Knowledge-based Decision Theory and the New Evil Demon*

20th February 2014

Abstract

The idea that being rational is doing what is best in view of what one knows has some intuitive appeal, fits nicely with attractive views of evidence, reasons and the application of decision theory, and promises to bring some theoretical unity to epistemology. However, it has hardly ever been seriously considered, because it appears to fly in the face of what epistemologists call the New Evil Demon problem. There is a loophole in the problem, however. While I show that it cannot be exploited by the most straightforward implementation of the view — Knowledge-based Decision

Theory (Hawthorne and Stanley, 2008, Weatherson, 2012) —, I sketch an alternative implementation — the Two-Step account — that shows more promise.

Being rational is, roughly, doing the reasonable, intelligent, smart, sensible thing; being irrational doing the senseless, stupid, idiotic, crazy thing (Parfit, 2011, 33). Being rational is not doing what is best. Some smart decisions are disastrous and some idiotic ones turn out good. Rather, being rational is, roughly, doing what is best *as far as one can see*. What is best as far as one can see is what appears best in view of the information one has. But the information one has is just what one knows. Thus being rational is, roughly, doing the best in view of what one knows.

The idea may sound like a platitude but it is rejected by all currently dominant theories of rationality. Subjective Bayesians take rationality to be a matter of doing what is best in view of one's (graded) *beliefs*. Many philosophers take it to be a matter of doing what is best in view of one's internal *evidence*, which consists in various internal mental states — seemings and feelings.[1] Others take it to be a matter of one's action or beliefs being the output of *reliable cognitive processes* or *virtues*, where the relevant processes are internal to one's person.[2] It is thus worth mentioning a few considerations that make the knowledge view attractive

---

[1]Parfit (2011) illustrates both views. On the one hand he says that it is rational to believe, want or do something when "what we believe would, if it were true, give us [decisive] reason" to believe, want or do it (111). When giving a more fine-grained account, however, he says that the rationality of some of our beliefs does not depend on our other beliefs (as the above suggested) but on our "evidence" (129).

[2]Along the lines of what Goldman (1986) says about justification. Goldman himself distinguishes justification from rationality. He doubts the latter turn captures any well-defined or useful notion (27). His stab at a positive characterization of it — being rational is reasoning well (317–8) — indicates that he takes rationality to be an internal matter.

and explaining why it is widely rejected. That will lead to the topic of this paper, namely, whether the rejection is warranted.

Two disclaimers are in order, to set aside irrelevant issues and to clarify why the view under consideration is that being rational is only *roughly* doing what is best in view what one knows.

First, what "best" amounts to will not matter much here. Perhaps rationality is a matter of doing what is *sufficiently good* rather than what is best — satisficing rather than maximizing. Perhaps it is matter of doing what is *expectably best* rather than what is best: for there are cases in which we know that an option does not have the best result and yet it is still rational to take it because it has the best expected results (Parfit, 2011, 159–60). There are plausibly different flavours of best: personal, moral, aesthetic and so on. These will yield corresponding flavours of rationality. There are also various localized notions of best: what is best in view of a certain goal or norm alone — which may or may not be best in a wider perspective. These will yield correspondingly localized notions of rationality. On some views, what counts as "best" for the purpose of evaluating an person's rationality also varies with what that person desires, prefers, wants and so on. Most importantly for us, a person's desires and like states may also affect what *appears best* to that person, for instance by affecting how they rank alternative state of affairs. To reflect that, we may have to say that being rational is, roughly, doing what appears best in view of what one knows *and of what one desires*. The precision would have no impact on the present discussion, however.

Second, some would prefer to say that being rational is doing what is best in

3

view of what one is *in a position* to know. Others that it is doing what is best in view of what one knows *by observation*, *directly* or some similar restriction. The differences between these variants of the view are interesting but lateral to our problems.

The paper proceeds as follows. Section 1 states five considerations in favour of the knowledge view. Section 2 states one main reason for its rejection: it appears to have blatantly false consequences in cases of non-culpable error. The problem is a version of what is known as the New Evil Demon problem in epistemology. Section 3 argues that it is not yet clear whether the view has such consequences. To progress on that score, section 4 spells out a straightforward implementation of the view: Knowledge-based Decision Theory. Sections 5–6 prove that the version does have the unpalatable consequences. Section 7 introduces a less straightforward implementation of the view: the Two-Step account. While that version is in better shape, it cannot yet be said to avoid the problem altogether.

1   Knowledge and rationality

Here are five considerations in favour of the idea that rationality is, roughly, doing what is best in view of what one knows.[3] The first comes from *ordinary folk appraisals* (Hawthorne and Stanley, 2008, 571). Claims about what is rational are routinely backed up by claims about knowledge. Explicitly so when we make such

---

[3]The claim is intended to cover both what is rational *ex ante* and what is rational *ex post*. Roughly, rationality *ex post* would require not merely doing what is best in view of what one knows but doing it *in the light of* (relevant aspects of) what one knows. See section 3 below on the distinction.

claims as "It wasn't sensible for Hannah to get so much food since she didn't know how many people would come". Explicitly, but indirectly so when we use instead knowledge-entailing constructions such as "seeing that", "remembering that" and so on.[4] Implicitly so when we make such claims as "It was silly for Hannah to buy bonds since the market was about to collapse" which arguably presuppose that Hannah knew or was in position to know that the market was about to collapse. One symptom of the presupposition is that challenges such as "How would she know that? All agencies gave the bonds stellar ratings!" are perfectly natural. Of course, some appraisals are instead backed up with claims about *desires and the like*. But as said above, we are not concerned with how rationality depends on those. Some appraisals are backed up by claims about *reasons*: "Hannah had no reason to think that the bonds would not be repaid". But these may be a matter of knowledge as well, as will be seen shortly. Some are backed up by claims about *what it is reasonable to think*: "It wasn't silly for Hannah to buy the bonds; in fact, it was perfectly reasonable for her to expect that they would be repaid". But that leaves open the possibility that what is reasonable to think is itself, roughly, believing what seems true in view of what one knows. Better counterexamples are appraisals backed up by mere claims of *belief*: "it was sensible for Hannah to get a lot of food, for she thought that many people would come". However, it is at least arguable that these presuppose that the belief in question was reasonable. If Hannah's belief that many people would come was wholly irrational, one would

---

[4]See Williamson (2000, chap. 1) for the view that these entail knowledge. The view is fairly orthodox but not uncontroversial (Bernecker, 2007; Turri, 2010; Pritchard, 2012, 27–34).

feel at least reluctant to accept that it was sensible for her to get a lot of food because she thought that many people would come. If it is presupposed that the belief is reasonable, then again its being so may be a matter of what one knows.[5]

The second consideration involves a principle linking rationality to *evidence*: being rational is doing what is best in view of one's evidence. Something along those lines is commonly endorsed in decision theory, formal epistemology and philosophy of science. As we noted, many philosophers take one's evidence to consist in internal mental states (or their contents). However a case can be made for the idea that one's evidence is just what one knows (Williamson, 2000, chap. 9). And though few decision theorists or philosophers of science seem to accept the idea that *all* one knows is evidence, they typically assume that one's evidence consists in a subset of what one knows: what is known by observation, for instance.

The third consideration involves a similar principle framed in terms of *reasons*: being rational is doing what is best in view of the reasons one has. Talk of reasons is notoriously flexible. There are marginal uses of "having a reason" on which

---

[5]A few more things can be said here. First, "ought" and "should" are sometimes used to talk about what is rational for one to do. When they are, they receive the kind of back up discussed above. See Hawthorne and Stanley (2008, 571–3) for relevant examples — though note that some of their cases involve a use of "ought" on which one may reasonably ("excusably") do something what one "ought" not to do. Second, these ways of talking are not restricted to the folk. They are also widespread in disciplines dealing with rationality, such as decision theory, game theory or economics — though admittedly, "knowledge" and "information" are there often used in ways that do not sharply distinguish knowledge and belief. Third, claims of rationality are rarely backed up with claims about what one experiences unless the latter indicate a reasonable belief. For instance, suppose that Hannah knows that a pair of gold sticks are of the same length though due to some optical illusion one looks longer than the other. "It was sensible for Hannah to take the left one since it looked longer" is bizarre if not outright false in such a scenario.

a "reason one has" may be something unknown to one. Two bankers talk about Hannah; one asks: "does she have any reason to pull her money out?", the other replies: "yes, her stock's value is about to collapse". That may be felicitous and true even though Hannah has no idea of the impending collapse.[6] But that is not a use on which the principle is plausible. If Hannah pulled out now, her doing so would not be made rational by the impending collapse. That being said, central uses of "reasons one has", "one's reason" and "reasons for which" appear to require knowledge. On their salient reading, "Hannah had a reason to complain: the banker lied to her", "Hannah's reason to change bank was that the banker had lied to her", "The reason for which Hannah changed bank was that the banker lied to her" all presuppose or entail that Hannah knew that the banker lied. They are intimately related to the reading of "Hannah left her bank because her banker lied" that presupposes the same. A case can be made for the principle stated with that notion of reason (Unger, 1975).

The fourth consideration is one of *systematic unity*. The line of thought is due to Williamson (2000). For decades, epistemologists have taken justification or rationality as their central notions and tried to account for knowledge in their terms. The project has not been successful.[7] That leaves us with three alternative pictures. The first is a two-headed epistemology, where knowledge and rationality are independent notions. Knowledge would belong to a group of interrelated

---

[6]See Parfit (2011, 31–2), though see Gibbons (2010) for a contrary view.

[7]Chisholm (1977) is paradigmatic. The programme is carried out in the literature of "undefeated justification" accounts of knowledge, which boomed in the 70s but was superseded by accounts of knowledge that used wholly different primitives in the 80s, such as relevant alternatives, reliability or sensitivity.

notions — including perhaps safety, sensitivity, modal stability, causal connexions between world and mind — and rationality to another — including perhaps reason, justification, defeat, epistemic norms, reasoning, reflexive access. The second is an epistemology with one trunk and two branches: knowledge and rationality are both explained in terms one third primitive — the current contenders are reliability and virtue. The third is a knowledge-first epistemology, where rationality and justification are accounted for in terms of knowledge. Those who find the first option theoretically unsatisfactory and the second unsuccessful are pulled towards the third. In recent years, knowledge-first accounts of justification have been actively developed.[8] Knowledge-first accounts of rationality are the obvious next step. The knowledge view — the idea that rationality is, roughly, doing what is best in view of what one knows — is not the only way to go this route.

The fifth consideration has to do with *how to apply decision theory in particular cases* (Weatherson, 2012, 77-82). We leave it aside for the moment but return to it in section 4.

Each consideration is debatable. The point here is not to establish the knowledge view conclusively. It is, after all, a controversial view. The point is rather to show why it would be very natural to consider it. That makes more striking the fact that, with few exceptions, theorists hardly give it a single thought.

---

[8]Williamson (2000), Sutton (2007), Haddock (2010).

2 The New Evil Demon problem

One major reason for theorists to dismiss the knowledge view is that it appears incompatible with some obvious facts about rationality. The facts in question have to do with what epistemologists call "the New Evil Demon problem" (Lehrer and Cohen, 1983, Cohen, 1984; the name appears in Sosa, 1991, 135).

Consider a pair of cases. In the Good case, Sarah remembers — and thereby knows — that her car is parked in front of the office. Her workday is over and she decides, reasonably, to walk towards that spot. In the Bad case, everything seems to Sarah as it does in the Good case but her car has been stolen. She does not know that her car is parked outside — it is not —, but she takes herself to do so. She decides to walk towards the same spot.[9]

Sarah's decision is rational in the Bad case just if it is rational in the Good case. That is obvious or at any rate extremely hard to deny. As Wedgwood (2002, 349) puts it, "when the word "rational" is used in the way that is most common among philosophers, [that] intuition is compelling".

The intuition appears incompatible with the knowledge view. What it is rational for Sarah to do in the Good and the Bad cases is the same. Yet her knowledge differs from one case to another. That does not entail, but strongly suggests, that rationality is not a function of knowledge. Rather, rationality appears to be a function of what remains constant across the cases: how things look to Sarah or Sarah's internal mental states.

---

[9]The Good case / Bad case terminology comes from Williamson (2000, 150).

The version of the problem familiar among epistemologists is more dramatic. In that version, the Good case is an ordinary case of knowledge, but the Bad case is one of massive deception: Sarah is the victim of a Cartesian Evil Demon, or she is a brain in vat stimulated to have experiences similar to the ones she would have in the Good case. Again, the claim is that her decision is rational in both cases. (As the case is more outlandish, the claim is perhaps less unassailable.) Since the cases differ drastically in terms of what is know, rationality does not seem to be a function of knowledge.

The Old Evil Demon problem was for certain views to explain how one *knows* in the Good case even though one does not in the Bad case. The New Evil Problem is for certain views to explain how one as *rational* and *justified* in the Bad case as one is in the Good case despite lacking relevant knowledge and failing in various other ways.[10]

While the New Evil Demon intuition is robust about many pairs of cases, it is not trivial to find a suitable generalization of it.[11] We call *case* a centred world, that is a possible world with a designated subject and time. When talking of a case we refer to its subject as *one* or *the agent* and to its time by using tenseless or present-tensed constructions. *Conditions* such as being seated, raising one's arm, believing that there are rocks obtain at cases.[12] We predicate *rational* of some conditions at cases: it is rational to intend to sell one's stock in some given case.

---

[10]In the dramatic scenario, not only Sarah knows little, but she also forms her beliefs in ways that are unreliable relative to the environment she is in. That is why epistemologists — whose originally targeted reliabilism about justification — focussed on the dramatic version.

[11]Thanks to ***Timothy Williamson for pressing this point.

[12]The terminology comes from Williamson (2000, 52).

Now we may initially try to generalize the intuition as follows:

> *NED-generalization — physical.* If the agents of two cases are physical duplicates, then something is rational in one if and only if it is rational in the other.

The generalization is mistaken. First, it captures too few cases. It is not essential to the pairs that they involve physical duplicates: some involve a healthy human being on one side and a brain in a vat on the other. The problem is not fixed by replacing "physical duplicates" by "cerebral duplicates": the intuition does not require that the two subjects are atom-per-atom duplicates. Second, it fails insofar as the contents of one's mental states do not supervene on the physical. It is rational for me to believe that water is wet, but it is not rational for my physical duplicate on Twin Earth to believe that water is wet. Similar problems arise if we replace "physical" by "intrinsic".

A second attempt is to say that *mental states* duplicates are rational duplicates. That does not capture the intended cases if knowledge is among the mental states (Williamson, 2000, chap. 1). For in the Good case one knows while in the Bad case one does not. The idea should rather be that duplicates in *internal* or *non-factive mental* states are rational duplicates. But again, that will capture too few cases. First, their are pairs of cases that differ on *irrelevant* mental states: the verdicts on any of the foregoing pairs are unchanged if we make one case so that one believes that the number of stars is even and the other so that one believes that it is odd. Second, many pairs involve different though *analogous* mental states that rationalize different but *analogous* beliefs or intentions. In the Good case I

11

see a dog d jumping around — it is my first encounter with it. In the Bad case I have a corresponding hallucination. On a standard view I am able to entertain thoughts about d in the Good case but not in the Bad case. Hence in the Good case it visually seems to me that d jumps around; not so in the Bad case. In the Good case it is rational for to believe that d jumps around; not so in the Bad case. Still, the intuition goes, in the Bad case it is rational for me to have an analogue belief (or belief-like state) that I would express by saying "that dog is jumping around".[13] And in both cases it is rational for me to believe that there is a dog jumping around. A NED-generalization framed in terms of duplicates is too stringent to capture such cases.

We can better recast the NED generalization with the help of two primitives. The first is *relevance* (to rationality): in a given case only some of one's mental states are relevant to whether a given belief or intention is rational. In the parked car cases Sarah's beliefs about the number of stars are irrelevant, for instance. The second is that of *counterparts* of internal mental states. My belief that water is wet and my Twin Earth counterpart's belief that twater is wet are counterparts. My belief that d is jumping around and my belief (or belief-like state) that "that" is a dog in the Bad case are internal counterparts. Similarly, my intention to catch d in the Good case may have an intention (or intention-like state) to catch "that dog" in the Bad case. We do not provide necessary and sufficient conditions for counterparthood, but the notion has a fairly intuitive application in the intended

---

[13]The state is not indisputably a belief since (on Russellian accounts) there is no singular proposition to be believed here.

12

cases.

> *NED-generalization — internal counterparts.* For every pair of be-
> liefs or intentions F, G and every pair of cases a, b, if F and the F-
> relevant mental states in a are counterparts of G and the $G-$relevant
> mental states in b, then it is rational to F in a iff it is rational to G in
> b.

Summing up, overwhelmingly plausible facts about NED pairs appear incompat-
ible with the view that rationality is, roughly, doing the best in view of what one
knows. In the NED pairs two cases are identical in some respect of rationality
even though they differ in relevant respects of knowledge. That suggests that ra-
tionality is not a function of knowledge. This is one of the main reasons why the
knowledge view has long been left off the table (Wedgwood, 2002).

3 The loophole

There is a loophole, however. Suppose that one's pay is a function of one's job; it
is still possible for two different jobs to pay the same. Similarly, rationality may
be a function of knowledge even though some people differing in knowledge do
not differ in rationality. The knowledge view can advert the New Evil Demon
problem if it is shown that what is known in the Bad case, though different from
the Good case, makes the same choices rational.

Lord (2013) pursues just such a strategy. On his view, whenever one knows p
in the Good case, one knows that *it appears that* p in the (dramatic version of the)

Bad case.[14] He argues that that knowledge plays an equivalent role in rationalizing one's actions and choices. In the Good case, it is rational for Sarah to walk towards the parking spot because she knows that the car is there. In the Bad case, it is also rational to do it, but because she knows that the car *appears to her* to be there.

At this point we need to distinguish two concepts of rationality: *ex ante* and *ex post*. What is *ex ante* rational is what it is rational for one to do, irrespectively of whether one does it or not and of why one does it. What is *ex post* rational is what one does rationally; that requires that one did it and depends on why one did it.[15] The two come apart when one does the reasonable thing for wrong reasons. Suppose it is reasonable for Sarah to buy health insurance but that she does so only because she has the pathological conviction that she caught a deadly disease. Is her action rational? Yes and no. It is (*ex ante*) rational for her to buy health insurance. But she did it (*ex post*) irrationally. The common wisdom is that rationality *ex post* requires doing the *ex ante* rational thing on a proper basis, that is, doing it in the light of what makes it *ex ante* rational.

As Lord (2013) stresses, the distinction raises a double New Evil Demon problem for the knowledge view.[16] The *primary* New Evil Demon problem is to explain why one is *ex ante* rational in the Bad case. The strategy here is to find pieces of knowledge in the Bad case that make the relevant action rational. Suppose that is done. Still, there is a *secondary* New Evil Demon problem, which is

---

[14]The proposal develops a suggestion made by Williamson (2000, 199–200).

[15]That is an analogue of the epistemologists' distinction between "propositional justification" and "doxastic justification". The labels "ex ante" and "ex post" are used by Goldman (1979).

[16]Not so for reliabilists, for instance.

to explain how can one act *ex post* rationally in a Bad case. For even if there is some knowledge in the Bad case that makes one's act *ex ante* rational, one does not appear to act on it. Rather, one acts on the basis of a state that does not constitute knowledge, *e.g.* a false belief. So we would have to say, implausibly, that actions in Bad cases are always ex post irrational. (And similarly for intentions, beliefs and other states.)

In broad outline, Lord's answer to the secondary problem is this. In the Bad cases in which one is *ex post* rational, one is sufficiently sensitive to the relevant pieces of knowledge for her action, intention or belief to count as based on them. Of course the action, belief or intention is also and more saliently based on a state that does not constitute knowledge, such as false belief. But that does not prevent it to be based on items of knowledge that rationalize it. The answer seems to me on the right track. At any rate, nothing will be added to it here. Our main focus is the primary problem, which raises more pressing concerns.

4 Knowledge-based decision theory

To see whether the loophole can be exploited, I focus on a more specific version of the knowledge view. I call it *knowledge-based decision theory*. It has been suggested, if not endorsed, by (Hawthorne and Stanley, 2008, 577–80) and endorsed by Weatherson (2012, 77–82).[17]

---

[17]Showing that Hawthorne and Stanley (2008) endorse the view requires a bit of exegetical work. They explicitly present their view as an alternative to dominant theories of *rationality* (571, 577 and other places). Yet when they state their view they do not use "rational" and cognates but instead "ought", "should", "appropriate" and "excusable". While they call the norm "treat p as a

A *choice situation — choice*, for short — is a case and a set of mutually exclusive and exhaustive *options* between which one must choose. In any concrete case we typically face several concurrent choices. Suppose Hannah stops at a coffee shop on her way to the bank: she faces the choice of what drink to have *and* the choice of whether to carry on with the plan to go to the bank. Typically we only give thought to one choice at a time, but that does not mean that we do not concurrently face several. There is a worry that concurrent choices are hardly ever independent: what to choose on one is always at least dimly relevant to what to choose on another. If that is so, any concrete case really is just one big choice between a myriad of highly specific options. Be that as it may, we will not worry much about the individuation of choice situations here. We assume that in the cases we will discuss, the individuation we use is at least an acceptable approximation of the proper one.

---

reason for acting only if you know p" a "norm of practical rationality" (577), it is unclear whether failing to comply with the norm entails being in any way irrational. This depends on whether "excusable" violations are always rational; but they stay silent on this — see 573, 577, 582, 586. Whichever way they go on this, they must endorse the idea that practical rationality is a function of the relevant types of appropriateness and excusability. If not, it is hard to see how their proposal is a theory of rationality. Since they think that appropriateness and excusability in the relevant sense are a function of what one knows, they endorse the generic knowledge view. More specifically, they take themselves to "gesture at" a type of decision theory in which knowledge delivers epistemic probabilities (580). It is very natural to assume that that knowledge is also meant to deliver epistemic possibilities, as in Weatherson's more explicit proposal.

The same page (580) misleads some readers into thinking that Hawthorne and Stanley distance themselves for the proposed decision theory. They write that while "[they] are by no means opposed to a perspective in which claims of practical rationality [...] are grounded in a decision theory of the sort [they] gestured at", "the need to integrate such a theory with reasons for action is still vital". The point of this, however, is not to distance themselves from the proposal; rather, it is to stress that the proposal only covers rationality *ex ante*. That is made clear in the next sentence: "For one thing, there are cases where one does one ought to do but for the wrong reasons, and this phenomenon needs explanation". The idea is not that knowledge-based decision theory is false; it is that it fails to cover rationality *ex post*.

A *decision table* is an abstract object consisting of a set of options, a set of states — and sometimes a probability distribution over them —, a set of outcomes that are functions of possibilities and options and a ranking of outcomes. *A decision theory* is a set of principles that selects options on decision tables. We assume that the selection sorts out options at least into acceptable and unacceptable ones. We say that a decision table *adequately represents what is rational* in a given choice just if what is acceptable on the table matches what is rational in the choice. We sometimes write simply that a table "adequately represents a choice".

We say that a table *reflects one's knowledge in a choice* if and only if the table's options match those of the choice and (a) all states on the table are compatible with what one knows in that case and (b) every possibility compatible with what one knows obtains in at least one state. If we represent one's knowledge as a set of possible worlds — the ones compatible with what one knows —, that amounts to saying that the table's states are a partition of one's knowledge.

Knowledge-based decision theory is the following view (compare Weatherson, 2012, 80–3):

(KBDT) If a decision table reflects one's knowledge in a choice, then that table is a rationality-table for that choice.

Some necessary refinements are left out.[18] The view straightforwardly imple-

---

[18]Not every way of partitioning one's knowledge will do. A gang of street kids threaten to scratch your car unless you pay them. If you divide the options compatible with what you know into *they will scratch the car* and *they will not* then paying will be dominated and hence deemed irrational no matter what. The table is wrong, however, because it fails to take into account the fact that your paying makes the former state less likely to obtain. A suitable partition should use states that are evidentially — on some views — or causally — on others — independent of your options.

ments the idea that it is rational to do what is best in view of what one knows. Decision-theoretic principles tell what options are "best" in view of the information provided in a decision table. When the information matches what one knows, the selected options are the best in view of what one knows. On the knowledge view, these are just the rational options.

The view is motivated by cases like the following (Weatherson, 2012, 79). Lucy chooses to put $1000 token down on 28 on a fair roulette with extremely poor returns: she merely wins double if her number comes out but looses if any of the 37 other numbers comes out. She is irrationally convinced that 28 will win, and in fact it will. We could represent her choice with a table like the following:

	28 comes out
put the token down	$2000
do not put it down	$0

On that table the selected option is to put the token down. The table may be an adequate representation of what Lucy *will do*, since it reflects her beliefs. It may also be an adequate representation of what option is *in fact best*, since it reflects the facts. But it does not adequately represent what is *rational* for her to do. For it is clearly irrational for Lucy to put the token down. On the knowledge view this is because she does not know that 28 will come out. In view of her knowledge, it is possible that other numbers come out and there is only $1/38$ chance that 28 comes out. The following table reflects her knowledge:

	28 comes out (1/38)	another number does (37/38)
put the token down	$2000	−$1000
do not put it down	$0	$0

On standard decision-theoretic assumptions putting the token down is not selected. Knowledge-based decision theory thus predicts that it is irrational to do so, which seems correct.

Weatherson (2012) further argues that of all the relations that one may bear to the information encoded in a decision table — such as knowledge, justified true belief, knowing that one knows, and so on —, knowledge is the most natural one that correlates with tables that represent rational choice adequately. That is the fifth consideration in favour of the knowledge view. It tells more particularly in favour of its implementation as knowledge-based decision theory.[19]

5 A proof that knowledge-based decision theory violates the NED intuition

Knowledge-based decision theory is well-motivated and simple. But it can be shown to violate the New Evil Demon intuition. Consider the following pair of cases. You are home and you need to grab a bottle of water to go on a bike trip. In the Good case, you know that:

---

[19]Surprisingly, Weatherson (2012) does not consider justified belief (in a sense of justified distinct from knowledge) or reasonable belief. The following suggests that they would have been better candidates.

b: there is a full bottle of water in the basement.

g: there is a full bottle of water in the garage.

Further details are as one would expect. You know that there is no other bottle of water in the house. You know that if there is a bottle in the basement and you go there, you will have a bottle; and that if there is a bottle in the garage and you go there, you will have a bottle. You know that it is good to have a full bottle — no matter which — and bad to have one. You also know that going down to the basement or the garage is a negligible cost. We claim that:

(1) In the Good case, the following table reflects your knowledge:

	$b\&g$
Go the basement	1
Go to the garage	1

In the Bad case, everything is as in the Good case, except that under exceptional circumstances, your partner has taken the bottle in the garage without warning you. So in the Bad case, you do not know that g. We claim that:

(2) In the Bad case, the following table reflects your knowledge:

	$b\&g$	$b\&\neg g$
Go the basement	1	1
Go to the garage	1	0

We further assume that our decision theory contains two principles. Let us say that an option is *unsurpassed* on a table iff its pay-offs are at least as good as any alternative at all states. For instance, in the first table above, the two actions

are unsurpassed; in the second, the first action is. And let us say that an option is *weakly dominated* iff there is some option whose pay-offs are at least as good as any alternative at all states and even better for one or more states. For instance, in the second table, the second action is weakly dominated. Now our two principles are:

(3) *Unsurpassed options* are acceptable.

(4) *Weakly dominated options* are unacceptable.

Both principles are uncontroversial dominance principles. Dominance principles do not apply to tables whose states are causally or evidentially dependent on one's actions. But nothing of the sort is going on here.

Finally, the following seems correct:

(5) It is rational to decide to go to the garage in the Good case if and only if it is rational to decide to go to the garage in the Bad case.

The claim is, of course, an instance of a New Evil Demon intuition.

(1)–(5) are inconsistent with (KBDT). By the decision-theoretic principles (3)–(4), going to the garage is selected on the first table but not on the second. By (1) and (2) the first table reflects what one knows in the Good and Bad cases, respectively. Hence by (KBDT) it is only rational to decide to go to the garage in the Good case, which contradicts (5).

Hence knowledge-based decision theory violates the New Evil Demon intuition. The result is not surprising but it is worth establishing. The argument can

be used against one way of spelling out Lord's view, for instance. Even if we add *it appears that g* in all columns of the Bad case table, going to basement still dominates going to the garage. So the view makes the wrong prediction that going to the garage is irrational in the Bad case. Hence if Lord's view is spelled out along the lines of Knowledge-based decision theory, it fails to avoid the New Evil Demon problem.

6   Objections and replies

A few attempts at blocking the proof are worth considering.

Attempt 1: scepticism. One may flat-out reject the claim that you know that b and g in the Good case, arguing that nothing is known of the external world.

Answer. External-world scepticism easily reconciles knowledge-based accounts of rationality and the New Evil Demon world intuition. It does so by holding that in Good cases one cannot know things that are not know in Bad cases, thereby removing asymmetries in knowledge between the cases. The difficulties for scepticism lie elsewhere: it is highly implausible, and when paired with a knowledge-based account of rationality, it has a hard time explaining why certain choices are rational (see Unger, 1975, chap. 4, who argues that sceptics must claim that none is). A discussion of scepticism is beyond our scope, however. The knowledge views we are concerned with are embedded in a resolutely non-sceptical outlook.

Attempt 2: in the Bad case, you do not know that b. One may try to argue that

in the Bad case, you lack knowledge of b as well as g. *Contra* (4), the proper table for Bad would thus be along the following lines:

	$b\&g$	$b\&\neg g$	$\neg b\&g$	$\neg b\&\neg g$
Go the basement	1	1	0	0
Go to the garage	1	0	1	0

Going to the garage is not weakly dominated on the table, so the proof fails.

Answer. The strategy is untenable for some variants of the case. Suppose you are currently standing in the basement, staring at the full bottle of water. The unusual happenings in the garage can hardly deprive you of your knowledge that there is water in the basement.

Attempt 3. Either you do not know b in the Bad case or it is irrational to go to the garage in the Good one. The attempt uses the previous one as part of a divide and conquer strategy. Variants of the pair are divided in symmetric and asymmetric ones. Symmetric ones are those in which b and g stand and fall together. For instance, you learnt about both bottles from the same source. Since the source is misleading in the Bad case, you fail to know both b and g in the Bad case. These case are dealt with along the lines of Attempt 2. In the other cases, it is now said, your standing to b is better. After all, your knowledge of b is retained in the Bad case. Because your standing to b is better, it is somewhat irrational to go the garage in the Good case. So it is not a problem that we classify it as irrational in the Bad case as well.

Answer. First, the decision theory needs revision in order to integrate "quality of knowledge" considerations. That is not trivial. Second, it is dubious that when-

ever one's standing to b is slightly better than one's standing to g, it is somewhat irrational to opt for the garage. Third and mainly, it is simply false that in asymmetric cases, one's standing to b is better. Consider the following. In the Good case, you are standing in the garage and looking at the bottle. You only know of the basement bottle by a fairly distant memory. In the Bad case, highly unusual circumstances have forced your partner to refill the water bottle you are looking at with petrol. The case is asymmetric, in the sense that the unusual happenings in the garage do not affect your knowledge of the basement. But it is not such in the Good case your standing to b is intuitively better than your standing to g in a way that would make it somewhat irrational to opt for the garage. If anything, your standing would make it somewhat irrational to opt for the one in the basement.

Attempt 4: in the Bad case, you do not know that the basement is at least as good as the garage. In the Good case, you know that the garage is no better and no worse than the basement. In the Bad case, you loose knowledge that the garage is no worse — for in that case it *is* worse. Perhaps you also loose knowledge that it is no *better*. You may fail to know that there is not more water or a fancier drink in the garage. If so, the proper table for Bad is something like:

	$b\&g\&r$	$b\&\neg g\&r$	$b\&\neg g\&\neg r$
Go the basement	1	1	1
Go to the garage	1	0	2

Where r stands for the proposition that the garage is no better than the basement.

Answer. Again, that is highly implausible for some versions of the case. Un-

usual happenings deprive you of some knowledge of the garage's contents, but not all of it. You retain knowledge that it does not contain a dozen elephants, for instance. You may equally well retain knowledge that the garage has at most one bottle of water.

Attempt 5. Zero Probability. Since knowledge entails truth, actuality is always compatible with what one knows. In the Bad case, this forces us to introduce a state that reasonably believes not to obtain: $b\&\neg g$. But everything else the agent knows (misleadingly) indicates that it does not obtain. So we could argue that the state receives a probability 0. And since the state has a zero probability, the fact that going to the basement is better in that state should not be taken into consideration. Hence we reject weak dominance (4): weakly dominated options are acceptable if they are only dominated in virtue of zero-probability states.

Answer. First, one has to motivate the probability function. If it is obtained by conditionalizing on knowledge, for instance, it requires a prior probability of 1 to *there is a bottle* conditional on *there appears to be a bottle*. That may appear implausible — especially if similar assignments must be made for all NED-style pairs. Second, the strategy faces a dilemma. Either it treats probability 0 as equivalent to leaving a state off the table or it does not. If the former, the strategy in effect gives up knowledge-based decision theory. If the latter, weak dominance applies.

Consider the latter horn first. Probability 0 and impossibility are thought to come apart in cases involving infinite sets of possibilities. A point-sized dart is randomly thrown on a disc; it has equal chance of falling on any point. The chance

25

that it hits the point-size bull's eye cannot be a real number greater than 0 — for the chances of each point is the same and they would add up to more than 1. So it must be 0. Still, it is possible that the dart falls on the bull's eye. Hence probability 0 is not impossibility; rather, it indicates that the ratio of positive cases to all cases is infinitely small. When probability 0 and impossibility are thus distinguished, weak dominance appears correct. It is not rational to prefer a bet that pays $1000 unless the dart falls on bull's eye to a straight offer of $1000. (If you doubt that, consider instead a huge but finite number of points scattered on the disc, say a trillion. It is not rational to prefer a bet that pays $1000 unless the dart falls on one of those to a straight offer of $1000. Yet the probability that the dart fall on one of those is still 0.)

To this, one may be tempted to reply that the infinite case and the Bad case differ. In the infinite case, it may be said, the probability is not 0 but infinitesimal — where infinitesimal numbers are (non-real) numbers whose infinite sums can add up to a finite number. By contrast, in the Bad case the true state should receive a proper probability 0. And here dominance does not apply.

If one goes this route, then for all intents and purposes one treats probability 0 as impossibility — that is, as equivalent to leaving a state off the table. But that is tantamount to reject knowledge-based decision theory: the true state is left out even though it is compatible with what one knows.

Attempt 6. Reject the decision principles. Finally one could object to the dominance principles (3) and (4), arguing that they fail in some cases.

Answer. The cases in which dominance principle appear to fail are well-

delimited: when the states are causally or evidentially dependent on one's choices or when reiteratively eliminating dominated strategies in game settings. Our case involves none of those. Those applications aside, the principles are perhaps the less controversial ones of decision theory. Rejecting them is a high cost.

Attempt 7: restrict the knowledge base. One may endorse a version of the knowledge view on which rationality depends on a *subset* of what one knows: what is known by observation, "directly" or some such restriction. The subset may suitably be called one's evidence. (KBDT) is revised accordingly: a table adequately represents rationality in a choice if it reflects one's *evidence* in that choice. Moreover, one's evidence in the same in both cases. Hence what is rational is the same.[20]

Answer. If evidence includes things observed, it is easy to produce pairs on which a structurally similar proof goes through. In the Good case, one may observe both that there is a bottle of water on the bar (b') and one on the ground (g'). In the Bad case, the ground bottle has been filled with petrol or one hallucinates it. So one does not have evidence g' but one still has evidence b'. The rest goes as before. Now the more evidence is restricted, the more contentious the pairs are. If evidence is restricted to propositions about one's presently occurring narrow mental states, it becomes less obvious that one can be mistaken about them — hence a convincing Bad case is harder to find. But with evidence so restricted, the

---

[20]We can say more. Since one's evidence is compatible with $b\&\neg g$ in the Bad case, it must be so as well in the Good case; since going to the basement does not dominate in the Good case, one's evidence must be compatible with $\neg b$. Hence the table for both cases should like the one in Attempt 2.

view, like scepticism, makes it harder time to explain why it is rational to act.

7   The Two-step account

Knowledge-based decision theory is not the only way to spell out the idea that rationality is, roughly, doing the best in view of what one knows. In this section, an alternative is sketched, the *Two-Step* account. Like Knowledge-based Decision Theory, it makes what is rational a function of what is known. But unlike it, it does not directly uses one's knowledge as input to decision theory. Rather, it proceeds in two steps. First, one's knowledge determines *what is supported by* what one knows. Second, what is supported in turns determines what is rational to do. For belief the second step is straightforward: it is rational to believe just what is supported by what one knows. For intentions and actions, what is rational is what is justified in view of what is supported by what one knows.

If what is supported by what one knows just *is* what one knows, the view boils down to knowledge-based decision theory. However, if what is supported is *more* than what one knows, the view has some hope of solving the New Evil Demon problem. The suggestion would be that what is supported by one's knowledge in the Bad case is the same as, or a counterpart of, what one knowledge supports in the Good case. In the basement / garage Bad case, for instance, even though you do not know that there is a water in the garage, what you know supports that there is one. Hence the one-column table of premise (1) would reflect *what is supported* in *both* cases. If rational choice is a matter of *what is supported* rather

than of what one knows, it will be the same across both cases.

On the proposed view what is rational is still a function of what is known, for it is a function of what is supported, and what is supported is itself a function of what is known. If what is supported is the same (or analogous) across Good and Bad cases, the New Evil Demon problem is averted.

It is useful to contrast the proposed view with Lord's (2013). On Lord's view what rationalizes actions are known facts; they differ across Good and Bad cases. On the present view what rationalizes actions are supported propositions; they are the same across Good and Bad cases. On Lord's view, what ultimately rationalizes action in the Bad case is a known fact about appearances. On the present view, it is whatever knowledge supports the relevant propositions. Lord's view requires a psychological state of appearing that p; the present view only requires that p is supported by what one knows. Note, however, that the proposed view cannot plausibly wholly avoid appeal to known appearances. In the dramatic Bad cases, one's relevant knowledge is limited to propositions such as *it appears to me that p* or *the present situation is like one in which p holds* (Williamson, 2000, 199). If p is supported, it will have to be supported by such propositions. The crucial difference between knowledge-based decision theory and the two-step view, however, is that in the former only *it appears that p* makes it way into decision tables, while in the latter p itself does. What Lord says about his own view does not adjudicate between the two options.

One could call what is supported "the evident" — with an eye towards Chisholm's (1977) use of the term. The evident is what is made evident by one's evidence,

namely what one knows. What one knows itself is evident, but the evident is not restricted to what one knows — see below. The term strains ordinary usage, however, because some things that are "evident" in that sense are false.

If we call what is supported *one's reasons*, the proposed view would yield the result that one's reasons are the same (or analogous) in the Good and Bad cases. However, that would require that in some cases one's reasons are false propositions. That strains ordinary usage.

Let us call what is supported *the scaffolding*. We can list some properties that the scaffolding must have for the view to succeed.

> *The scaffolding supervenes on knowledge.* What is supported in a case is a function of what is known in that case.
>
> *The scaffolding includes knowledge (reflexivity).* If one knows p, p is supported.[21]
>
> *The scaffolding goes beyond knowledge (ampliativity).* For some p, one does not know that p (and p does not follow from what one knows) but p is supported. For instance, in our Bad case it is supported that there is a bottle in the garage, though one does not know that.
>
> *The scaffolding is indistinguishable from knowledge.* If p is supported, one does not know that one does not know p. Equivalently,

---

[21]Thus the view entails that if one knows p, it is reasonable to believe that p. That has been challenged by Lasonen-Aarnio (2010).

if one knows that one does not know p, then p is not supported.[22]

The scaffolding is less than the "unknown unknowns". For some p, one does not know that one does not know p, but p is not supported. For instance, if p is a proposition that one cannot even conceive, one does not know that one does not know p. Yet p is not supported.[23]

The first four properties can be summed up as follows: the scaffolding lies between the known the "unknown unknowns" but cannot be identified with either.[24] Finally, the last property is:

New Evil Demon compatibility. The scaffoldings in Good cases / Bad cases pairs are counterparts of each other.[25]

Some further properties that are not mandatory but desirable are *consistency* (the scaffolding never entails both p and not-p), *deductive closure* (the scaffolding is closed under logical entailment), *non-monotonicity* or *defeasibility* (strict increases in what is known need not preserve what is supported). To illustrate the latter: if in the Bad case you learned that the neighbour asked your partner some water in emergency, the proposition that there is a bottle of water in the garage

---

[22]That entails that if one knows that one does not know p, it is not reasonable to believe that p. That is challenged by those who think that it is reasonable to believe that a lottery ticket will loose even though one does not know it.

[23]I *do* know that *I don't know facts that I cannot conceive*. But that does not make it the case that I know that *I don't know p*, where p is a particular fact that I cannot conceive.

[24]"Between" makes sense because the known is necessarily a subset of the "unknown unknowns": if you know p, it is not the case that you do not know p, so you do not know that you do not know it.

[25]Strictly speaking, our counterpart relation applies to internal mental states. We may extend it to propositions by saying that p in case a is a counterpart of q in case b if and only if a belief that p in case a would be counterpart of a belief that q in case b.

would cease to be supported. Non-monotonicity is required if the scaffolding function is reflexive, ampliative and consistent.[26] A further property is *cut:* what supported propositions would support if known is itself supported.[27]

The notion of support generalizes to *supported probabilities*. Some associate an *evidential* (or *epistemic*) *probabilities* with what one knows (Williamson, 2000, chap. 10). Supported probabilities would be a function of those. The simplest proposal is that supported probabilities are obtained by conditionalizing evidential ones on supported propositions. Supported probabilities — rather than evidential ones — would then be fed in decision tables that represent what it is rational for one to choose in a case.[28]

It is far from trivial to *build* a scaffolding function that has the required properties. But that does not show that there is none. One simple refutation one would

---

[26]By ampliativity, there is a body of knowledge X that does not entail p but that supports p. By reflexivity, $\neg p$ is apparent on $X \cup \neg p$. By consistency, p is not supported by $X \cup \neg p$. So a strict increase in knowledge fails to preserve a supported proposition.

[27]The idea is this: suppose that a body of knowledge, K, supports p. And suppose that taken as a body of knowledge, K and p together would support q. Then q is supported by K itself. That may seem intuitive. Some word of warnings are in order, however. Cut is notoriously violated in probabilistic settings — it may be that given p, it is likely that q, and given p and q, it is likely that r, and yet given p alone, it is not likely that r. And, as Weisberg (2010) argues, failures of Cut seem precisely to be what appears to go wrong in "bootstrapping" reasoning. So one may reject Cut if one invokes some probabilistic machinery to account for the scaffolding function and one may need to reject it anyway to classify "bootstrapping" reasoning as irrational. We do not explore these issues here.

[28]On the view just sketched, a lot of things get (apparent) probability 1: everything you know and then some more. Many philosophers think that it is irrational to assign probability 1 to anything except a highly restricted group of propositions. They are worried, first, that assigning probability 1 to a proposition means that one will be disposed to bet on it at any odds. And, second, that assigning probability 1 to a proposition means that one is disposed to remain certain of it whatever evidence comes. With respect to second problem, it should be clear from the above that given the non-monotonicity of supported propositions, increases in one's knowledge can demote a supported probability of 1. The first is beyond the scope of the present paper. See Hawthorne and Stanley (2008, 589) for some discussion.

be to exhibit a pair of cases which are identical in knowledge but differ in rational belief, in violation of the supervenience claim. But none seems forthcoming. Furthermore, as long as the New Evil Demon property is not well specified, it is difficult to prove that no such functions exists — just as it is difficult to prove that one does. The jury is still out.

8 Conclusion

The idea that rationality is, roughly, doing the best in view of what one knows has some intuitive and appealing features. But it appears to fly in the face of the New Evil Demon intuition. The intuition is that rationality is similar across cases that differ in knowledge but that are suitably similar in internal mental states. That suggests that rationality is not a function of knowledge. It does not entail it, however. For it is in principle possible that different bodies of knowledge rationalize the same thing.

The most natural way to spell out the idea more rigorously is *knowledge-based decision theory*. That is the view that what is rational to do is determined on a table whose states are a partition of one's knowledge. We have shown that the view does violate the New Evil Demon intuition. To many, that definitely disqualifies it as a theory of rationality.

That is not the end of the knowledge view, however. For one may retain the idea that what is rational is a function of what one knows without plugging knowledge directly onto decision tables. The *Two–Step* view we introduced does exactly

this. On that view, what one knows determines what is supported in one's case, and the latter is what gets onto decision tables. The view avoids the problem of the simple knowledge-based decision theory approach. We have delineated the properties that a notion of the supported must have for the view to work. We have not built one, however, and it remains to be seen whether a suitable notion exists.

References

Bernecker, S. (2007). Remembering without knowing. *Australasian Journal of Philosophy*, 85:137–56.

Chisholm, R. M. (1977). *Theory of Knowledge*. Prentice Hall, Englewood Cliffs, NJ.

Cohen, S. (1984). Justification and truth. *Philosophical Studies*, 46(3):279–295.

Gibbons, J. (2010). Things that make things reasonable. *Philosophy and Phenomenological Research*, 81(2):335–61.

Goldman, A. I. (1979). What is justified belief? In Pappas, G. S., editor, *Justification and Knowledge*, pages 1–23. Reidel, Dordrecht.

Goldman, A. I. (1986). *Epistemology and Cognition*. Harvard University Press.

Haddock, A. (2010). Knowledge and action. In Haddock, A., Millar, A., and Pritchard, D., editors, *The Nature and Value of Knowledge: Three Investigations*, pages 191–259. Oxford University Press.

Hawthorne, J. and Stanley, J. (2008). Knowledge and action. *Journal of Philosophy*, 105(10):571–590.

Lasonen-Aarnio, M. (2010). Unreasonable knowledge. *Philosophical Perspectives*, 24(1):1–21.

Lehrer, K. and Cohen, S. (1983). Justification, truth, and coherence. *Synthese*, 55(2):191–207.

Lord, E. (2013). Defeating the externalist's demons. draft of june 2013.

Parfit, D. (2011). *On What Matters*, volume 1. Oxford University Press.

Pritchard, D. (2012). *Epistemological Disjunctivism*. Oxford University Press.

Sosa, E. (1991). *Knowledge in Perspective*. Cambridge University Press.

Sutton, J. (2007). *Without Justification*. MIT Press.

Turri, J. (2010). Does perceiving entail knowing? *Theoria*, 76:197–206.

Unger, P. (1975). *Ignorance : a Case for Scepticism*. Oxford University Press.

Weatherson, B. (2012). Knowledge, bets and interests. In Brown, J. and Gerkken, M., editors, *Knowledge Ascriptions*. Oxford University Press.

Wedgwood, R. (2002). Internalism explained. *Philosophy and Phenomenological Research*, 65(2):349–369.

Weisberg, J. (2010). Bootstrapping in general. *Philosophy and Phenomenological Research*, 81(3):525–548.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press.