidence is greater than or equal to the expected value of performing an act that one prefers relative to one's current evidence. Now, some writers have tried to see if Good's inequality holds when we relax these idealizing assumptions.

Consider assumption (ii), according to which the agent must have the same unique and precise *ur-prior* function and the same utility function at present as well as in the future. Some writers (including Good himself) have pointed out that when probabilities are imprecise, Good's inequality doesn't hold.[8]

Other writers have explored the consequences of rejecting assumption (iv), which says that the agent must satisfy both CONDITIONALIZATION and EXPECTED VALUE MAXIMIZATION. Some writers have tried to generalize Good's result to cases where the agent doesn't update her credences according to CONDITIONALIZATION.[9] Others have shown that when an agent exhibits a form of risk-aversion that is incompatible with EXPECTED VALUE MAXIMIZATION, Good's inequality may no longer hold.[10]

In the rest of this essay, I shall examine whether Good's inequality can be preserved without assumption (i), namely, the assumption that the agent satisfies FACTIVITY, POSITIVE ACCESS, and NEGATIVE ACCESS at present as well as in the future.

[8] See Good [1974] and Kadane, Schervish and Seidenfeld [2008].
[9] See, for example, Skyrms [1990] and Huttegger [2014].
[10] See Buchak [2010].

2 Evidence Externalism and the Access Principles

Amongst FACTIVITY, POSITIVE ACCESS, and NEGATIVE ACCESS, FACTIVITY seems to be the least controversial. According to FACTIVITY, an agent's evidence only entails truths. There are plenty of arguments in favour of FACTIVITY; here is one. If FACTIVITY were false, then an agent's evidence could entail a falsehood, and therefore an agent could have conclusive evidence for a falsehood. But an agent cannot have conclusive evidence for a falsehood; for the evidential support for any false claim could indeed be defeated by some further piece of evidence that shows that the claim is false, and the evidential support for a claim which has been conclusively established cannot be so defeated. So, FACTIVITY must be true.[11]

Whatever you make of this argument, it is worth pointing out that if we reject FACTIVITY, VALUE OF INFORMATION would be extremely difficult to establish; for, if our future evidence entails falsehoods, then gathering new evidence and using it for decision-making may have disastrous consequences. Hence, it may indeed be instrumentally rational for us to avoid gathering new evidence. This connection between VALUE OF INFORMATION and FACTIVITY is so obvious that there doesn't seem to be anything theoretically interesting about calling VALUE OF INFORMATION into question by rejecting FACTIVITY. Therefore, let us grant that FACTIVITY is true.

What about POSITIVE ACCESS and NEGATIVE ACCESS? There is a certain conception of evidence on which both these access principles might seem quite natural. According to a Cartesian picture of evidence, an agent's evidence consists only of facts concerning her current phenomenal states, i.e., facts about *what it's like* for her at that time. It is commonly thought that an agent cannot be misled about such states and their absence: if such states obtain, the agent learns by introspection that they do, and if they don't obtain, the agent learns by introspection that they don't. On this picture, therefore, when the agent's evidence includes (or doesn't include) a certain proposition, her evidence entails that her evidence includes (or doesn't include) that proposition. Therefore, both POSITIVE and NEGATIVE ACCESS are true.

However, the Cartesian picture of evidence pushes us towards scepticism about the external world. For the Cartesian, we only ever have conclusive evidence for propositions about our mental lives. So, when we undergo our first perceptual experience at the beginning of our epistemic careers, our evidence comes to include only the proposition that such an experience has occurred. In order for that evidence to support any claim about the external world, we would have to have independent evidence for thinking that the relevant experience is veridical. Now, *ex hypothesi*, we don't have any other empirical evidence in that stage. Since we cannot have non-empirical evidence for taking our experiences to be veridical, we cannot form any justified belief about the external world under those circumstances. But this means that we cannot ever form justified beliefs about the external world: for any subsequent experience that we may undergo, we won't have any independent empirical or non-empirical evidence for believing

[11] For other arguments in favour of FACTIVITY, see Williamson [2000], Littlejohn [2012], Byrne [2013]. For dissent from FACTIVITY, see Joyce [2004], Goldman [2009], and Leite [2013].

that the experience is veridical. This will lead to scepticism about the external world.[12]

Unless we want to embrace full-fledged scepticism about the external world, we ought to adopt an account of evidence, on which an agent can acquire evidence not only about her own phenomenal states, but also about the external world. This supports:

> EVIDENCE EXTERNALISM. An agent's evidence may include non-trivial propositions about the external world.[13]

However, if both FACTIVITY and EVIDENCE EXTERNALISM are true, then NEGATIVE ACCESS cannot be saved.

Suppose I am looking at a white wall that is lit up with red light, but I have no reason to think that this is the case. Since I am undergoing an experience as of there being a red wall before me, I have strong misleading evidence for thinking that the wall before me is red. If EVIDENCE EXTERNALISM is correct, then, plausibly, I can gain conclusive evidence about the external world from my veridical and reliable perceptual experiences. So, when I have strong misleading evidence for thinking that the wall is red, I may have strong evidence for thinking that I have conclusive perceptual evidence that the wall is red; for I have no reason to suspect that my perceptual experience is unreliable or non-veridical. Thus, my evidence won't entail that my evidence doesn't entail that the wall is red. However, by FACTIVITY, my evidence won't entail that the wall is red, because that claim is false. Therefore, NEGATIVE ACCESS will fail.[14]

More generally, the idea is this. Even when a claim P about the external world is false, an agent may have strong misleading evidence for thinking that P is true. If EVIDENCE EXTERNALISM is correct, the agent may in such a scenario have strong misleading evidence for thinking that P is part of her evidence (provided that she also thinks that other conditions for P to be part of her evidence are satisfied). However, by FACTIVITY, the agent's evidence cannot entail P. Therefore, NEGATIVE ACCESS fails.

It is worth noting that some evidence externalists also take POSITIVE ACCESS to be false. Consider, for example, what Timothy Williamson [2000] calls the E=K thesis, i.e., the thesis that all and only known claims are part of an agent's evidence. If this view is correct, then this entails that if what an agent knows entails X, then what she knows entails that what she knows entails X. This is questionable for the same reasons that cast doubt on the KK principle, i.e., the principle that if an agent knows a claim, she is in a position to know that she knows it. Since knowledge requires reliability, we might think that an agent can reliably believe a claim, without being able to reliably determine that she reliably believes it; if so, she can

[12] I will consider some responses to this argument in §7.

[13] Typical examples of EVIDENCE EXTERNALISM include Williamson's [2000] E=K thesis, McDowell's [2011] view that when one undergoes a veridical perception, one's evidence includes the proposition that one sees that such-and-such is the case, and Goldman's [2009] view that one's evidence includes the deliverances of reliable non-inferential cognitive processes.

[14] For similar complaints about the negative introspection principle about knowledge, see Hintikka (1962, p. 106) Williamson (2000, pp. 23-27), and Stalnaker (2009, p. 400). Some externalists like Goldman [2009] don't accept FACTIVITY. However, Goldman might still reject NEGATIVE ACCESS. On his view, it is possible for an agent to have misleading evidence about the reliability of a cognitive mechanism; so, an agent may reasonably take her evidence to entail a certain proposition when it in fact doesn't entail it.

know without being able to know that she does.[15] Other writers, however, have resisted this argument.[16] Therefore, EVIDENCE EXTERNALISM need not be straightforwardly incompatible with POSITIVE ACCESS.

Let us take stock. If FACTIVITY is true, EVIDENCE EXTERNALISM is incompatible with NEGATIVE ACCESS. In what follows, I argue that this very tension also creates a tension between EVIDENCE EXTERNALISM and VALUE OF INFORMATION.

3 Good's Inequality Without Positive and Negative Access

I will begin with the observation that Good's inequality doesn't always hold when both POSITIVE and NEGATIVE ACCESS fail.

Consider the following example.

> *Example 2.* You are about to enter a room, and look at a wall. Your current evidence entails that the wall is going to be one of three shades of red: crimson, rusty red, and cardinal red. On the basis of your current evidence, you are 0.1 confident that the wall is going to be crimson, 0.8 confident that it is going to be rusty red, and 0.1 confident that it is going to be cardinal red. You are also certain that you can discriminate crimson from cardinal red, and vice-versa, but you can't discriminate rusty red from either crimson or cardinal red. You know you will be offered a gamble where you stand to gain \$100 if the wall is rusty red, and lose \$450 if it's not. Should you make a decision about this before you enter the room?

In this scenario, your future evidence doesn't satisfy POSITIVE or NEGATIVE ACCESS.[17] When the colour of the wall is crimson, the agent's evidence entails that the claim that the wall isn't cardinal red, but her evidence doesn't entail that her evidence entails this claim. So, POSITIVE ACCESS fails. When the colour of the wall is rusty red, the agent's evidence doesn't entail that the claim that the wall isn't cardinal red, but her evidence doesn't entail that her evidence doesn't entail this claim. So, NEGATIVE ACCESS fails. First, consider the frame $\langle W, P\rangle$ that represents your current evidence. Let $W = \{cr, ru, ca\}$, where *cr* is the world in which the wall is *cr*imson, *ru* is the world in which the wall is *ru*sty red, and *ca* is the world in which the wall is *ca*rdinal red. Since all these three worlds are compatible with your current evidence in every world, $P(cr) = P(ru) = P(ca) = \{cr, ru, ca\}$ (Figure 3.1).

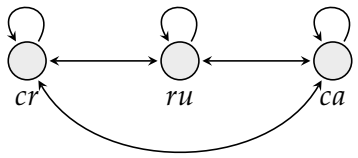

Figure 3.1: Your Current Evidence in *Example 2*

[15] For such complaints against the KK principle, see Alston (1980, pp. 140141), Williams (1991, p. 96), Antony ([2004], p. 12) and Dretske (2004,section 2), and Williamson [2000].

[16] See, for example, Stalnaker (2006, 2009, 2015), Greco [2014], and Das and Salow [forthcoming].

[17] This example is similar to a case of "improbable knowing" discussed by Timothy Williamson [2011].

$\langle W, P\rangle$ is partitional; for every world is compatible with your evidence in every world. Therefore, you currently satisfy FACTIVITY, POSITIVE ACCESS, and NEGATIVE ACCESS.

Compare this to your future evidence, which is represented by $\langle W, Q\rangle$. In *cr*, your future evidence after entering the room eliminates *ca*, but not *cr* and *ru*. In *ca*, your future evidence after entering the room eliminates *cr*, but not *ca* and *ru*. But in *ru*, your evidence eliminates none of the worlds. So, $Q(cr) = \{cr, ru\}$, $Q(ca) = \{ru, ca\}$, and $Q(ru) = \{cr, ru, ca\}$ (Figure 3.2).

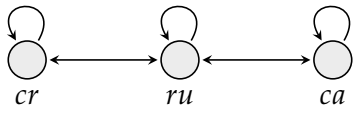

Figure 3.2: Your Future Evidence in *Example 2*

At *cr*, the world *ca* isn't compatible with your evidence, but *ru* is. However, at *ru*, the world *ca* is compatible with your evidence. As a result of this failure of transitivity, POSITIVE ACCESS fails. Moreover, at *ru*, both *cr* and *ca* are compatible with your evidence, but *cr* isn't compatible with your evidence in *ca*. Hence, euclideanness fails. That is why NEGATIVE ACCESS also fails.

In this example, since you don't satisfy POSITIVE ACCESS and NEGATIVE ACCESS in the future, Good's inequality fails. Let $\langle W, P, A, \pi, \mu, f\rangle$ and $\langle W, Q, A, \pi, \mu, g\rangle$ be the decision-problems based on $\langle W, P\rangle$ and $\langle W, Q\rangle$ respectively.

First, let us ask what your credences relative to your current and future evidence are.

> *Current Credences*. By the setup of the story, you assign 0.8 credence to *ru* before you see the wall, and 0.1 each to *cr* and *ca*.
>
> *Future Credences*. After you see the wall, your credences in *ru* remain unchanged; for, in that world, you gain no evidence. In *cr*, your credence in *ru* increases to $\frac{0.8}{0.1+0.8} = \frac{8}{9}$; as a result, you assign credence $\frac{1}{9}$ to *cr* and credence 0 to *ca*. Analogously, in *ca*, your credence in *ru* increases to $\frac{0.8}{0.1+0.8} = \frac{8}{9}$; as a result, you assign credence $\frac{1}{9}$ to *ca* and credence 0 to *cr*.

So, your current and future credences stand as in Table 4.

	Worlds		
Credence Functions	*cr*	*ru*	*ca*
$\pi(.\|P(cr))$	0.1	0.8	0.1
$\pi(.\|P(ru))$	0.1	0.8	0.1
$\pi(.\|P(ca))$	0.1	0.8	0.1
$\pi(.\|Q(cr))$	1/9	8/9	0
$\pi(.\|Q(ru))$	0.1	0.8	0.1
$\pi(.\|Q(ca))$	0	8/9	1/9

Table 4: Your Credences in *Example 2*

Now, let's fix the payoffs. In this example, you have two options: the option of accepting

the gamble—call it *Accept*—and the option of rejecting the gamble—call it *Reject*. Assuming that you value money linearly, the values of the utility function μ stand as in Table 5.

	Worlds		
Acts	*cr*	*ru*	*ca*
Accept	-450	100	-450
Reject	0	0	0

Table 5: Payoffs for *Example 2*

Now, we are in a position to see what your current and future preferences might be.

Current Preferences. By your current lights, the expected value of accepting the gamble is less than that of rejecting the gamble: the expected value of accepting the gamble is $0.8 \times 100 + 0.2 \times (-450) = -10$, while the expected value of rejecting the gamble is 0. Assuming that you satisfy Bayesian rationality, your current preferred act in every world is to reject the gamble.

Future Preferences. In *ru*, since you gain no new evidence, you prefer to reject the gamble in the future. However, in *cr* and *ca*, the expected value of accepting the gamble is $\frac{8}{9} \times 100 + \frac{1}{9} \times (-450) = \frac{350}{9}$, which is greater than the expected value of rejecting the gamble, i.e., 0. Therefore, in *cr* and *ca*, you prefer to accept the gamble in the future.

Assuming that f and g are the functions that reflect your current and future preferences, your preferred acts stand as in Table 6.

	Worlds		
Preference Functions	*cr*	*ru*	*ca*
f	*Reject*	*Reject*	*Reject*
g	*Accept*	*Reject*	*Accept*

Table 6: Your Preferred Acts in *Example 2*

How does all this bear on Good's inequality? We can see that, for any $w \in W$,

$$\begin{aligned}\sum_{w' \in W} \pi(w'|P(w))\mu(f(w'), w') &= \pi(cr|P(w))\mu(f(cr), cr) + \pi(ru|P(w))\mu(f(ru), ru) \\ &\quad + \pi(ca|P(w))\mu(f(ca), ca) \\ &= 0.1 \times \mu(Reject, cr) + 0.8 \times \mu(Reject, ru) + 0.1 \times \mu(Reject, ca) \\ &= 0.1 \times 0 + 0.8 \times 0 + 0.1 \times 0 \\ &= 0.\end{aligned}$$

$$\begin{aligned}\sum_{w' \in W} \pi(w'|P(w))\mu(g(w'),w') &= \pi(cr|P(w))\mu(g(cr),cr) + \pi(ru|P(w))\mu(g(ru),ru) \\ &\quad + \pi(ca|P(w))\mu(g(ca),ca) \\ &= 0.1 \times \mu(Accept,cr) + 0.8 \times \mu(Reject,ru) + 0.1 \times \mu(Accept,ca) \\ &= 0.1 \times (-450) + 0.8 \times 0 + 0.1 \times (-450) \\ &= -90.\end{aligned}$$

Since, by lights of your current credences, the expected value of performing the act that you prefer relative to your current evidence is greater than the expected value of performing the act you prefer relative to your future evidence, Good's inequality fails in this case.

This shows that in the absence of POSITIVE and NEGATIVE ACCESS, VALUE OF INFORMATION needn't be true.

4 Good's Inequality without Negative Access

In *Example 2*, POSITIVE ACCESS fails: when the colour of the wall is crimson, the agent's evidence entails that the claim that the wall isn't cardinal red, but her evidence doesn't entail that her evidence entails this claim.

One might think that this has unpalatable consequences. Since the agent's evidence doesn't entail that her evidence entails that the wall isn't cardinal red, it is rational for the agent to be uncertain about whether she can in fact rule out the possibility that the wall is cardinal red. And, presumably, she can know this. So, she can assert, "Well, the wall isn't cardinal red, but I'm not sure whether I can rule out the possibility that it is!" This seems strange.[18] Such failures of POSITIVE ACCESS therefore are implausible. Hence, one might not take *Example 2* very seriously.

This, in turn, might give the evidence externalist some hope of saving VALUE OF INFORMATION. She might think that even though Good's inequality fails when both access principles are false, we can still validate Good's inequality by holding on to POSITIVE ACCESS and by replacing NEGATIVE ACCESS with a weaker condition.[19] In the rest of this section, I am going to flesh out this strategy in further detail.

4.1 Nestedness

Geanakoplos [1989] describes a property of frames which, together with FACTIVITY and POSITIVE ACCESS, can make Good's inequality come out true. The property is called *nestedness*.

[18] I am not claiming that this straightforwardly establishes POSITIVE ACCESS; in fact, some writers, such as Marusic [2013] and Benton [2013], have questioned similar arguments about the KK principle. However, I think such arguments put some intuitive pressure on us to accept POSITIVE ACCESS.

[19] The project of finding theoretically interesting classes of frames for representing knowledge that don't validate NEGATIVE ACCESS isn't new. See, for example, Bacharach [1985], Geanakoplos (1988, 1994), Samet [1990], and Shin [1993].

Definition 4. A frame $\langle W, P\rangle$ is *nested* if and only if for any two worlds w, w' in W, if the agent's total evidence in w, $P(w)$, isn't disjoint from her t evidence in w', $P(w')$, then either $P(w)$ entails $P(w')$, or $P(w')$ entails $P(w)$. More formally,

$(\forall w \in W)(\forall w' \in W)(P(w) \cap P(w') \neq \emptyset \Rightarrow (P(w) \subseteq P(w') \vee P(w') \subseteq P(w)))$.

Geanakoplos offers a natural interpretation of nestedness as a property of memory. Suppose I am about to appear for a chemistry exam for which I have to memorize the periodic table. Here, the only propositions of interest are of the form, "The ith element on the period table is x." Now, the learning technique that I use allows to me to remember a particular element on the period table only if I can remember all the previous ones: for example, I can't remember what the twentieth element on the table is unless I also recall what the first, the second,..., the eighteenth, and the nineteenth elements are.

Now, suppose my memory is bad, so I can only at most remember the first three elements on the periodic table. Therefore, the possible bodies of evidence that I could end up with are F, FS, and FST, where F is the proposition that hydrogen is the first element on the periodic table, FS is the proposition that hydrogen and helium are the first and second elements on the table, and FST is the proposition that hydrogen, helium, and lithium are the first, second, and third elements on the periodic table. These bodies of evidence could be represented as follows:

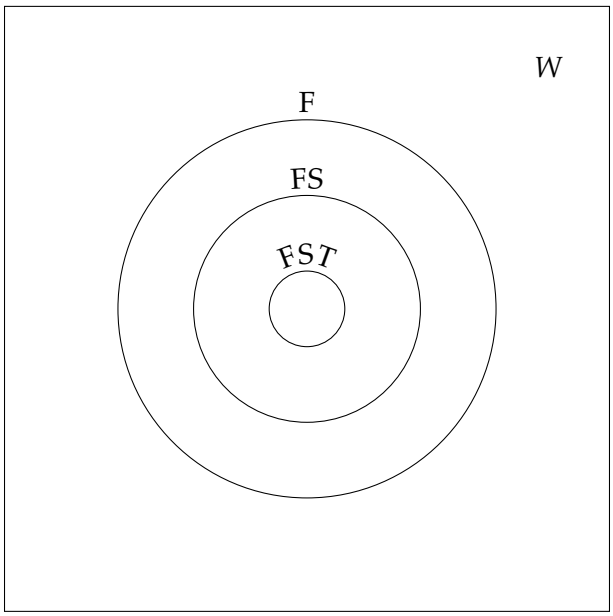

Figure 4.1: My Possible Bodies of Evidence Before the Chemistry Exam

The diagram shows why the frame representing my evidence is *nested*: my possible bodies of evidence demarcate regions that form concentric circles around each other. More formally, nestedness reflects the following property of memory: for any agent, there is a list of propositions $\{P_1, P_2, ..., P_k\}$, such that if the agent remembers P_i from that list, then, for every P_j with $j \leq i$, she remembers P_j.[20] This means that if there are two different possible scenarios where the agent has distinct bodies of evidence $E_1 = P_1 \cap P_2 \cap ... \cap P_m$ and $E_2 = P_1 \cap P_2 \cap ... \cap P_n$,

[20]For a different interpretation, see Shin [1989].

either $m \leq n$ or $n \leq m$, i.e., either $E_2 \subseteq E_1$ or $E_1 \subseteq E_2$.

It is easy to check that all partitional frames are nested. In such frames, for any two worlds w, w' in W, if the agent's body of evidence in w, $P(w)$, isn't disjoint from her body of evidence in w', $P(w')$, then $P(w) = P(w')$. However, not all nested frames are partitional. For example, consider the frame represented in Figure 4.2, where $W = \{w_1, w_2\}$, such that $P(w_1) = \{w_1\}$, but $P(w_2) = \{w_1, w_2\}$.

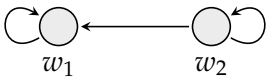

Figure 4.2: A Non-Partitional Nested Frame

This frame is nested because $P(w_1) \subseteq P(w_2)$. Even though it is reflexive and transitive, it is not euclidean: though both w_1 and w_2 are compatible with the agent's evidence in w_2, w_2 isn't compatible with the agent's evidence in w_1. Hence, the combination of reflexivity, transitivity, and nestedness is a weaker property of frames than partitionality.

Since nestedness doesn't require euclideanness, it seems well-suited to represent scenarios where NEGATIVE ACCESS fails. In the remainder of this section, we will see how nestedness allows us to preserve Good's inequality in the absence of NEGATIVE ACCESS.

4.2 Nestedness and Good's Inequality

John Geanakoplos [1989] proves that even if an agent's future evidence doesn't satisfy NEGATIVE ACCESS, Good's inequality will hold if the frame representing the agent's future evidence is nested.

> **Geanokoplos' First Theorem (Theorem 1, Geanakoplos 1989).** Suppose $D_1 = \langle W, P, A, \pi, \mu, f \rangle$ and $D_2 = \langle W, Q, A, \pi, \mu, g \rangle$ are two decision problems which satisfy Bayesian rationality, such that $\langle W, P \rangle$ is partitional and P is coarser than Q. If the frame $\langle W, Q \rangle$ satisfies reflexivity, transitivity, and nestedness, then Good's inequality holds for D_1 and D_2.[21]

Here is an example which illustrates this theorem.

> *Example 3.* You are about to enter a room and see a wall. You don't know for sure what the colour of the wall is, but it is 0.99 likely by lights of your current evidence that the wall is red. If the wall is red, your evidence after entering the room will entail that it is red. However, there is a small probability of 0.01 that it is white, but lit up with red right. In that case, your evidence will remain the same as before. You also know that immediately afterwards, you will be offered a gamble where you stand to gain \$100 if the wall is red, and lose \$10000 if the wall isn't red. Should you make a decision about the gamble before entering the room?

[21] It is worth noting that Geanokoplos claims a converse of this theorem: namely, if the the frame $\langle W, Q \rangle$ doesn't satisfy reflexivity, transitivity, or nestedness, then it is possible to construct decision problems based on that frame for which Good's inequality is strictly reversed.

In this scenario, NEGATIVE ACCESS fails. In the scenario where the wall is white but lit up with red light, the agent has misleading evidence for thinking that the wall is red, and therefore, for thinking that her evidence entails that the wall is red.

For simplicity, let $W = \{r, w\}$ where r is the world where the wall is red, while w is the world where it is white. The scenario here can be formalized using two frames $\langle W, P\rangle$ and $\langle W, Q\rangle$.

Let the frame $\langle W, P\rangle$ represent your current evidence. In each world, your current evidence is $P(r) = P(w) = \{r, w\}$ (Figure 4.3).

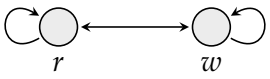

Figure 4.3: Your Current Evidence in *Example 3*

$\langle W, P\rangle$ is partitional: since every world is compatible with the agent's evidence in every world, the frame is reflexive, transitive, and euclidean.

Compare this to the frame $\langle W, Q\rangle$ which represents your future evidence. After entering the room, if you see a red wall, your future evidence entails the claim that the wall is red. But, in the world w where you see a white wall lit up with red light, your evidence doesn't rule out the world r where the wall is red. So, $Q(r) = \{r\}$ and $Q(w) = \{r, w\}$ (Figure 4.4).

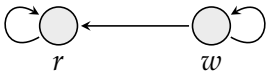

Figure 4.4: Your Future Evidence in *Example 3*

$\langle W, Q\rangle$ isn't partitional, because it is not euclidean: both r and w are compatible with the agent's evidence in w, but w isn't compatible with the agent's evidence in r. That is why NEGATIVE ACCESS fails. But $\langle W, Q\rangle$ is nested; for $Q(r) \subseteq Q(w)$.

We can now show that Good's inequality holds in this scenario. Let $\langle W, P, A, \pi, \mu, f\rangle$ and $\langle W, Q, A, \pi, \mu, g\rangle$ be the decision-problems based on $\langle W, P\rangle$ and $\langle W, Q\rangle$ respectively.

Start by asking what your current and future credences are.

> *Current Credences*. At both r and w, your current evidence is the same. In each world, you assign credence 0.99 to r, and credence 0.01 to w.
>
> *Future Credences*. In r, your future evidence is $\{r\}$, i.e., the claim that the wall is red; so, if you satisfy Bayesian rationality, you assign credence 1 to r and credence 0 to w. By contrast, in w, your future evidence remains the same before, so your credences remain the same as before.

Accordingly, Table 7 specifies your current and future credences in *Example 3*.

Now, let's look at the payoffs. In this scenario, you have two options: rejecting the gamble (*Reject*) or accepting it (*Accept*). Let us say that the value of rejecting the gamble is 0, while the

	Worlds	
Credence Functions	r	w
$\pi(.\|P(r))$	0.99	0.01
$\pi(.\|P(w))$	0.99	0.01
$\pi(.\|Q(r))$	1	0
$\pi(.\|Q(w))$	0.99	0.01

Table 7: Your Credences in *Example 3*

value of accepting the gamble in r is 100 and the value of accepting it in w is $-10,000$. So, the values of the utility function μ stand as in Table 8.

	Worlds	
Acts	r	w
Accept	100	-10,000
Reject	0	0

Table 8: Payoffs for *Example 3*

Next, we ask which acts you prefer relative to your current and future bodies of evidence.

Current Preferences. In both r and w, by lights of your current credences, the expected value of taking the gamble is $0.99 \times 100 + 0.01 \times (-10000) = -1$, and the expected value of rejecting the gamble is 0. Therefore, if you satisfy Bayesian rationality, you will prefer to reject the gamble in each world.

Future Preferences. In r, by lights of your future credences, the expected value of taking the gamble is $1 \times 100 + 0 \times (-1000) = 100$ and the expected value of rejecting the gamble is 0. Therefore, provided you satisfy Bayesian rationality, you will prefer to accept the gamble in r. In w, your credences remain the same as before, so you will prefer to reject the gamble in w.

Therefore, your current and future preferences stand as in Table 9.

	Worlds	
Preference Functions	r	w
f	*Reject*	*Reject*
g	*Accept*	*Reject*

Table 9: Your Preferred Acts in *Example 3*

We can now plug in the values from Tables 7-9, and check whether Good's inequality

holds. For any $w^* \in W$,

$$\begin{aligned}
\sum_{w' \in W} \pi(w'|P(w^*))\mu(f(w'), w') &= \pi(r|P(w^*))\mu(f(r), r) + \pi(w|P(w^*))\mu(f(w), w) \\
&= 0.99 \times \mu(Reject, r) + 0.01 \times \mu(Reject, w) \\
&= 0.99 \times 0 + 0.01 \times 0 \\
&= 0. \\
\sum_{w' \in W} \pi(w'|P(w^*))\mu(g(w'), w') &= \pi(r|P(w^*))\mu(g(r), r) + \pi(w|P(w^*))\mu(g(w), w) \\
&= 0.99 \times \mu(Accept, r) + 0.01 \times \mu(Reject, w) \\
&= 0.99 \times 100 + 0.01 \times 0 \\
&= 99.
\end{aligned}$$

Since the expected value of performing the act you prefer relative to your future evidence is greater than the expected value of performing the you prefer relative to your current evidence, Good's inequality holds. Therefore, if we replace NEGATIVE ACCESS with nestedness, we can preserve Good's inequality. This might give us some hope of reconciling EVIDENCE EXTERNALISM with VALUE OF INFORMATION.

5 The Argument from Fallibility

In this section, I will argue that the reason for which the externalist rejects NEGATIVE ACCESS also makes nestedness unacceptable. That is why the externalist shouldn't try to save VALUE OF INFORMATION by replacing NEGATIVE ACCESS with nestedness.

5.1 An Example

In *Example 1*, when the wall is white but lit up with red light, you undergo a visual experience as of there being a red wall before you. Thus, your visual system malfunctions, and you gain no new evidence. But you are rationally confident that you are in the scenario where the wall is red, and your visual system has given you evidence that the wall is red. So, you can't eliminate the possibility that your evidence entails that the wall is red; as a result, NEGATIVE ACCESS fails. Therefore, it is your *fallibility*—i.e., the tendency of your information-gathering mechanisms to malfunction without giving you any warning that this has happened—which leads to the failure of NEGATIVE ACCESS in this scenario.[22]

Since this scenario involves just one source of information, i.e., your vision, there is just one scenario in which your information-gathering mechanism malfunctions. However, in scenarios where there are multiple sources of information involved, those sources of information could malfunction independently of each other, thereby misleading the agent in independent

[22]For elaboration of this point, see Salow [ms.].

ways. I am going to argue that in such scenarios, nestedness is bound to fail. So, the fallibility-based considerations that tell against NEGATIVE ACCESS in *Example 1* also tell against nestedness.

Consider the following example.

> *Example 4.* You are about to go into a room and encounter a wooden wall. You don't know for sure what the colour of the wall is, but you are rationally 0.99 confident that the wall is red. If the wall is red, you will see that it is red; so your evidence after entering the room will entail that the wall is red. However, there is a small probability of 0.01 that it is white, but will be lit up with red right when you enter the room. If that happens, your evidence will remain the same as before.
>
> You also don't know for sure what kind of wood the wall is made of, but you are rationally 0.99 confident that the wall is made of sandalwood. If the wall is made of sandalwood, you will be able to tell by smelling the wall that it is made of sandalwood; so, your evidence after entering the room will entail that it is made of sandalwood. However, there is a small probability of 0.01 that the wall is only be made of ordinary wood, but smeared with sandalwood perfume. If that happens, your evidence will remain the same as before.
>
> Relative to your current credence function, the possibility of the wall's being red is probabilistically independent of the possibility of its being made of sandalwood, while the possibility of the wall's being white is probabilistically independent of the possibility of its being made of ordinary wood. You also know that immediately afterwards, you will be offered a gamble where you stand to gain \$100 if the wall is red and made of sandalwood, and lose \$5,000 if the wall is either white or not made of sandalwood. Should you make a decision about the gamble before entering the room?

In this scenario, there are two different sources of evidence—vision and smell—which could malfunction independently of each other. When one malfunctions, you are rationally misled into thinking that you are in the scenario where it doesn't: when your vision malfunctions, you are rationally confident that the wall is red; similarly, when your sense of smell malfunctions, you are rationally confident that the wall is made of sandalwood.

We can see how this leads to a failure of nestedness. Let $W = \{rs, ro, ws, wo\}$ where rs is the world where the wall is red and made of sandalwood, ro is the world in which the wall is red and made of ordinary wood, ws is the world where the wall is white and made of sandalwood, and wo is the world where it is white and made of ordinary wood.

Let the frame $\langle W, P\rangle$ represent your current evidence. In each world w, your current evidence is $P(w) = \{rs, ro, ws, wo\}$ (Figure 5.1)

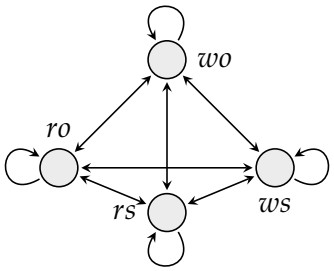

Figure 5.1: Your Current Evidence in *Example 4*

Since every world is compatible with the agent's evidence in every world, $\langle W, P\rangle$ satisfies reflexivity, transitivity, and euclideanness, and therefore is partitional.

Now, let the frame $\langle W, Q\rangle$ represent your future evidence. In the world *rs* where the wall is red and made of sandalwood, your evidence after you enter the room will rule out the worlds *ro*, *ws* and *wo*, where the wall is either not red or not made of sandalwood. But, in the world *ro* where you see a red wall made of ordinary wood, you evidence will include worlds where the wall is red, i.e., *rc* and *ro*, but not the worlds where it is white, i.e., *ws* and *wc*. Similarly, in the world *ws* where you see a white wall made of sandalwood, your evidence will include worlds where the wall is made of sandalwood, i.e., *rs* and *ws*, but not the worlds where the wall is made of ordinary wood, i.e., *ro* and *wo*. However, in *wo*, you gain no evidence, so your evidence won't rule out any of the worlds. Therefore, your future evidence is given by $Q(rs) = \{rs\}$, $Q(ro) = \{rs, ro\}$, $Q(ws) = \{ws, rs\}$, and $Q(wo) = \{rs, ro, ws, wo\}$ (Figure 5.2).

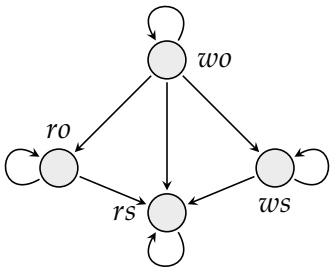

Figure 5.2: Your Future Evidence in *Example 4*

It should be clear from Figure 5.2 that the frame $\langle W, Q\rangle$ isn't partitional: it is reflexive and transitive, but not euclidean. Hence, FACTIVITY and POSITIVE ACCESS are true, but NEGATIVE ACCESS fails. More importantly for our purposes, it is not nested; for even though $Q(rs)$ and $Q(wc)$ aren't disjoint, neither of them is a subset of the other.

This failure of nestedness prevents Good's inequality from being true. Let $\langle W, P, A, \pi, \mu, f\rangle$ and $\langle W, Q, A, \pi, \mu, g\rangle$ be the decision-problems based on $\langle W, P\rangle$ and $\langle W, Q\rangle$ respectively.

We start by calculating your current and future credences under the assumption that you satisfy Bayesian rationality.

> *Current Credences*. Since the possibility of encountering a red wall and the possibility of encountering a wall made of sandalwood are probabilistically independent, your current credence in *rs* is $0.99 \times 0.99 = 0.9801$. Since the possibility of encountering a white wall and the possibility of encountering a wall made of

ordinary wood are probabilistically independent, your current credence in wo is $0.01 \times 0.01 = 0.0001$. Since it is 0.99 likely by your lights that you will encounter a red wall, your credence in ro is $0.99 - 0.9801 = 0.0099$. Similarly, since it is 0.99 likely that you will encounter a wall made of sandalwood, your current credence in ws is $0.99 - 0.9801 = 0.0099$.

Future Credences. In rs, your future evidence is $Q(rs) = \{rs\}$, i.e., the proposition that the wall is red and made of sandalwood. So, you assign credence 1 to rs and 0 to every other world. In ro, your future evidence is $Q(ro) = \{rs, ro\}$. So, you assign credence $\frac{0.9801}{0.9801 + 0.0099} = 0.99$ to rs, credence 0.01 to ro, and credence 0 to ws and wo. Analogously, in ws, your future evidence is $Q(ws) = \{rs, ws\}$. So, you assign credence $\frac{0.9801}{0.9801 + 0.0099} = 0.99$ to rs, credence 0.01 to ws, and credence 0 to ro and wo. Finally, in wo, your future evidence is $Q(wo) = \{rs, ro, ws, wo\}$; since your evidence remains the same as before, your credences also remain the same as earlier.

Table 10 summarizes the discussion above.

	Worlds			
Credence Functions	rs	ro	ws	wo
$\pi(.\|P(rs))$	0.9801	0.0099	0.0099	0.0001
$\pi(.\|P(ro))$	0.9801	0.0099	0.0099	0.0001
$\pi(.\|P(ws))$	0.9801	0.0099	0.0099	0.0001
$\pi(.\|P(wo))$	0.9801	0.0099	0.0099	0.0001
$\pi(.\|Q(rs))$	1	0	0	0
$\pi(.\|Q(ro))$	0.99	0.01	0	0
$\pi(.\|Q(ws))$	0.99	0	0.01	0
$\pi(.\|Q(wo))$	0.9801	0.0099	0.0099	0.0001

Table 10: Your Credences in *Example 4*

Let us now look at the payoffs. In this scenario, you have two options: rejecting the gamble (*Reject*) or accepting it (*Accept*). Let's say that, for every $w \in W$, the value of rejecting the gamble is 0; in rc, the value of accepting the gamble is 100, but for every other world, the value of accepting the gamble is $-5,000$. So, the values of the utility function μ stand as in Table 11.

	Worlds			
Credence Functions	rs	ro	ws	wo
Accept	100	-5,000	-5,000	-5,000
Reject	0	0	0	0

Table 11: Payoffs for *Example 4*

Next, we check what your current and future preferences are.

Current Preferences. By lights of you current credences, the expected value of accepting the gamble is $0.9801 \times 100 + 0.0099 \times (-5000) + 0.0099 \times (-5000) + 0.0001 \times$

$(-5000) = -1.49$. But the expected value of rejecting it is 0. If you satisfy Bayesian rationality, you will prefer to reject the gamble in every world.

Future Preferences. In *rs*, by lights of future credences: accepting the gamble maximizes expected value: the expected value of rejecting the gamble is 0, the expected value of accepting it is $1 \times 100 + 0 \times (-5000) = 100$. In *ro*, taking the gamble maximizes expected value: while the expected value of rejecting the gamble is 0, the expected value of taking the gamble is $0.99 \times 100 + 0.01 \times (-5000) = 49$. The same is true for *ws*. Finally, in *wo*, since your credences remain the same as before, rejecting the gamble maximizes expected value. If you satisfy Bayesian rationality, in the worlds *rs*, *ro*, and *ws*, you prefer to accept the gamble relative to your future credences, but in *wo*, you prefer to reject it.

Table 12 lays out your preferred acts relative to your current and future bodies of evidence, given by the functions f and g respectively.

	Worlds			
Preference Functions	*rs*	*ro*	*ws*	*wo*
f	*Reject*	*Reject*	*Reject*	*Reject*
g	*Accept*	*Accept*	*Accept*	*Reject*

Table 12: Your Preferred Acts in *Example 4*

It is easy to see now that Good's inequality doesn't hold in this scenario. For any $w \in W$,

$$\begin{aligned}
\sum_{w' \in W} \pi(w'|P(w))\mu(f(w'), w') &= \pi(rs|P(w))\mu(f(rs), rs) + \pi(ro|P(w))\mu(f(ro), ro) \\
&\quad + \pi(ws|P(w))\mu(f(ws), ws) + \pi(wo|P(w))\mu(f(wo), wo) \\
&= 0.9801 \times \mu(Reject, rs) + 0.0099 \times \mu(Reject, ro) \\
&\quad + 0.0099 \times \mu(Reject, ws) + 0.0001 \times \mu(Reject, wo) \\
&= 0.9801 \times 0 + 0.0099 \times 0 + 0.0099 \times 0 + 0.0001 \times 0 \\
&= 0. \\
\sum_{w' \in W} \pi(w'|P(w))\mu(g(w'), w') &= \pi(rs|P(w))\mu(g(rs), rs) + \pi(ro|P(w))\mu(g(ro), ro) \\
&\quad + \pi(ws|P(w))\mu(g(ws), ws) + \pi(wo|P(w))\mu(g(wo), wo) \\
&= 0.9801 \times \mu(Accept, rs) + 0.0099 \times \mu(Accept, ro) \\
&\quad + 0.0099 \times \mu(Accept, ws) + 0.0001 \times \mu(Reject, wo) \\
&= 0.9801 \times 100 + 0.0099 \times (-5,000) + 0.0099 \times (-5,000) + 0.0001 \times 0 \\
&= -0.99.
\end{aligned}$$

Hence, the expected value of performing the act you prefer relative your future evidence is less than the expected value of performing the act you relative to your current evidence.

5.2 The Argument

In this example, Good's inequality fails because there are two distinct scenarios—i.e., the worlds *ro* and *ws*—where the agent rationally confident that the wall is red and made of sandalwood, when it isn't so. In these two scenarios, the agent accepts the gamble and loses money as a result. In the scenario where the wall is red and made of sandalwood, the agent is certain that the wall is red and made of sandalwood. In this scenario, too, she accepts the gamble, and makes some money. However, the money she expects to make in this scenario is too little to make up for losses she expects to incur in the other scenarios. That is why the expected value of performing the act she prefers according to her future evidence is less than the expected value of performing the act she prefers according to her current evidence.

A crucial feature of *Example 3* is that the agent is misled about the truth in multiple independent ways. In *ro*, she is misled about what the wall is made of, because her sense of smell isn't reliable. In *ws*, she is misled about the colour of the wall, because her vision malfunctions. In each of these scenarios, the agent has conclusive evidence about one subject-matter, but gains no evidence about the other. As a result, the agent's bodies of evidence in these two scenarios are not disjoint from each other, but still don't entail each other. That is why nestedness fails.

Note, however, things wouldn't be the same if we were to assume that vision and smell cannot malfunction independently of each other. To see this, consider a modification of $\langle W, Q \rangle$ where, in *ro* and *ws*, the agent's evidence includes all the worlds in W. The frame will now look like this.

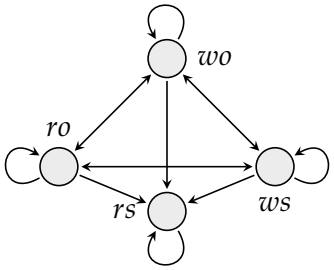

Figure 5.4: Your Future Evidence Without Independent Malfunctioning of Vision and Smell

In this scenario, when the agent encounters either a red wall made of ordinary wood or a white wall made of sandalwood, she cannot gain any evidence from vision or smell: the agent's visual system cannot independently provide evidence about colour when her sense of malfunctions, and her sense of smell cannot provide evidence about what the wall is made of when her sense of vision malfunctions.

This, effectively, makes the frame nested, thereby preserving Good's inequality. In *wo*, *ro* and *ws*, the agent doesn't gain any new evidence, and therefore prefers to reject the gamble relative to her future evidence. By contrast, in *rs*, the agent learns that the wall is red and made of sandalwood, and prefers to accept the gamble. Hence, if the agent performs the act that she prefers according to her future evidence, she won't lose any money in *ro*, *ws*, and *wo*, but will definitely gain money in *rs*. That is why it is better for her to act according to her

future preferences. This reveals that, in *Example 4*, Good's inequality fails, because the agent's senses of vision and smell malfunction independently of each other.

The lesson I want to draw is this. The strongest reason for rejecting NEGATIVE ACCESS has to do with our fallibility: even when a source of information malfunctions and doesn't yield any evidence, an agent could still have misleading evidence which suggests that the information conveyed by this source counts as evidence. If we take on board possibilities where an agent is misled about the truth due to the malfunctioning of *one* source of information, we should have no qualms countenancing scenarios where an agent is misled about the truth in multiple independent ways due to the independent malfunctioning of *several* sources of information. But, now, in scenarios where there are multiple sources of information which could malfunction independently of each other, the agent's fallibility might not only make NEGATIVE ACCESS false, but could also lead to failures of nestedness. Therefore, there is no principled position which rejects NEGATIVE ACCESS on grounds of such fallibility, but accepts nestedness.

In response to this argument, the externalist might be tempted to defend nestedness by appealing to Geanakoplos' interpretation of nestedness as a property of memory. Geanakoplos' rough idea was that the propositions that are part of an agent's evidence are always derived from a list of propositions $P_1, P_2, ..., P_m$ such that, if P_i is part of an agent's evidence, then, for every P_j with $j \leq i$,P_j is also part of her evidence. This means an agent cannot possess a piece of evidence from the list without also possessing propositions that occur earlier in the list. That is why it is not possible for the agent to be in two scenarios where she possesses two bodies of evidence such that the two overlap with each other, but one of them isn't identical to or stronger than the other.

This constraint is implausible in scenarios where an agent can acquire evidence about multiple orthogonal subject-matters, e.g. about the colour of the wall and the stuff that the wall is made of, using independent sources of information. Suppose, in *Example 4*, the two propositions on the list are $\{rs, ro\}$ and $\{rs, ws\}$, i.e., the claim that the wall is red and the claim that the wall is made of sandalwood. Since the agent's vision and smell can malfunction independently of each other, the agent could learn each of these propositions without learning the other. Therefore, the agent's evidence could entail the first proposition without entailing the second, or could entail the second proposition without entailing the first. Yet, at the same time, the agent may have misleading evidence for thinking that she is in the scenario where her senses of vision and smell don't malfunction, and therefore that her evidence entails both the claim that the wall is red and the claim that the wall is made of sandalwood. When that happens, nestedness will have to fail.

Thus, nestedness turns out to be an implausible constraint on frames for representing evidence.

5.3 A General Strategy

I have shown two things. First, I have shown that when both POSITIVE and NEGATIVE ACCESS fail, Good's inequality can fail. Then, I have considered a proposal that doesn't reject POSITIVE ACCESS, but tries to save Good's inequality by replacing NEGATIVE ACCESS with nestedness. In response, I have shown that the same fallibility-based considerations that tell against NEGATIVE ACCESS also tell against nestedness.

Still, we might wonder whether there is a property of frames weaker than nestedness, which, together with FACTIVITY and POSITIVE ACCESS, preserves Good's inequality. Let me now sketch a general strategy for arguing against any such proposed substitute. Whatever that proposed substitute might be, it must rule out scenarios like the one described in *Example 4*. In other words, it must rule out the possibility that an agent could have multiple sources of evidence which can malfunction independently of each other, thereby misleading the agent in independent ways about the truth.

Such a condition should be incompatible with any plausible version of EVIDENCE EXTERNALISM.[23] Here is what I take to be the naïve externalist conception of evidence: we get evidence about the external world using multiple independent sensory channels, which could malfunction independently of each other. Hence, it is possible for our sense of smell to malfunction independently of our visual system, and vice-versa. Yet, in a scenario where only one of them is malfunctioning, we may indeed have misleading evidence for thinking that both our senses of vision and smell are functioning properly. So, the kind of scenario where Good's inequality fails is built into our naïve externalist conception of evidence. Therefore, the burden is upon the defender of Good's inequality to tell us why this naïve picture is false. Unless she does so, she cannot reconcile Good's inequality with EVIDENCE EXTERNALISM.

6 Objections from Self-Evidence

In this section, I shall entertain two related objections that the evidence externalist might raise against the arguments presented in §3 and §5. In stating these objections, I shall appeal to a property of propositions called *self-evidence*.

Definition 5. A proposition X is *self-evident* relative to a frame $\langle W, P\rangle$ if and only

[23] Nestedness and its weaker substitutes might be considered especially incompatible with a conception of evidence on which our evidence consists all and only of propositions that we know. Some writers, such as Lenzen [1978] and Stalnaker [2006], think that the class of frames that best represent knowledge are reflexive, transitive, and strongly convergent, where *strong convergence* is defined as follows:

> A frame $\langle W, P\rangle$ is *strongly convergent* if and only if, for any world $w \in W$, there exists a world $w' \in W$ such that, for every $w'' \in W$ which is compatible with $P(w)$, w'' is compatible $P(w')$.

Stalnaker [2006] motivates strong convergence with reference to scenarios like *Example 4* where the agent could gain knowledge from multiple sources, which might in turn malfunction independently of each other. In fact, the non-nested frame that represents the agent's evidence in *Example 4* is reflexive, transitive, and strongly convergent. Therefore, those who think that our evidence consists of all and only those propositions that we know and also take knowledge to be best represented by reflexive, transitive, and convergent frames, should reject nestedness and all its weaker substitutes.

if, for any $w \in W$, if $w \in X$, then the agent's evidence $P(w)$ at w entails X. More formally,

$$(\forall w \in W)(w \in X \Rightarrow P(w) \subseteq X).$$

Self-evidence is the evidential analogue of what Williamson [2000] calls *luminosity*. A condition C is luminous for an agent if and only if whenever it obtains, the agent is in a position to know that it obtains. Similarly, a proposition X is self-evident if and only if whenever it is true, the agent's evidence entails that it is true.

6.1 The Objection from Self-Evident Preferences

We might think that our preferences are always self-evident in this sense: if we prefer a certain act, then we would be inclined to act in certain ways, which in turn might give us guaranteed evidence about what we prefer. So, we may think that all decision problems satisfy a property that we may call *self-evident preferences*.

Definition 6. A decision problem $D = \langle W, P, A, \pi, \mu, f \rangle$ satisfies *self-evident preferences* if and only if, for any $w \in W$, if an agent's preferred act at w relative to her evidence is $f(w)$, then her current evidence entails that it is $f(w)$. More formally,

$$(\forall w \in W)(P(w) \subseteq \{w' : f(w') = f(w)\}.$$

If this seems like a plausible constraint on decision problems, then *Example 2*—the example where both POSITIVE ACCESS and NEGATIVE ACCESS fail—may seem unrealistic. In that scenario, when the wall is either crimson or cardinal red, you prefer to accept the relevant gamble relative to your future evidence, but it is compatible with your future evidence that the wall is rusty red, and therefore that you don't prefer to accept the gamble. Similarly, when the wall is rusty red, you prefer to reject the bet relative to your future evidence, but it is compatible with your evidence that the wall is either crimson or cardinal red, and therefore that you prefer to accept the gamble. So, your preferences in *Example 2* are not self-evident.

Now, consider a version of *Example 2*, where once you form your preferences, they become evident to you. So, when the wall is either crimson or cardinal red, and you prefer to accept the relevant gamble, then you become certain that you prefer to accept the relevant gamble. From that, you should conclude that the wall is either crimson or cardinal red, but not rusty red. But, then, you wouldn't prefer to accept the gamble any more. Now, if you were to act according to this revised future preference, you wouldn't lose any money. Thus, Good's inequality would hold in that scenario.

This objection isn't promising. First of all, if we like Williamson's [2000] anti-luminosity argument, then we may argue that our preferences cannot always be self-evident. Typically, defenders of EVIDENCE EXTERNALISM, such as Williamson [2000] himself, require our evidence to include only reliably or safely acquired information. But our ability to discriminate our preferences is limited. Therefore, there may indeed be cases where an agent like us prefers to perform an act, but cannot safely or reliably determine whether she prefers to perform that act.

In such cases, the agent's preferences won't be self-evident.

Moreover, even if we accept the thesis that an agent's preferences are always self-evident, Good's inequality may still fail. To see why, focus on *Example 4*. In that example, after you enter the room, there is only one possible scenario where your evidence doesn't entail what your preferences are: namely, the world *wo* where the wall is white and made of ordinary wood. In that scenario, you prefer to reject the gamble, because you receive no new evidence about the colour of the wall or the stuff that it is made of. Yet, it remains compatible with your future evidence that you prefer to accept the gamble; for you can't rule out the possibility that you are in a scenario where the wall is either red or made of sandalwood.

Now, consider a variant of this example, where you not only undergo the same experience upon the entering the room as you did in the original example, but also get an extra piece of evidence from a trustworthy informant in *wo*. In *wo*, you learn from the informant that you are in *wo*. Therefore, in *wo*, your evidence rules out every world other than *wo*. Everywhere else, your future evidence remains the same as it was in the original example. So, your future evidence can be represented by the frame $\langle W, Q\rangle$ such that $Q(wo) = \{wo\}$ and $Q(ro) = \{rs, ro\}$, $Q(ws) = \{rs, ws\}$, and $Q(rs) = \{rs\}$ (Figure 6.1).

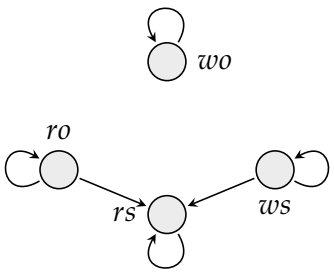

Figure 6.1: Your Future Evidence in the Modified Version of *Example 4*

This frame satisfies *self-evident preferences*. In *wo*, your future evidence entails that the wall isn't red or made of sandalwood; so, you prefer to reject the gamble relative to your future evidence. Since *wo* is the only world that is compatible with your future evidence in *wo*, your evidence entails that you prefer to reject the gamble. In every other world, by the same reasoning as in the original example, you prefer to accept the gamble relative to your future evidence. Since the only worlds compatible with your future evidence in those cases are also worlds where you prefer to accept the gamble, your future evidence entails that you prefer to accept the gamble.

Importantly, in this variant of *Example 4*, your future preferences in each world remain the same as it was in the original example. Therefore, the expected value of performing an act recommended by your future evidence is lower than the expected value of performing an act that you prefer according to your current evidence. Hence, appealing to self-evidentness of preferences doesn't help us save Good's inequality.

We can offer a more principled diagnosis of why this strategy cannot succeed in general. Suppose an agent's preferences are self-evident. When her future evidence satisfies FACTIVITY, but doesn't satisfy one or both of the access principles, Good's inequality can hold only if the frame that represents her future evidence also satisfies a further property which Geanakoplos

[1989] calls *positive balancedness*.

> **Definition 7.** Let, for any set of worlds X, I_X be the characteristic function of X, i.e., $I_X(w) = 1$ if $w \in X$, and 0 otherwise. A frame $F = \langle W, P \rangle$ satisfies *positive balancedness* with respect to a proposition E if and only if there exists a function $\lambda : \{P(w) : w \in W\} \to \mathbb{R}_{\geq 0}$ such that for all $w \in W$,
>
> $$\sum_{P(w') \subseteq E} \lambda(P(w')).I_{P(w')}(w) = I_E(w).$$[24]

Let a frame satisfy *positive balancedness* if and only if it is positively balanced with respect to every self-evident proposition. Roughly, a frame is positively balanced if and only if, for any self-evident E, every body of evidence $P(w')$ which entails E has an intensity $\lambda(P(\omega'))$ such that, for any $w \in E$, the sum of the intensities of every $P(w')$ which doesn't eliminate w is 1. This is intended to be a generalization of partitionality: all partitional frames satisfy positive balancedness, but not all positively balanced frames are partitional.[25] Therefore, the combination of reflexivity and positive balancedness is weaker than partitionality.

Geanakoplos [1989] proves the following theorem:

> **Geanokoplos' Second Theorem (Theorem 4, Geanakoplos 1989).** Let $D_1 = \langle W, P, A, \pi, \mu, f \rangle$ and $D_2 = \langle W, Q, A, \pi, \mu, g \rangle$ be two decision problems which satisfy Bayesian rationality and self-evident preferences, such that $\langle W, P \rangle$ is partitional and P is coarser than Q. If the frame $\langle W, Q \rangle$ satisfies positive balancedness and reflexivity, then Goods inequality holds. Conversely, if the frame $\langle W, Q \rangle$ does not satisfy reflexivity or does not satisfy positive balancedness, then there is a decision problem D_1 relative to which Goods inequality is strictly reversed.

Now, the frame $\langle W, Q \rangle$ depicted in Figure 6.1 satisfies self-evident preferences, but doesn't satisfy positive balancedness.[26] That is why Good's inequality fails in this example.

---

[24] My definition of positive balancedness, borrowed from Brandenburger, Dekel, and Geanakoplos (1992, p. 185), is a simplified version of the original definition that Geanakoplos [1989] proposes.

[25] *Proof*: The second conjunct is easily proved: the frame $\langle W, Q \rangle$ depicted in Figure 3.2 is positively balanced, but is neither transitive nor euclidean. In that frame, there is one self-evident proposition, namely, $W = \{cr, ru, ca\}$. If we let $\lambda(Q(ru)) = 1$ and $\lambda(Q(cr)) = \lambda(Q(cr)) = 0$, then the frame is positively balanced with respect to W.

Let us now show that every partitional frame is positively balanced. Suppose $\langle W, P \rangle$ is partitional. Now, consider any proposition E that is self-evident relative to $\langle W, P \rangle$. For any w in W, either w is in E or it isn't. If w is not in E, then $I_E(w) = 0$. But then, there is no $P(w^*) \subseteq E$ such that $w \in P(w^*)$. So,

$$\sum_{P(w^*) \subseteq E} \lambda(P(w^*)) I_{P(w^*)}(w) = 0 = I_E(w).$$

If w is in E, then $I_E(w) = 1$. Since E is evident, $P(w) \subseteq E$. By the partitionality of $\langle W, P \rangle$, for any $P(w^*) \subseteq E$, $w \in P(w^*)$ if and only if $P(w^*) = P(w)$. Let λ be the function such that, for any $P(w^*) \subseteq E$, $\lambda(P(w^*)) = 1$. So, for any $P(w^*) \subseteq E$, $\lambda(P(w^*)) I_{P(w^*)}(w) = 1$ if and only if $P(w^*) = P(w)$. For any other $P(w^*) \subseteq E$, $\lambda(P(w^*)) I_{P(w^*)}(w) = 0$. This means that

$$\sum_{P(w^*) \subseteq E} \lambda(P(w^*)) I_{P(w^*)}(w) = 1 = I_E(w).$$

Therefore, the frame is positively balanced.

[26] *Proof*: Suppose, for *reductio*, that the frame $\langle W, Q \rangle$ is positively balanced. Now, consider the self-evident proposition $E = \{rs, ro, ws\}$. Then, there exists a function λ with non-negative values such that, for every w,

$$\sum_{Q(w') \subseteq E} \lambda(P(w')).I_{Q(w')}(w) = I_E(w).$$

Now suppose $w = rs$. Since rs is compatible with the agent's evidence in each possibility in E, $I_{Q(rs)}(rs) =$

6.2 The Objection from Self-Evident Credences

We might think that an agent who has sufficiently powerful capacities for introspection should always be able to learn what her own credences are. So, we may think that an agent's credences are always self-evident. Therefore, we may be tempted to impose on all decision problems a constraint that we may call *self-evident credences*.

> **Definition 9.** A decision problem $D = \langle W, P, A, \pi, \mu, f \rangle$ satisfies *self-evident credences* if and only if, for any $w \in W$, if an agent's credence function in w is $\pi(.|P(w))$, then her current evidence entails that it is $\pi(.|P(w))$. More formally,
>
> $$(\forall w \in W)(P(w) \subseteq \{w' : \pi(.|P(w)) = \pi(.|P(w'))\}.$$

This constraint is incompatible with our description of *Example 4*. In *Example 4*, when the wall is either not red or not made of sandalwood, you are rationally less than certain that it is red and made of sandalwood. Yet, it remains compatible with your evidence that your evidence entails that claim. So, you cannot rule out the possibility that you are rationally certain that the wall is red and made of sandalwood. Therefore, your credences are not self-evident.

Now, consider a variant of in *Example 4*, where your credences are self-evident. When you are less than certain that the wall is red and made of sandalwood, your evidence also entails that you are less than certain about that claim. If that happens, you would also learn that you are not in the scenario where the wall is red and made of sandalwood. Therefore, you won't accept the gamble, and lose money in those cases. Thus, Good's inequality would hold.

We can respond to this worry in the same way as we did to the objection from *self-evident preferences*. Using the same strategy implicit in Williamson's anti-luminosity argument, we might argue that a rational agent may not always be a position to safely determine what her own credences are. So, her evidence may not always entail what her credences are.

It is also worth noting that the requirement of self-evident credences has disastrous consequences when it is combined with Bayesian rationality.

> **Theorem**. If a decision problem $D = \langle W, P, A, \pi, \mu, f \rangle$ satisfies self-evident credences and Bayesian rationality, then $\langle W, P \rangle$ is both transitive and euclidean.[27]

$I_{Q(ro)}(rs) = I_{Q(ws)}(rs) = 1$. Moreover, $I_E(rs) = 1$. So,

(a) $$\sum_{Q(w') \subseteq E} \lambda(P(w')).I_{Q(w')}(rs) = \lambda(Q(rs)) + \lambda(Q(ro)) + \lambda(Q(ws)) = 1.$$

Suppose $w = ro$. Then, since ro is incompatible with the agent's evidence in ws and rs, $I_{Q(ws)}(ro) = I_{Q(rs)}(ro) = 0$. But $I_{Q(ro)}(ro) = 1$. Moreover, $I_E(ro) = 1$. So,

(b) $$\sum_{Q(w') \subseteq E} \lambda(P(w')).I_{Q(w')}(ro) = \lambda(Q(ro)) = 1.$$

But now suppose $w = ws$. Then, since ws is incompatible with the agent's evidence in rs and ro, $I_{Q(rs)}(ws) = I_{Q(ro)}(ws) = 0$. But $I_{Q(ws)}(ws) = 1$. Moreover, $I_E(ws) = 1$. So,

(c) $$\sum_{Q(w') \subseteq E} \lambda(P(w')).I_{Q(w')}(ws) = \lambda(Q(ws)) = 1.$$

Since the values of λ are non-negative, (c) is inconsistent with (a) and (b). Therefore, the frame is not positively balanced.

[27] *Proof.* Suppose, for *reductio*, that there exists a decision problem $D = \langle W, P, A, \pi, \mu, f \rangle$ which satisfies self-

Hence, a decision problem which satisfies both self-evident credences and Bayesian rationality validates both POSITIVE ACCESS and NEGATIVE ACCESS.

Less formally, the argument is as follows. In cases where POSITIVE ACCESS NEGATIVE ACCESS fails, the agent's credences won't be self-evident to her. Focus on failures of NEGATIVE ACCESS like *Example 3*. If you satisfy Bayesian rationality, and you learn that the wall is red, then, according to CONDITIONALIZATION, you will assign credence 1 to the claim that the wall is red. When the wall is white but lit up with red light and you satisfy Bayesian rationality, you won't assign credence 1 to the claim that the wall is red (provided that your ur-prior is regular, i.e., assign non-zero probability to all non-empty propositions). In that scenario, it remains compatible with your evidence that your evidence entails that the wall is red. But, if you are also certain that you adopt rational doxastic attitudes, then, it is compatible with your evidence that you are certain that the wall is red. Therefore, when the wall is white, your credences won't be self-evident to you.

The lesson is this. If the evidence externalist wants to save Good's inequality by requiring the agent's credences to be self-evident to her even in cases where NEGATIVE ACCESS fails, she must either give up CONDITIONALIZATION or the assumption that the agent's ur-prior is regular. In *Example 3*, there are two ways of making your future credences self-evident to you: either we make it irrational for you to assign credence 1 to the claim that the wall is red when your evidence entails that the wall is red, or we allow to you assign credence 1 to the claim that the wall is red even when the wall is white. The first strategy licenses violations of CONDITIONALIZATION. The second requires your ur-prior to be irregular; for, as long as you satisfy CONDITIONALIZATION, you can only assign credence 1 to the proposition that the wall is red when it is white if your ur-prior attaches probability 0 to the possibility that it might be white.

However, none of these strategies seem reasonable. With respect to the first strategy, it is hard to see how it could be irrational for an agent to be certain in a claim for which she has conclusive evidence. In relation to the second, it is hard to see how it could be rational for you to assign credence 0 to the possibility that the wall is white independently of all empirical investigation. So, the evidence externalist cannot save VALUE OF INFORMATION by just appealing to self-evident credences.

---

evident credences and Bayesian rationality, and is based on a reflexive frame $\langle W, P\rangle$, but $\langle W, P\rangle$ is not transitive. Then, there are three worlds w, w', w'', such that $w' \in P(w)$ and $w'' \in P(w')$, but $w'' \notin P(w)$. Since π is regular by the definition of a decision problem, $\pi(w''|P(w')) > 0$, but $\pi(w''|P(w)) = 0$. However, then, self-evident credences is violated. Contradiction.

Suppose, for *reductio*, that there exists a decision problem $D = \langle W, P, A, \pi, \mu, f\rangle$ which satisfies self-evident credences and Bayesian rationality, and is based on a reflexive frame $\langle W, P\rangle$, but $\langle W, P\rangle$ is not euclidean. Then, there are three worlds w, w', w'', such that $w' \in P(w)$ and $w'' \in P(w)$, but $w'' \notin P(w')$. Since π is regular by the definition of a decision problem, $\pi(w''|P(w)) > 0$, but $\pi(w''|P(w')) = 0$. Once again, self-evident credences is violated. Contradiction.

7 Despair

This concludes my defence of the claim that EVIDENCE EXTERNALISM is in tension with VALUE OF INFORMATION.

How should we respond to this claim? One response might be to reject EVIDENCE EXTERNALISM, and adopt a Cartesian picture of evidence. But, then, the defender of the Cartesian picture would face the sceptical challenge that we raised earlier: she would have to say how we ever come to form justified beliefs about the world in the absence of prior empirical evidence about the reliability of our perceptual and cognitive mechanisms. Some writers, such as Vogel [1990] and Pryor [2000], have responded to this challenge. For Vogel, we can have reason to believe certain claims about the external world by inference to the best explanation alone, irrespective of our background evidence. On the view that Pryor defends—what he calls *dogmatism*—perceptual experiences provide *prima facie* justification for beliefs about the external world even in the absence of any background evidence about the veridicality of such experiences.

Unfortunately, these views are not able to save VALUE OF INFORMATION. Imagine an agent who, on the basis of inductive evidence, is rational to believe that she will undergo an experience as of there being a red wall before her. According to both Vogel and Pryor, if she does undergo the experience, she will be rational to believe the claim that the wall is red. Therefore, the agent in question is rational to believe that she will be rational in the future to believe that the wall is red. However, both Vogel and Pryor want to say that, prior to actually undergoing the experience, the agent needn't have any reason to be confident that there will in fact be a red wall before her when she undergoes the "red wall" experience. In fact, it may indeed be more likely on her evidence that, conditional on her undergoing an experience as of there being a red wall before her, she will be facing a white wall lit up with red light, than that she will be facing a red wall. Hence, there is good reason for her to believe that she will only get misleading evidence from her "red wall" experience. Therefore, it might be instrumentally rational for her not to have that experience. This conflicts with VALUE OF INFORMATION.

The only Cartesian account that doesn't conflict with VALUE OF INFORMATION is the one defended by writers like Crispin Wright [2004] and Roger White [2006]. Both of these writers deny the assumption that we need empirical evidence in order to be justified in taking our perceptual and cognitive faculties to be reliable. For Wright [2004], we are *epistemically entitled* on purely non-epistemic grounds (e.g., for pragmatic reasons, for reasons having to do with our practices of inquiry, etc.) to accept the claim that our ordinary methods of belief-formation are reliable. By contrast, for White [2006], we are justified *a priori* in ruling out sceptical possibilities where our perceptual and cognitive faculties mislead us. Both these views seem to entail that it is rationally permissible for us to have a bias against a class of contingent hypotheses, namely those on which our faculties provide misleading information, independently of all empirical evidence whatsoever. These views might strike us as counterintuitive. Wright's view seems to run contrary to a widely accepted evidentialist approach to epistemic rationality, which requires all agents to proportion their doxastic attitudes to the evidence they possess. White's

view, by contrast, licenses a strong form of rationalism, on which we have *a priori* justification for believing certain contingent claims about the world.

If we can preserve VALUE OF INFORMATION only by accepting such views, it might just be better to accept the seemingly implausible conclusion that gathering and using cost-free information isn't always instrumentally rational. There is little hope for VALUE OF INFORMATION.

References

William P. Alston. Level-confusions in epistemology. *Midwest Studies in Philosophy*, 5(1):135–150, 1980.

Louise Antony. A naturalized approach to the a priori. *Philosophical Issues*, 14(1):1–17, 2004.

Michael Bacharach. Some Extensions of a Claim of Aumann in an Axiomatic Model of Knowledge. *Journal of Economic Theory*, 37(1):167–190, 1985.

Matthew A. Benton. Dubious objections from iterated conjunctions. *Philosophical Studies*, 162 (2):355–358, 2013.

Adam Brandenburger, Eddie Dekel, and John Geanakoplos. Correlated equilibrium with generalized information structures. *Games and Economic Behavior*, 4(2):182–201, 1992.

Lara Buchak. Instrumental rationality, epistemic rationality, and evidence-gathering. *Philosophical Perspectives*, 24(1):85–120, 2010.

Alex Byrne. Perception and evidence. *Philosophical Studies*, 170(2):101–113, 2013.

Nilanjan Das and Bernhard Salow. Transparency and the KK principle. *Noûs*, forthcoming.

Fred Dretske. Externalism and modest contextualism. *Erkenntnis*, 61(2-3):173–186, 2004.

John Geanakoplos. Common knowledge, Bayesian learning, and market speculation with bounded rationality. Technical report, mimeo, Yale University, 1988.

John Geanakoplos. Game theory without partitions, and applications to speculation and consensus. Technical report, 1989. Cowles Foundation Discussion Paper No. 914, Yale University.

John Geanakoplos. Common knowledge. In R. Aumann and S. Hart, editors, *Handbook of Game theory With Economic Applications*, volume 2, pages 1437–1496. Leiden: Elsevier, 1994.

Alvin Goldman. Williamson on knowledge and evidence. In Patrick Greenough and Duncan Pritchard, editors, *Williamson on Knowledge*, pages 73–92. Oxford: Oxford University Press, 2009.

I. J. Good. On the principle of total evidence. *British Journal for the Philosophy of Science*, 17(4): 319–321, 1967.

I. J. Good. A little learning can be dangerous. *British Journal for the Philosophy of Science*, 25(4): 340–342, 1974.

Daniel Greco. Could KK be OK? *Journal of Philosophy*, 111(4):169–197, 2014.

Jaakko Hintikka. *Knowledge and Belief*. Ithaca: Cornell University Press, 1962.

Simon M. Huttegger. Learning experiences and the value of knowledge. *Philosophical Studies*, 171(2):279–288, 2014.

James Joyce. Williamson on evidence and knowledge. *Philosophical Books*, 45(4):296–305, 2004.

Joseph B. Kadane, Mark Schervish, and Teddy Seidenfield. Is ignorance bliss? *Journal of Philosophy*, 105(1):5–36, 2008.

Adam Leite. But that's not evidence; it's not even true! *Philosophical Quarterly*, 63(250):81–104, 2013.

Wolfgang Lenzen. Recent work in epistemic logic. *Acta Philosophica Fennica*, 30(2):1–219, 1978.

Clayton Littlejohn. *Justification and the Truth-Connection*. Cambridge: Cambridge University Press, 2012.

Berislav Maruši. The self-knowledge gambit. *Synthese*, 190(12):1977–1999, 2013.

John McDowell. Knowledge and the internal. *Philosophy and Phenomenological Research*, 55(4): 877–93, 1995.

John Henry McDowell. *Perception as a Capacity for Knowledge*. Marquette University Press, 2011.

Charles S Peirce. Note on the theory of the economy of research. *Operations Research*, 15(4): 643–648, 1967.

James Pryor. The skeptic and the dogmatist. *Noûs*, 34(4):517–549, 2000.

Frank P Ramsey. Weight or the value of knowledge. *The British Journal for the Philosophy of Science*, 41(1):1–4, 1990.

Bernhard Salow. Elusive Externalism. ms. Unpublished Manuscript.

Dov Samet. Ignoring ignorance and agreeing to disagree. *Journal of Economic Theory*, 52(1): 190–207, 1990.

Hyun Song Shin. Non-partitional information on dynamic state spaces and the possibility of speculation. Technical report, Michigan-Center for Research on Economic & Social Theory, 1989.

Hyun Song Shin. Logical structure of common knowledge. *Journal of Economic Theory*, 60(1): 1–13, 1993.

Brian Skyrms. The value of knowledge. *Minnesota Studies in the Philosophy of Science*, 14:245–266, 1990.

Robert Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128(1):169–199, 2006.

Robert Stalnaker. On Hawthorne and Magidor on assertion, context, and epistemic accessibility. *Mind*, 118(470):399–409, 2009.

Robert Stalnaker. Luminosity and the KK thesis. In Sanford C. Goldberg, editor, *Externalism, Self-Knowledge, and Skepticism*, volume 1, pages 167–196. Cambridge: Cambridge University Press, 2015.

Jonathan Vogel. Cartesian skepticism and inference to the best explanation. *Journal of Philosophy*, 87(11):658–666, 1990.

Roger White. Problems for dogmatism. *Philosophical Studies*, 131(3):525–57, 2006.

Michael Williams. *Unnatural Doubts: Epistemological Realism and the Basis of Scepticism*. B. Blackwell, 1991.

Timothy Williamson. *Knowledge and its Limits*. Oxford: Oxford University Press, 2000.

Timothy Williamson. Improbable knowing. In Trent Dougherty, editor, *Evidentialism and its Discontents*. Oxford: Oxford University Press, 2011.

Crispin Wright. Warrant for nothing (and foundations for free)? *Aristotelian Society Supplementary Volume*, 78(1):167–212, 2004.