Unstable preference, unstable deontic modals

February 26, 2016

Abstract

I defend the thesis that deontic modals and preferences are unstable. I start by offering a conceptual argument for instability of deontic modal beliefs based on the connection between instability and the contextual presence of distinct outcomes. Updated contexts can make available information that guarantees new outcomes, and insofar as these outcomes are differently valued than the ones guaranteed by our initially preferred choice, it becomes appealing to switch to the new choices (that give rise to the new outcomes). I present several ways in which deontic modal reasoning is unstable, and their corresponding definitions of instability. I then give an empirical argument for instability, and address some methodological challenges. I conclude by showing that recent defences of stability can be more appropriately reconstructed as arguing for a limited version of stability (stepwise stability), which is compatible with my instability thesis.

Consider the similarities between universally quantified statements (a-b) and sentences containing deontic necessity modals (b-c).

(1) *The quantified statement (b) follows from (a), but not vice-versa. Likewise for the modal statements (d) and (c)*

a. All students in the university are young males.
b. All students in the university are males.
c. You should go to the university and talk to the students.
d. You should talk to the students.

The inferences from subsets (e.g. young males) to supersets (e.g. males) are called superset-inferences or upward monotone inferences. The above monotone inferences seem to be correct, but this may be an isolated case. It is worth asking under what conditions monotone inferences hold. Monotonicity permits a systematic characterisation of natural language sentences containing operators such as the universal *all*, negation *not* and many others.[1] Since modal expressions like *should* can be thought of as special kinds of universal quantifiers (namely, as expressions talking about all the possibilities), modals may also be characterisable systematically in terms of monotonicity. I will be concerned here with the monotonicity of deontic modals as a constraint on linguistic competence and common reasoning.

As speakers of natural languages, we use deontic modals to express our commitments regarding what we should do. Our deontic modal commitments relate to our preferences. When we face a choice between several alternatives, and we prefer one of them, we often say we *should* or *ought to* choose that alternative.[2] It is also the case that our preferences depend on the information we have about the options available and their consequences. Information determines how likely we are to reach our goals after deciding for a course of action. In this paper, we will look at the varieties of monotonicity of deontic modals through a study of how our preferences evolve with the acquisition of information.

The property of monotonicity for deontic modals has been discussed under different names, e.g. persistence, stability, or independence. We call it Stability here. My thesis—which I call *Unstable doxastic preference*—is that the reasoning with necessity deontic modals and preferences is unstable. The thrust and scope of this thesis are insufficiently acknowledged in the discussions of deontic modality.[3]

(2) **Unstable doxastic preference**: *The preference expressed by deontic modals is unstable under changes of the decision maker's belief state*

[1] For an extensive discussion of the monotonicity properties of quantifiers see Peters and Westerstahl (2006). For an intriguing empirical application of the notion of monotonicity see Chierchia (2013).

[2] This is not to say that, in general, preferences coincide with deontic commitments, but only that they coincide in some cases that can, in turn, tell us something about stability.

[3] Here is a representative sample of the recent positions on stability. Cariani (2011) rejects stability by positing a non-standard semantics for deontic modals. von Fintel (2012) retorts that the deontic modal semantics should be seen as stable, and that non-stability is obtained at the pragmatic level. Charlow (2013) defends a stable semantics, although he restricts the notion of stability. My account resembles the accounts that deny stability, but its motivation is different and I maintain that my account provides a more comprehensive view of stability and its empirical underpinnings. Moreover, I think that some of the arguments for stability are mistaken.

i. The evaluation of deontic necessity modals is unstable.
ii. The instability of preference is not only relative to knowledge, but also to degreed beliefs or probabilistic information.

This thesis—which will be further unpacked in the first sections of the paper—points to the instability of preference and deontic modal commitments, and to one of its important (but yet neglected) aspects. If stability means that deontic modal commitments and preferences do not change when the domain of possibilities (representing our knowledge) contracts or expands, instability will amount to saying that deontic modal commitments and preferences *do* change under both contractions and expansions of the domain. My aim in this paper is to show that Instability is a constraint on any account of deontic modals.

Stability is what drives the success of subset- and superset-inferences, where the sets in question consist of possibilities (or possible worlds). Superset-inferences with deontic modals do sometimes succeed, as shown by the sentences (1c-d), and similar cases can be produced in favour of subset-inferences. Now, the thesis (2) denies that any of these inferences are generally successful. So instability amounts to denying the validity of such inferences. But in discussing instability it will be convenient to go beyond characterising it in terms of subset- or super-set inferences, and define instability in terms of concepts that are more familiar in the analysis of deontic modals and preferences.

I express the basic fact of instability in several equivalent ways. Following the phrasing of (2 i), I say that deontic modals are unstable, in the sense that what is deontically necessary (or what we should do) changes with our knowledge or state of information. This is equivalent to saying that the preference for an act—the proposition in the scope of the modal—changes with the information state. This is also equivalent to saying that the relative desirability of possible worlds changes with newly acquired information. And instability can be equally well attributed to the main component in the meaning of a deontic modal: the deontic function. A deontic function—a function from sets of worlds to the most desirable or ideal worlds in those sets—is unstable in case it does not yield a constant set of ideal worlds when it takes as arguments subsets or supersets of a given information state (viz. the set of possible worlds representing an agent's knowledge).

The question of stability, even if at first sight an abstract one, is interesting in several respects. First, as suggested, it promises to offer a fruitful systematisation of modal operators similar to the systematisation of quantificational operators. Second, the question has some epistemological import. Stability has been seen as a normative principle guiding choice and thus as a principle that rational preferences should satisfy (cf. Sen (1993)). Given the close connection between what we prefer and what we should do, an inquiry into the stability property of deontic modals can yield a robust conclusion about the normative force of such a principle. Finally, in the recent literature there have been a number of opposing views about stability, and it is worth asking how fundamental is this controversy and whether there are ways out of the presumed disagreement.

Since my main efforts will go into defining and defending the instability of deontic modals and preferences, I have straightforward proposals regarding the (last two) issues to do with the normativity status of stability and the substantial disagreement over stability. I argue that stability has no normative force and that there is no substantial disagreement between accounts that incorporate stability and those that eschew it. I cannot fully address the first issue, since my argument for instability turns on pragmatic facts, whilst a systematic linguistic classification of deontic modals (with respect to monotonicity) will also have to delve more deeply into semantic matters.

1 Unstable preference

1.1 Semantic assumptions

I frame my argument in possible worlds semantics. I assume a restrictor semantics for modality (Kratzer (2012), von Fintel and Heim (2011)) and a Ramsey view of conditionals (Ramsey (1929), Stalnaker (1968), Arlo-Costa (2014)).

According to the restrictor semantics, modals are quantifiers over (covertly or overtly) restricted domains of possible worlds. Modals place two restrictions on their domains. The first restriction, known as the modal base, provides the type of worlds to be quantified over; in our case, these worlds represent epistemic possibilities. A good paraphrase of a restricted deontic modal is: *in view of what we know*, *we should q*, where *q* is an act or the proposition in the scope of the modal (also known as the prejacent of the modal). A second restriction, known as the ordering source, determines the 'best' or ideal or most desirable worlds in the modal base. This ordering is implicitly guided by norms and values that arrange possibilities (or possible worlds) from the most desirable to the less desirable possibility.

When we are asserting a necessity deontic modal statement we say of these most desirable worlds that they are in the set of worlds determined by the prejacent. Otherwise put, the necessity modal expresses a deontic function that takes the two restrictors as arguments and returns the most desirable worlds meeting the constraints imposed by these restrictors. Then the modal matches the most desired worlds against the worlds of the prejacent, and if the former are a subset of the latter, the modal statement is true.

According to the Ramsey view of conditionals, in assessing a conditional we take its antecedent as known, and we judge its consequent in light of our present knowledge, which includes the newly added antecedent. A conditional of the form *if p*, *should q* is evaluated as follows: we take p as known along with the other pieces of background knowledge we may posses, and in light of the resulting state of knowledge we evaluate q. To evaluate the conditional, we take the antecedent as known and examine how well the consequent is supported by our knowledge.

Putting together the restrictor view and the Ramsey view, we can say the following. When the consequent of the conditional is (explicitly or implicitly) modalised, the antecedent of the conditional constrains the first restrictor of the modal in the antecedent. Thus, the worlds over which the modal quantifies are the worlds in which the antecedent is true and—in line with the Ramsey view—worlds in which we *know* that the antecedent is true. Here are some examples of modal statements and their interpretation on the present assumptions.

(3) *Paraphrases of modals statements according to the restrictor-view of modals and the Ramsey-view of conditionals*

a. We should block neither shaft.
≃ All the epistemically possible worlds that are ideal with respect to our norms and values are worlds in which it is true that we block neither shaft.

b. If there are more yellow balls than blue ones, we have to choose yellow.
≃ All the epistemically possible worlds where there are more yellow balls than blue ones and which, moreover, are ideal with respect to our norms and values, are worlds where we are choosing yellow.

1.2 Decision-theoretic assumptions

To be able to assess the stability properties of modal statements such as the ones above, we have to think through some concrete cases. Two decision problems will serve this purpose. The first decision problem is well known in the recent philosophical literature.

- *The miners' puzzle* (M): ten miners are trapped into one of two shafts, we don't know which. We have sandbags to block one of the shafts. In order to save the miners, we have to decide between blocking neither shaft, blocking shaft A, or blocking shaft B. (Kolodny and MacFarlane 2010)

On the basis of this scenario, we can express the preferences and modal claims below.

(4) a. It is preferable to block neither shaft (rather than one of the shafts).
b. We ought to block neither shaft (rather than one of the shafts).

The second decision problem is this.

- *The Ellsberg paradox* (E): we have to choose a ball from an urn which contains ten coloured balls made of diamond on the inside, and we get to keep the ball we choose just in case we guess its colour. We know that there are three red-coloured balls, and the other seven are either blue or yellow, but we don't know in which proportion. (cf. Ellsberg 1961)

Several intuitive deontic judgements can be made in context (E).

(5) a. It is preferable to choose red (rather than blue).
b. We ought to choose a red ball (rather than a blue one).
c. A blue or yellow ball is preferable to a red or yellow ball.
d. We ought to choose a ball which is blue or yellow rather than one which is red or yellow.

The two scenarios give rise to decision problems, as they raise the question of which course of action we need to commit to. Should we block one shaft or no shaft at all? Should we choose a red ball or a blue one?

A decision problem has the following ingredients. A decision maker has to decide between alternative acts. As a consequence of choosing to act in certain ways the decision maker produces certain outcomes in certain states of the world. States are seen as sets of possible worlds. The states in which acts are to produce outcomes are e.g. the state in which the miners are in shaft A, or the state in which the ball drown is blue. As for the *acts*, they are the courses of actions that we may take in the two scenarios: e.g. blocking one shaft, viz. $B_a \vee B_b$, blocking neither shaft, viz. $B_n = \neg(B_a \vee B_b)$, or choosing red, C_r. The outcomes of acts are e.g. saving nine miners, S_9, or winning a diamond ball. In this framework, which borrows key notions from decision theory, the desirability of an act depends on the value of the outcome as well as on the chances of that outcome's occurrence. I will say that we ought to do an act as long as it is maximally desirable (in the sense of offering the best trade-off between probability and desirability). A maximally desirable act will be preferable to any alternative act.

It is thus a fact that acts determine outcomes in states. The facts themselves, however, are of no help if we do not possess information about them. We cannot take decisions if we lack information about which act produces which outcome in which state. Take a tuple of the form $\langle a_1, s_1, o_1 \rangle$ to represent a certain act that produces, in a certain state, such and such outcome. Then, we need to have information about which tuples of the form $\langle a_i, s_i, o_i \rangle$ are more likely to occur when we are acting in the actual circumstances of the world. It is possible, and indeed quite common, to be ignorant about the specific circumstances and outcomes of an action. [4] For instance, even if we think that saving ten miners is the best outcome in (M), it does not follow that we have to block one shaft. That is, even if we value o_1 (e.g. saving ten miners) above all the other possible outcomes, this does not mean that we have to act in order to produce a tuple of the form $\langle a_1, s_1, o_1 \rangle$ (whereby act a_1, e.g. blocking one shaft, is the way that we can bring about our most valued aim). It is important for us, as decision makers, to *know* or have *good evidence* that act a_1 brings about the outcome o_1. What the agents know or have information about when they start to deliberate will be represented by a set of possible worlds, the information state F (also known as the epistemic modal base in the standard modal semantics).

[4] We can also be ignorant about the fact that we are acting, but I shall put aside such cases.

1.3 The argument

Given this value of information, I argue that the principles according to which desirabilities of acts vary—the deontic functions delivering the most desirable worlds—are unstable. This is essentially thesis (2). I take the core of the thesis to be the following fleshed out version of (2 i).

(6) **Instability**: A deontic function that represents the preferences of a decision maker is unstable iff the decision maker's preference between two options can change as a result of learning or acquiring new information. (2 i)

Our thesis *Unstable doxastic preference* (2) can be thus spelled out as Instability (6), provided that we understand information as degreed belief, in line with (2 ii). I put forward two principles that will support Instability.

(7) **Shifting Outcomes**: A deontic function (or preference) is unstable for an agent if it is contextually sensitive to at least two different outcomes o_1 and o_2, and the agent values one above the other. [5]

Shifting Outcomes spells out a necessary condition for Instability: the existence of two non-equivalent outcomes. We can then add other conditions on top of Shifting Outcomes in order to obtain necessary and sufficient conditions for Instability. The different conditions we might add yield different forms of instability, and so different ways of understanding (6). (I will examine the main types of instability in the next section.)

Since I focus on *subjective* necessity modals—modals that are true from the perspective of the agent's information, rather than objective, as when modal statements are true according to how things stand independently of our knowledge—I assume that for an outcome to be an incentive for action, not only should the action produce that outcome, but the deliberating agent should have reliable information that the action produces the outcome.

(8) **Reliable Information**: The deliberating agents have, in principle, access to reliable information about the outcomes of their acts (and information consists of degreed beliefs along the lines of (2 ii)).

[5] The outcomes can be viewed, in the standard modal semantics, as ordering source propositions in $g(w)$ compatible with the modal base $\cap f(w)$.

Putting our assumptions together, the existence of two outcomes about which the deliberating agents have reliable information determine unstable deontic functions (and unstable preferences) for these agents.

(9) **Theorem**. *Shifting Outcomes and Reliable Information imply Instability.*

For ease of exposition, I only sketch the proof here and relegate a fuller demonstration to the appendix. The basic argument is easy to grasp once we acknowledge the set-theoretic relations between acts, outcomes, and the agent's knowledge. From a possible worlds perspective, learning produces a contraction or an expansion of the set of worlds representing the agent's epistemic states. And an act guarantees (or entails) an outcome just in case the former is a subset of the latter.

By the Shifting Outcomes hypothesis, there are two differently valued yet comparable outcomes. These outcomes take the form of two non-null and partially disjoint sets of worlds. By Reliable Information, we have enough information to guarantee that these two outcomes come about and thus we know how to act in order to obtain the two outcomes. In possible worlds terms, there exist two other sets of worlds (representing acts) that are disjoint subsets of the two outcomes. What's crucial is that Reliable Information can *alternatively* offer warrants for any of the differently valued outcomes. And this allows for the decision maker's shift in preference and deontic modal commitments.

In other words, the agent can acquire information that strengthens (or weakens) her information state and thus rule out (or rule in) some key worlds in such a way that the agent remains with a set of worlds in which either (i) a new act comes to guarantee a new outcome, or (ii) an old act no longer guarantees a formerly favourite outcome. To put it in yet another way, according to the agent's new information state the relations between acts and outcomes evolve in either of the following two directions: (i) the agent gathers reliable information that a better outcome can be achieved or (ii) the agent retracts information to the effect that the best outcome can be achieved. Therefore, the decision maker's preference for an act—or for a world to be in a certain way—is not constant or stable under acquisition of information.

To visualise the argument, consider the sketch in figure 1. The shaded *pivotal areas* represent the moves that determine an unstable switch in preference according to our argument. When the pivotal area in blue shade (left) becomes available (viz. is repopulated with worlds) as a result of having learnt that one of the assumptions that led us to believe that o_1 is guaranteed by a_1 is false, then outcome o_1 (and thus the set d_1) loses its ideality status.[6] When the pivotal area in red shade (right) becomes unavailable (and we remove all the worlds from it) as a result of having learnt that a_2 guarantees o_2, set d_2 gets to be the new ideal and act a_2 will be preferred instead of a_1. (Once more, it is crucial to keep in mind that the subset relation between an act and an outcome means that the latter is guaranteed by the former—this being a precondition for the outcome to accede to ideality status.[7])

The argument for instability demonstrates the possibility of a reliable outcome that reverses the initial preference for a certain act. The argument's premises, *Shifting Outcomes* and *Reliable Information*, are incontrovertible. Decision making is unthinkable without the existence of different outcomes. It also beyond doubt that the context of deliberation *can* make available information relevant to which outcomes are desirable and warranted.

That the two conditions are easily satisfied can be seen from examining the miners' scenario (M). This scenario satisfies Shifting Outcomes because we have (at least) two outcomes; for instance: saving nine miners and saving ten miners. Also, the statement of the Miners' Puzzle relies on the possibility of acquiring relevant information, since (we can reasonably assume that) if we learn that the miners are in one of the shafts, blocking the shaft they are in ensures the better outcome (saving ten miners), whilst not acquiring this information leaves us in the position of securing the less appealing outcome (namely, saving nine miners). Thus, Reliable Information is also satisfied. We can then conclude via the above argument that the preferences in (M) are unstable. (I will make a similar case about the preferences in (E).)

2 Kinds of (in)stability

Shifting Outcomes can be combined with other principles to yield different versions of instability. This is because the existence of two outcomes (a necessary condition for instability) is oblivious to how else we choose to individuate the propositions representing our knowledge (the modal base F), what is learned (ϕ), and the acts available (a_1 and a_2). We can then require F to represent epistemically necessary

[6] This is because $(a_1 - o_1)$-worlds are as good as $(a_2 - o_2)$-worlds.

[7] If the information state consists of degreed beliefs rather than knowledge, one way to see the corresponding notion of guaranty—reliable guaranty—is the following. An outcome is reliably guaranteed if there is an act that overlaps to a high degree with the outcome. See the section on Reliable Information.

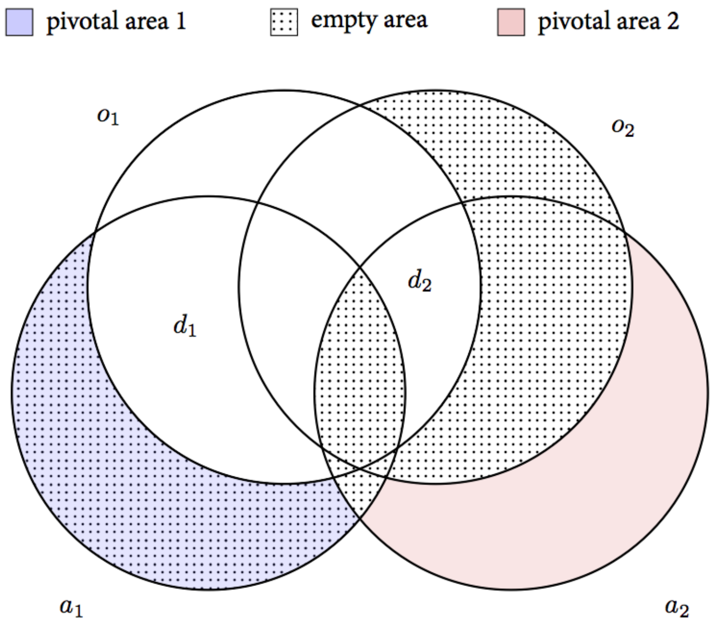

Figure 1: This is a sketch of the argument for (9). Outcome o_1 is more valuable than outcome o_2, $o_1 \subset o_2$ (this is required by Shifting Outcomes). Act a_1 is incompatible with act a_2 because only one act can be chosen: $a_1 \cap a_2 = \emptyset$. Act a_1 guarantees outcome o_1 (that is, $a_1 \subset o_1$), whilst a_2 does not guarantee (viz. o_2, i.e., $a_2 - o_2 \neq \emptyset$). So a_1 is preferable to a_2, and d_1 is the set of best worlds. The non-overlapping worlds in a_1 and a_2 are equally good. According to our argument for (9), there are two *pivotal areas* that can change the status quo (d_1) and confer ideality on a different set, d_2. If either of the pivotal areas change their status from being empty to being non-empty (or vice versa), the set of ideal worlds changes, and with it, our preferences change too.

worlds, or circumstantially available worlds. The latter are seen as representing facts, independent of our knowledge. Likewise, the proposition that a_i is conducive to outcome o_i can be an epistemically or circumstantially necessary truth. As we have said, since we are interested in subjective modals and preferences, the possible worlds and the relevant entailment relations are epistemic.

2.1 Permissive instability

We obtain more (or less) restrictive notions of instability if we require ϕ to include (or allow it to exclude) worlds that were best in F. A more restrictive notion of instability is one that requires that a world that is desirable in an information state fails to be desirable in an updated information state, *if that world exists* in the updated information state. For the sake of generality, I adopt a less restrictive conception of instability. On our less restrictive definition, instability covers situations in which the addition or retraction of propositions is incompatible with the action that was the most desirable in the original information state.[8] In other words, the update with these propositions rules out the most desirable worlds in the original information state. This happens when an option should be retracted though it was best among its alternatives. (For instance: we are about to order our favourite dish and the waiter lets us know that it is unavailable.) Whatever other option we may come to prefer, the act that we initially preferred is no longer an option. Hence, the worlds that were formerly considered ideal are not even epistemic possibilities after retracting that option.

It is an advantage to include such cases in the scope of one's account of necessity deontic modals, since such cases are common. Moreover, they are genuine cases of instability, since they too capture how preference is reversed with the acquisition of new information.

So *Instability*, as defined in (6), is a general principle. But for characterising some cases, a restricted version of the principle is more adequate. We can restrict it by requiring that ϕ contain some desirable worlds in the original information state. In doing so, we recover the (more restricted) notion of instability put forward by other theorists (e.g. Kolodny and MacFarlane (2010)). See figure 2.

This figure represents two types of updates. When we learn ϕ_2, some of the formerly ideal worlds d_2 are still epistemically possible, albeit no longer ideal. In contrast,

[8] Willer (2014, p. 10) also adopts the less restrictive definition of instability in a dynamic setting.

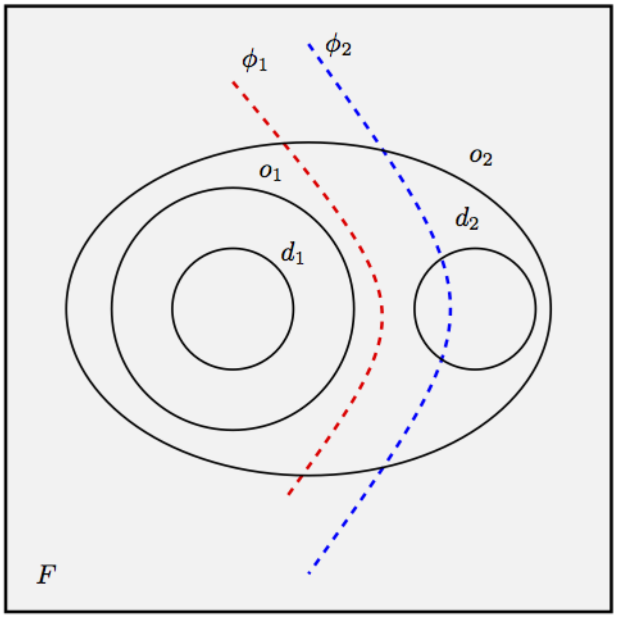

Figure 2: Two sub-cases of instability

when we update with ϕ_1, the formerly ideal worlds d_1 disappear completely, as they are no longer epistemically possible according to what we learned. Only a less restrictive notion of instability allows for the second type of update.

To see the use of the more restrictive notion of instability, consider what it denies:

- *Stability*: A deontic function is stable iff a world that is desirable in a given information state p preserves its desirability in a contracted information state r (with $p \subset r$) under learning new information, if it exists in that contracted state.

In the (M) context, Stability should be denied. This is because when learning that the miners are, e.g., in shaft A, the most desirable act is to block A, rather than the previously best act of blocking neither shaft. Given that the worlds where the miners are in shaft A include worlds in which we block neither shaft, and the latter were previously best but stopped being so in the current context, the existence condition imposed by *Stability* is observed, although the preservation of desirability fails.

So *Stability* should be denied in the miners' scenario, and a restrictive notion of Instability (incorporating the existence condition) captures the main characteristics of the scenario. (Although the less restrictive notion is a natural extension of the idea of instability, from now on I will focus on the restrictive notion, since it is the more commonly invoked notion. My claims can be easily extended to the less restrictive notion of instability.)

2.2 Downward and upward instability

Instability is, in yet another sense, a stronger principle than similar principles assumed in the literature. For instance, *Instability* is a strengthening of a principle that Kolodny and MacFarlane (2010) called serious information dependency, which is a denial of *Stability*. But we can define stability in a stronger form: *Bidirectional Stability*, or *Bi-Stability* for short.

- *Bi-Stability*: A deontic function is stable iff a world that is desirable in a given information state p preserves its desirability, if it exists in a contracted information state r (whereby $p \subset r$) under learning new information *or* in an expanded information state q (with $q \subset p$) obtained by retracting some proposition previously taken for granted.

Bi-Stability is a conjunction of two claims: downward stability ($\downarrow S$), which is equivalent with *Stability* above, and upward stability ($\uparrow S$), which is added by the disjunct in the new definition. And here we depart from the current views about stability. We claim that both $\uparrow S$ and $\downarrow S$ should be denied. [9] In contrast, Kolodny and MacFarlane (2010), Cariani et al. (2013), and Willer (2014) deny only their versions of $\downarrow S$.

By rejecting Bi-Stability, we are endorsing Instability. Since Bi-Stability is a conjunction of two types of stability, Instability, which is equivalent to the denial of Bi-Stability, is a disjunction of two types of instability (viz. a disjunction of the negations of downward- and upward-stability).

2.3 A restricted version of stability

In arguing for Instability, I'm not opposing restricted versions of stability. Indeed we can ask about the stability features of preferences relativised to particular steps in the learning process. I call this notion *stepwise stability*. Stepwise stable deontic reasoning is best understood in opposition to reasoning that does not have this property. (I will give this notion a more precise statement later on, when it will come to play a crucial role in my argumentation.) Contrast two pieces of reasoning that may occur in context (M).

(10) a. If the miners are in shaft *A*, we ought to block shaft *A*. We know that the miners are in shaft *A*. *And we know that the sandbags have just arrived.* Therefore, we should block shaft *A*.

b. If the miners are in shaft *A*, we ought to block shaft *A*. We know that the miners are in shaft *A*. *But we find out that the sandbags contain a substance that produces an asphyxiating gas in contact with water and that this gas could kill all the miners*. Therefore, we should not block shaft *A*.

The point of this contrast, which crucially depends on the italicised statements, is the following. Even if a deontic modal claim (or preference) is stable relative to the addition of a particular piece of knowledge (e.g. the information that the sandbags

[9] Another view that denies upward stability is that of Cariani (2011), but Cariani does so only implicitly (by his endorsement of a principle he calls Coarseness) in the course of arguing for a non-mainstream semantics for deontic modals.

have just arrived), it may not be stable relative to the update with another piece of knowledge (e.g. the information that the sandbags contain a mix of sand and a certain toxic substance). An update's impact on the contextual preferences may preserve these preferences only when considered in isolation from other pieces of information whose learning does *not* preserve the preferences. In other words, we can have stability, as long as it is relative to a particular step (or set of steps) in the learning process, but we cannot have it relative to *any* step (or set of steps).

3 Reliable information

The discussion so far focused on instability of modals and preferences in contexts where the best outcomes are certain. The second point of the *Unstable doxastic preference* thesis (2) proposes a generalisation of instability to cases where acts can be preferred even if they are not conducive to certain outcomes. The point can be defended by observing the probabilistic aspect of our second case study.[10] In case (E), we do not possess factual knowledge of the outcomes, but only knowledge about the probability of the outcomes. We are not certain about what ball-colour (red, blue or yellow) will be drawn. The probability of these outcomes according to our partial knowledge in setting (E) is all we have to go on in decision making.

This sets case (E) apart from (M). The best decision in (M) is based on a certain outcome (saving nine miners), whilst the best decision in (E) is based on an uncertain outcome (e.g. choosing red). The difference is represented in figure 3.

As the second matrix suggests, in choosing red we take into consideration its probability, which is 0.3. This is not the case with the Miners' scenario, represented in the first matrix, where blocking neither shaft guarantees the saving of nine miners, and thus S_9 is certain (all the B_n-worlds are S_9-worlds). So probabilities do not play a substantial role in (M), but they are essential for decision making in (E). By giving them a substantial role—in line with the second point of (2)—we can obtain

[10] Several other theorists, e.g. Cariani (forthcoming), Carr (2012), Lassiter (2011), make assumptions similar to *doxastic preference* about deontic modals in the sense of allowing for 'oughts' supported by credences less than 1. I remain neutral with respect to how best to conceive the contents of such probabilistic beliefs. Nor do I want to prejudge the issue as to whether these (probabilistic) doxastic states can constitute, at the same time, probabilistic knowledge (Moss 2013). Instead, my point will be that there is a small step—via scenarios such as (E)—from accepting such doxastic preferences and viewing them as unstable.

	B_a	B_b	B_n
A	S_{10}	S_0	S_9
B	S_0	S_{10}	S_9

	C_r	C_b	C_y	
R	1	0	0	$P(R) = 3/10$
B	0	1	0	$P(B) = (7 - n)/10$
Y	0	0	1	$P(R) = n/10$

Figure 3: Partitions with cells representing acts, states and outcomes for cases (M) (left) and (E) (right). Consider the first cell, viz. the leftmost cell in the upper row, in the (M) matrix. That cell represents the set of worlds where the miners are in shaft *A*, act B_a is done, and the outcome S_10 is obtained. The first cell in matrix (E) represents the worlds where we choose red (C_r) and we actually select a red ball (*R*) with the happy outcome of keeping that diamond ball (1). Since we know that in setting (E) there are exactly 3 red balls (out of a total of 10), the probability of selecting a red ball is 0.3. Moreover, given that the 7 remaining balls are blue or yellow, choosing *n* to represent the number of yellow balls, where *n* is between 1 and 6 (since we don't know the exact number of yellow balls), we get the probabilities of yellow ($n/10$) and of blue $7 - n/10$.

a generalisation of the accounts of deontic modality. [11]

An account of probability has to be combined with an account of desirability to obtain a comprehensive decision-theoretic account, which yields estimated desirabilities for propositions (in particular, acts and outcomes). Here are the basic ideas of such an account. Suppose that the probabilities of states and the (actual and estimated) desirabilities of acts and outcomes are expressed numerically. Suppose, moreover, that these numerical features are properties of possible worlds. Then, if an act has numerical estimated desirability *n*, it can be represented by the set of possible worlds in which the act is done with the effect of ensuring a desirability with value *n*. [12] Since on this picture probabilities and desirabilities are encoded by possible worlds, we can formulate the set-theoretic relations (over sets of possible worlds) needed in the argument for (9), and the thesis (9) itself will carry over to probabilistic cases such as (E).

Alternatively, we can encode information about desirability in a richer structure formed from sets of worlds and numerical desirability functions mapping sets of worlds to their numerical desirabilities. Once we have an account that assigns desirabilities to sets of worlds, we can formulate Shifting Outcomes in terms of two outcomes with different desirabilities, i.e. two outcomes mapped to different numerical desirabilities.[13] And this will lead to an equivalent of (9) for decision problems where probabilities matter.

We are now in a position to draw the moral of this section. Stability becomes even more difficult to maintain if we are making full use of *Reliable Information*. According to this condition, the information state may not represent knowledge

[11] This is a descriptive point about the additional structure needed to tackle (E). As such, (2 ii) and Reliable Information are auxiliary assumptions supporting our core instability thesis. Kolodny and MacFarlane (2010, 131, fn. 26) explicitly acknowledge the possibility of enriching their framework by taking into account the probabilities attached to propositions. In the context of their discussion of (M), representing probabilities is not necessary. Consequently, their assumption that an information state represents factual knowledge and that 'oughts' supervene on (known) facts is also sufficient for their purposes. However, neither of these assumptions can yield a good understanding of cases such as (E), since these cases give rise to 'oughts' that are not based on factual knowledge, but rather on information about the probabilities of situations or propositions (cf. Cariani (forthcoming, pp. 10-11)).

[12] The desirability of a proposition in a possible world has to be in part a property of the deliberating agent in that world.

[13] This alternative is more in line with the strategy pursued in the literature. For accounts of (epistemic) modality that incorporate probabilistic information, see e.g. Moss (2015), Yalcin (2007), and Cariani (forthcoming). For accounts of (deontic) modality that also incorporate desirabilities see Goble (1996), Lassiter (2011), and my paper on modality and insensitivity.

(or epistemic necessities) but just reliable beliefs (or high credence beliefs). So the agent can give up pieces of information that she no longer considers reliable in view of incoming evidence. Thus, belief revision has a more destabilising effect than revision of our stock of factual knowledge, since factual knowledge is not retractable, whilst beliefs are. For instance, we may find out through a more careful calculation that, in the miners scenario, blocking neither shaft does not save nine miners, but just about two. We should then retract previous information incompatible with the result of the new calculation. The new set of guaranteed outcomes will look different, since it won't contain the outcome of saving nine miners. In contrast, on the view that takes 'oughts' to require knowledge, the outcome of saving nine miners will be always guaranteed under any acquisition of information, and will always constitute a reason for action.

4 Evidence for Instability

We have examined several types of Instability which are common in deontic reasoning. We have seen that Instability carries over to decisions and deontic judgements whereby uncertainty has a more substantial role. Now, I would like to present further empirical evidence for the Instability thesis, and reject several arguments for stability.

4.1 Modus ponens and downward stability

There is some controversy in the literature about the relation between the validity of modus ponens and the validity of stability. Kolodny and MacFarlane (2010) reject a classic version of modus ponens and downward stability. By making modus ponens the main culprit of puzzles generated by scenario (M), they suggest that the failure of modus ponens entails instability. On the other hand, there are arguments that severe the link between the failure of modus ponens and instability. Charlow (2013) argues that neither modus ponens nor downward stability are problematic.[14] If this is true, the failure of modus ponens cannot be taken as a data-point for Instability.

I think that modus ponens fails precisely because it entails stability. Arguments to the contrary misdiagnose the significance of modus ponens in the miners' puzzle,

[14] A similar view of modus ponens in the context of the miners' puzzle is defended by Willer (2012).

and they end up encouraging too sanguine a view about the prospects of stability ($\downarrow S$).

There is no way around the fact that modus ponens entails instability (in the restrictive sense). Consider modus ponens in the context of (M): $\phi, \phi \rightarrow \Box_d\psi \vDash \Box_d\psi$. Following Kolodny and MacFarlane (2010), we take → to represent an indicative conditional (which is true if, at any world, the consequent is true whenever the antecedent is). Initially, the most desirable worlds are those in which the deliberating agents block neither shaft, that is, $d(F) = B_n$. Now suppose that the miners are in shaft A, and that this is what ϕ stands for. Once we restrict our attention to $F + \phi$ and apply modus ponens, the most desirable worlds appear to be those in which we block shaft A, which are B_a-worlds. So $d(F + \phi) = B_a$. If this is the case, Instability is violated, since ϕ-worlds (i.e., A-worlds) contain some of the worlds that were most desirable in the previous information state—viz. worlds where we block neither shaft thus saving nine miners—and these worlds fail to be best in the new state $F + \phi$.

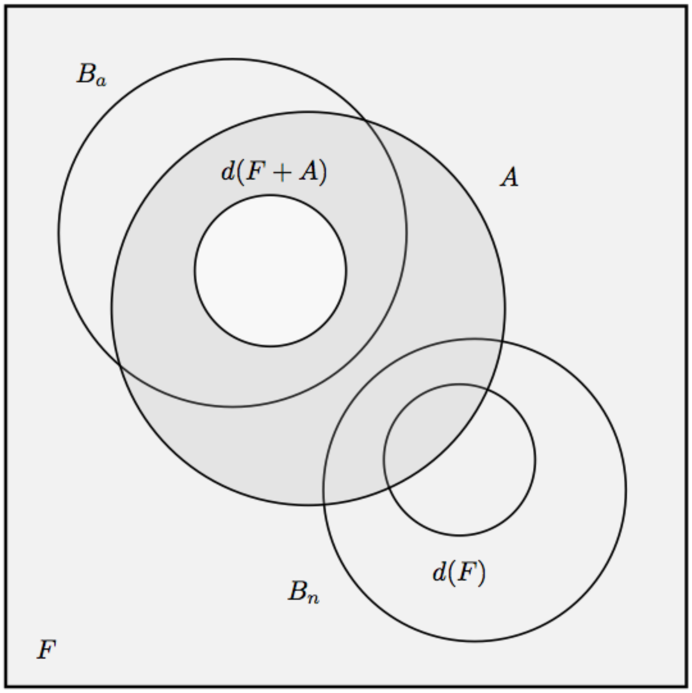

Figure 4: The failure of modus ponens

Charlow (2013) retorts that we get inconsistent modal recommendations even on assuming a shifty (or Ramsey) interpretation of the conditional, and thus of modus ponens, as long as we assume downward stability. However, I believe that it is impossible to get such a contradiction from the (M) setting. This is because on a shifty interpretation of the conditional, it is impossible to make the two premisses of modus ponens true. The shifty interpretation of the conditional $\phi \Rightarrow \Box_d\psi$ amounts to saying that we ought to ψ if we know ϕ. Then, for modus ponens to apply at all we have to assume that we know ϕ. But knowledge of ϕ is something we cannot assume in (M), since in (M) we know neither that the miners are in A nor that the miners are in B.[15] (We know that they are in A or B, but we can neither infer that we know that the miners are in A nor that we know that the miners are in B. It is perfectly possible to know a disjunction without coming to know any of the disjuncts. For instance, I know that at this moment the number of people leaving in India is either odd or even, although it is not true that I know that number is odd or that I know that it is even.)

I conclude that a failure of (classic) modus ponens entails lack of downward stability, and since we have independent reasons for denying modus ponens[16], we have equally good reasons for denying downward stability.[17]

[15] Charlow (2013, p. 2302) not only claims that modus ponens is not at issue but also advances a proof of this claim. The proof purports to show that a contradiction can be obtained out of the assumptions in scenario (M) under a shifty interpretation of the conditional. If I'm right, this proof must be mistaken. I suspect that the mistake lies in line 5 of Charlow's proof, in particular in his application of 'context-invariance'.

Given Charlow's way of defining a shifting context (p. 2301), but also plausible assumptions about modal reasoning in general, we are not allowed to infer from the statement that world i is a world where the miners are in shaft A to the statement that the world i is a world where it is true that the miners are in shaft A *and we know (or it is epistemically necessary that) they are in* A. This reasoning is definitely wrong, by Charlow's own admittance (cf. p. 2301). Yet, his application of 'context-invariance' gives rise to precisely this problematic reasoning.

[16] See Kolodny and MacFarlane (2010) and references therein.

[17] There is yet another argument against the centrality of denying modus ponens in order to solve the miners' puzzle. Willer (2012) argues that modus ponens is not longer problematic on a dynamic conception of logical consequence, and that the dynamic conception is better suited for an account of deontic modal reasoning than the classical notion used by Kolodny and MacFarlane (2010). While I find the dynamic approach to the miners' puzzle appealing, I think that Kolodny and MacFarlane (2010) 's point still stands: modus ponens fails *under a classical interpretation of validity and logical consequence*, no matter how good/bad the classical interpretation is for the purposes of modal semantics in general. The key point of these theorists, I take it, is that the classical interpretation serves as a convenient foil for determining some notable semantic properties of conditional and deontic modal reasoning.

4.2 Disjunction introduction and upward stability

As a consequence of rejecting Bi-Stability, what fails is not only modus ponens (due to the failure of $\downarrow S$), but also other inference rules that lead to deontic modal conclusions, e.g. disjunction introduction in modal contexts (due to the failure of $\uparrow S$). Here are two problems related to rules involving disjunction.

- In the miners' context it is true that $\Box B_n$, but it is not true that $\Box(B_n \vee (B_a \vee B_b))$, which is obtained from the former by disjunction introduction within the scope of the modal: if we ought to block neither shaft (B_n), we ought to block neither or block one of them ($B_n \vee (B_a \vee B_b)$). But in the miners' context the latter disjunct cannot make true the deontic modal claim, although this claim can be made true by the second disjunct.
- In the Ellsberg context it is true that we prefer (and we have) to choose red rather than blue: $\Box(C_r \succ C_b)$. Yet it is not true that we prefer (or have to) choose a ball which is red or yellow rather than a ball which is blue or yellow. So $\Box(C_r \vee C_y \succ C_b \vee C_y)$ comes out false. The latter is derived by introducing a disjunct in both terms of the inequality.

Let us focus on case (E). Intuitively, the inference from *we ought to choose red rather than blue* to *we ought to choose red or yellow rather than blue or yellow* is judged to be true because we think that the introduction of the same disjunct in both terms of the comparison (*choose red* and *choose blue*) cannot change the truth-value of the deontic modal claim. After all, we (disjunctively) add the same amount of preference to both terms, and this, the reasoning goes, shouldn't interfere with our initial preference for C_r (the act of choosing red). However, as the intuitive judgements in (E) make clear (see (5)), this reasoning is incorrect: red or yellow is a worse option than blue or yellow.

We can see the disjunction problem in yet another way. If C_r is better than C_b in a fixed domain, $C_r \cup C_b$, then for any C_y, $C_r \vee C_y$ will be better than $C_b \vee C_y$ in the same domain. C_y's intersection with the domain is null, and hence the disjunctive addition of C_y to each term doesn't affect the relative preference between these terms. This however is not the case when we extend the domain to $C_r \cup C_b \cup C_y$ (by adding C_y). The previous inequality does no longer hold in this extended domain. Instead, it becomes true in the new domain that $C_r \vee C_b$ is a worse option than $C_b \vee C_y$.[18] Otherwise put, the statements in (11) appear to be true:

[18] The (E) puzzle we are concerned with can be presented in conjunctive (i) or disjunctive form (ii).

(11) a. If only a red or blue ball yields a prize, a red or blue ball is preferred to a blue or yellow one.
b. If any ball (i.e., a red, blue, or yellow ball) yields a prize, a red or blue ball is dispreferred to a blue or yellow one.

A new disjunct introduced in the antecedent of (11 a) reverses the order of preference between two options, as evidenced by (11 b). Recall that on our assumptions about modals and conditionals the antecedents of conditionals restrict the information state (or epistemic modal base). The restricted information state in which the (consequent) modal statement is evaluated in (11 b) is a superset of the corresponding information state in which the (consequent) modal statement is evaluated in (11 a). This is a failure of upward stability.

Upward stability fails in even more unexpected cases. Willer (2014), like Kolodny and MacFarlane (2010) and Cariani et al. (2013), invalidates $\downarrow S$ and validates a restricted version of $\uparrow S$ through his commitment to what he calls *Uniformity* (see Willer (2014, p. 9)).

(12) *Uniformity*: If $B(\sigma) \subseteq \tau$ and $B(\tau) \subseteq \sigma$ for a fixed deontic function $B(\cdot)$, then $B(\sigma) = B(\tau) = B(\sigma \cup \tau)$.

Willer (2014, p. 16) takes uniformity to be a well-established principle of dynamic reasoning. While I do not intend to dispute Willer's applications of the principle to the cases that concerned him (e.g. Chisholm's paradox and the miners' puzzle), this principle is not adequate for modal reasoning in general, since it fails in probabilistic contexts. To see this, let us distil two claims from (12). Assuming the premise of (12) to the effect that the ideal σ-worlds are also τ-worlds and vice-versa, then (12) comes down to two independent conditions, which are each necessary, and, taken together, sufficient for uniformity:

i. $B(\sigma) = B(\tau)$ is ideal in $\sigma \cap \tau$ (symmetry condition)
ii. $B(\sigma) = B(\tau)$ is ideal in $\sigma \cup \tau$ (restricted form of $\uparrow S$)

(i) When we can choose more than one colour, choosing blue *and* yellow is preferable to choosing red and yellow. (ii) When our choice specifies conditions for the drawn ball to satisfy in order to bring out the best outcome, choosing a blue *or* yellow ball—viz. a ball that satisfies the condition of being blue or yellow—is better than choosing a red or yellow ball. The conjunctive form of the puzzle introduces a violation of $\downarrow S$, while the disjunctive form of the puzzle introduces a violation of $\uparrow S$. (See my paper on disjunctive modality for discussion of disjunction.)

The former condition is a symmetry condition, and, as such, does not interfere with the notion of stability.[19] The second condition is more relevant to our concerns. It is a restricted form of upward stability, since it projects the most desirable worlds from subsets (σ, τ) to a superset ($\sigma \cup \tau$). I think that not even this restricted version of upward stability is correct.

Here are two scenarios that provide counter-models to condition (ii).

- *Ellsberg*: In setting (E), red balls are preferred to any other ball (blue or yellow) in one-to-one comparisons, but they are not preferred to blue balls and yellow balls taken together. So red is better than any other competitor in any of the two relevant subsets (namely, the set of balls which are red or blue, and the set of balls which are red or yellow), but it is worse than its competitor in the superset (that is, the set of balls which are red or blue or yellow).
- *Additivity*: We have to choose among sets made up of one hundred tokens of equal value, say lottery tickets. The tokens are grouped as follows: one set S_1 of 40 tickets and two sets S_2 and S_3 of 30 tickets each. In the first step, we compare each of S_2 and S_3 to S_1 and we choose the best. Plausibly, each time we would prefer S_1. So S_1 is ideal in both $S_1 \cup S_2$ (a set which plays to role of σ in (12)) and $S_1 \cup S_3$ (τ in (12)). S_1 itself is thus identical to $B(\sigma)$ and $B(\tau)$. Now, is S_1 ideal in the larger set, $S_1 \cup S_2 \cup S_3$? If we allow mixed choices under the form of union of sets of tickets, it isn't. Indeed, $S_2 \cup S_3$ will be preferred to S_1 and hence S_1 is no longer the most desirable option in the newly enlarged domain.

These scenarios show that an option that was ideal in a given domain becomes non-ideal in a superset of that domain. Consider scenario (E): red is preferred to any other

[19] That said, I think that this symmetry condition is too strong. Consider the following scenario, which we may call *Incompatible persons*. Bob and Nina's parents need help. The parents need someone to come over and fix their computers. Both Bob and Nina are able to help with computers. But Bob and Nina do not get along, and a scandal would come out of their meeting, which would disturb everyone. In this situation, if the mother's computer is broken (σ), it is desirable for Bob to come $B(\sigma)$, and if the father's computer is broken (τ), it is desirable that Nina come $B(\tau)$. Note that it is also the case that it is equally desirable for both parents if the other kid come to their aid. (No parent has a particular preference for one of the kids; they would be glad to see any of them and get the needed help.) In this situation though, it is not desirable that *both* Bob and Nina come to help their parents. The invalidity of the symmetry condition follows from Shifting Outcomes (7). This is because the intersection of two sets, $B(\sigma)$ and $B(\tau)$, is not desirable in the Incompatible persons scenario despite $B(\sigma)$ and $B(\tau)$ being locally desirable on their own in τ and σ respectively

alternative (i.e., it is preferred to choosing blue and choosing yellow, respectively) in $C_r \cup C_b$-worlds and in $C_r \cup C_y$-worlds, but is no longer preferred in $C_r \cup C_b \cup C_y$, since in that set it is competing with $C_b \cup C_y$, which is a far better alternative. Likewise for Additivity. The locally desirable outcome $S_1 = B(\sigma) = B(\tau)$ in the initial step of Additivity stops being desirable when the set of options is broadened to $\sigma \cup \tau$.

This completes our probabilistic case against upward stability. (See the next section for non-probabilistic cases.) I have shown that Uniformity, and, in particular, its implication of a restricted form of upward stability, cannot be right in probabilistic contexts. In our terms, the restrictive form of upward stability is not valid, because the premise of Uniformity is compatible with the existence of two distinct outcomes, and, moreover, with the existence of an update (in our case, a broadening of the universe) which would make the better outcome available, and thus would make the switch in preferences possible, and, indeed, necessary.

4.3 Rational choice desiderata and stability

Several theorists claim that stability is well supported by rational-choice theory (Charlow (2013, p. 2037), Lassiter (2011, pp. 139-141)), and that it is a property well-worth keeping for the purposes of natural language semantics (e.g. Silk (2014, p. 703), von Fintel (2012, p. 14)).

Charlow (2013, p. 2307) cites an example of Sen's that purports to show that stability[20] is appealing for rational choice theory. Sen (1969, p. 384) suggests that stability is as sure as the following piece of intuitive reasoning: *If the world champion in some game is a Pakistani, then he must also be the champion in Pakistan*. This, making some plausible background assumptions about such competitions, is indeed an intuitive piece of reasoning. Even so, I find the analogy between our cases and Sen's inference problematic, because this inference does not depend on any modal feature of the sentence. Rather, the inference relies on the properties of the description *the world champion* and its restrictor (*world champion*). However appealing this inference might be, invoking it here, in the context of deciding on the linguistic properties of deontic modals begs our starting question: To what extent does the monotonicity of modal statements resemble the monotonicity of non-modal, quantificational ones?

Universal quantifiers behave differently from deontic necessity modals with respect to superset-inferences like the ones below.

[20] Sen calls it Property α or Independence of Irrelevant Alternatives

(13) a. Each student bought a cheap ticket.
b. ⇒ Each student bought a ticket.

(14) a. Students ought to buy a cheap ticket.
b. ??/?⇒ Students ought to buy a ticket.

We observe that upward monotonic inferences succeed more easily with quantifiers than with deontic necessity modals, because the latter are more sensitive to the shift of context (linguistic or otherwise). In (14), it is easy to get a reading of sentence (b) on which it is permissible to buy *any* ticket, and this reading is not entailed by sentence (a). This behaviour of superset-inferences in the scope of deontic modals (*a ticket* is a superset of *a cheap ticket*) is due to the fact that it is possible to determine the domain of quantification of *a ticket* in (b) not relative to the set of cheap tickets introduced by (a), but relative to the domain of tickets in general. [21]

So simply invoking an inference relying on non-modal properties of the restrictor of a description will not do as an argument for (or against) the monotonicity proprieties of deontic modals. Deontic necessity modals depend on our beliefs and cognitive states in a way regular universally quantified expressions don't.

This observation points to another obstacle in the way of borrowing principles from rational choice theory. In that field, stability was not claimed to deal with the *impact of knowledge or lack thereof* on choice (see e.g. Sen (1993), Savage (1972)), and the theories of rational choice do not in general assume the dependence of preferences on states of knowledge. Indeed it is precisely when knowledge is taken into consideration that stability is called into question.[22] There is then no reason to suppose that cherished principles in the theory of choice will carry over to modal reasoning involving knowledge- and information-sensitivity.

Most importantly, there are independent reasons—well-known in the theory choice literature—for rejecting Stability (and even Bi-Stability) as a principle of rationality. These reasons constitute empirical evidence for our central claim.

Sen (1993, 500ff.) points out that we may have to re-evaluate (downward) stability when confronted with real life choices (cf. Sen (1969, p. 384)). The so-called

[21] Cf. the discussion in von Fintel (2012, p. 15) and references therein.

[22] There is a long decision-theoretic tradition that proposes relaxations of (Bi-)Stability. The corresponding decision theories are reactions to axiom of independence in the expected utility theory of von Neuman, J. and Morgenstern, O. (1944), and its equivalent—viz. a version of the Sure-Thing principle—put forward by Savage (1972). See e.g. Machina (1982), Sen (1985), and Shafer (1986) for discussion and further references.

positional choice phenomena offer us another angle on the failure of stability. [23]

(15) *Preferences based on the principle: Do not choose the last remaining apple!*

a. I prefer having nothing (x) to having the last remaining apple in the fruit basket (y). ($x \succ y$)
b. But if another apple (z) is added to the basket, I prefer to have the apple y to having nothing (x). ($y \succ x$)

(16) *Preferences based on the principle: Never pick the largest slice of cake!*

a. I prefer having the smallest piece x to having a medium sized piece y. ($x \succ y$)
b. But if a larger piece z becomes available, I prefer having y to having x. ($y \succ x$)

The decision maker that adopts a positional choice strategy maximises the outcomes of her acts only after removing the best act from the menu of options. This makes her relative ranking of options liable to change when a better option is added to the menu. There are several other ways in which the decision maker can change her mind. Particularly relevant to our discussion of the role of knowledge in decision making are cases that bring out what Sen (1993, p. 502) calls the *epistemic value of the menu*. Let us follow one of Sen's examples.

Suppose that a distant acquaintance invites us over to his home to have tea and get to know each other better. This acquaintance also lets us know beforehand about the choices of refreshments. So we have to choose between having these refreshments in the company of our acquaintance and refusing to visit him altogether. Consider two particular sets of options that we may face in this scenario and the corresponding preferences that we would reasonably express in the two cases.

(17) a. If we have to choose between having tea or orange juice with our host, and not going at all, we prefer going and having tea to not going.
$\simeq$ Options: tea, orange juice, not going $\Rightarrow$ Preference: having tea $\succ$ not going
b. If we have to choose between having tea, orange juice, or *cocaine* with our host, and not going at all, we prefer not going to having any other

[23] The examples are adapted from Sen (1993, 501ff.).

option on the menu.
$\simeq$ Options: tea, orange juice, *cocaine*, not going $\Rightarrow$ Preference: not going $>$ having tea

Our conditional preferences seem to be coherent. Nevertheless, we rank having tea above not going in (17 a), and not going above having tea in (17 b). The deciding factor is that the menu in (17 b) includes an additional item: the option of having cocaine. The very availability of this option suggests that our acquaintance may be a trouble maker, and the reversal of choice is determined by our not being willing to associate with such a person. Thus it turns out that we can reassess the value of some options on the menu in light of the information we receive from an additional option. It would be a waste for decision makers not to avail themselves of this subtly conveyed information, and sometimes such subtle information determines unexpected turnarounds of their preferences and deontic modal commitments.

All of these cases are pathologically unstable. When an additional option is made available, the relative preference relation between two initial options changes. If initially we ought to do act a_1 rather than act a_2, when presented with additional options we appear to flip our preferences and deontic modal requirements. We thus ought to a_2 rather than a_1. In possible worlds terms, when the epistemic possibilities widen with the addition of the worlds where the additional option is available, a certain world that was considered ideal in the original domain, e.g. a world where we are having a certain cake or we are having tea, ceases to be ideal when the domain expands.

The examples above provide counterexamples to upward stability. It is easy to construct similar cases against downward stability by modifying the case (17), where the epistemic value of the menu matters. Instead of adding an option, as we did in (17), let us think about cases where one of the existing option is altered by adding a conjunct to it. (The effect of adding a conjunct is that the information state contracts, and thus our example becomes pertinent to the evaluation of downward stability.) So let the initial choice be between having tea or orange juice in the company of our acquaintance and not seeing him at all. Subsequently we are offered a choice between having tea, having orange juice *and cocaine*, and not meeting the acquaintance at all.

(17') a. If we have to choose between having tea or orange juice with our host, and not going at all, we prefer going and having tea to not going.
$\simeq$ Options: tea, orange, not going $\Rightarrow$ Preference: having tea $>$ not going

b. If we have to choose between having tea, orange juice, or *cocaine* with our host, and not going at all, we prefer not going to having any other option on the menu.
$\simeq$ Options: tea, orange, *cocaine*, not going $\Rightarrow$ Preference: not going $\succ$ having tea

Faced with the initial choice, we would prefer having tea to not going. This preference is reversed after one of the options is replaced by the cocaine and orange juice cocktail. Tweaking some option on the menu can dramatically change our preference between two other options. Therefore, downward stability is not in a better position in cases where the value of the menu matters.

4.4 Beyond deontic modals

Thus far we have focused on deontic modals and preferences. But once we start looking for Instability, we find that this property is a more general feature of modal reasoning. Epistemic modals, knowledge ascriptions, and counterfactuals are also *loci* of unstable inferences.[24] Moreover, Instability is a basic feature of non-monotonic logics (cf. e.g. Horty (1997), Hawthorne and Makinson (2007)).

Hence the empirical domain of Instability can be broadened beyond deontic modality with constructions like the following.

(18) a. If Sophie had gone to the parade, she would have seen Pedro.
b. But if Sophie had gone to the parade and been stuck behind a tall person, she would *not* have seen Pedro.

(19) a. I know that I'm going to fly back to Europe this summer.
b. I do *not* know that I'm not going to die the while. [25]

These constructions have been discussed in relation to the semantics of counterfactuals and epistemic operators. For our purposes, it suffices to note that they license non-monotonic, unstable inferences. That is, we are not allowed to infer from a proposition constituting a superset (e.g. the set of worlds where Sophie goes to the

[24] See von Fintel, Kai (2001), Moss (2012), Alonso-Ovalle (2009) and Lewis (2001).

[25] Cf. *(For all I know) he must fly back to Europe this summer*, where the modal *must* receives an epistemic reading.

parade) to a proposition constituting a subset (the set of worlds where Sophie goes to the parade *and* some tall guy sits in front of her). So (18 a,b) reflects a failure of downward stability. As for failures of upward stability, note that accepting (19 a) and then (19 b) as true tells against an inference from a subset (viz. a set of worlds where I'm flying to Europe) to a superset (viz. a set of worlds where I'm not dying).

To sum up, we have gathered some robust and fairly well-known data whose significance is easy to overlook. However, once we get clear about the meaning of stability, the general direction suggested by such data is clear. Instability is ubiquitous. It is well supported by ideas from rational-choice theory, and there are good logic, linguistic, and epistemic reasons for adopting it.

5 Stepwise stability

Our efforts on behalf of Instability notwithstanding, there are accounts that succeed in preserving some version of downward stability at the cost of more intricate parametrisations of the semantic apparatus. For instance, Charlow (2013, pp. 2313-2318) relativises the deontic function to what he calls practical ends (our guaranteed outcomes), and Silk (2014, pp. 702-4) has the modal base determine the ordering according to which the worlds' desirability is evaluated.[26] Such strategies open up new ways of obtaining stability. This shows a tension concerning the role of instability in modal reasoning, and a point of disagreement between stability-accounts and instability-accounts. So it is worth getting clear about this tension. Are there substantial disagreements between stability-views and instability-views? How can we make the stability accounts palatable to an advocate of Instability?

We have seen that there is nothing sacrosanct about stability[27], and so positing

[26] It is not clear to me that this is exactly what Silk has in mind, but it is nevertheless instructive to consider this strategy of saving stability. A similar strategy is suggested by von Fintel (2012), who suggests that failures of stability turn on context shift (and thus on the shift of the information state). I think that this strategy is reasonable for semantic purposes, but, as we will see, it is not an argument against Instability. Silk (2014, 703; fn. 18) also suggests that a model of modality that takes into account the expected utilities of the acts (or modal prejacents) will meet the stability requirement. But this is not so, as we have seen in our analysis of (E) and (11), (15)-(16), and (17). Some further restrictions are needed, and fixing a modal base is the only restriction that would do given Silk's explicit semantic assumptions.

[27] Sen (1993), for instance, explicitly denies the existence of what he calls *internal consistency constraints*—which notably include an equivalent of the Stability constraint—on choice functions. Sen argues (convincingly) that these principles should be re-evaluated when considered in the

semantic structure just for the purposes of salvaging this principle won't do. But perhaps the semantic structure posited by the defenders of stability is independently motivated. I'm not in a position to argue against this possibility here. In the rest of the paper, I elude this issue altogether, and make a more fundamental point about the notion of stability that is implicit in the stability-accounts. I argue that, whatever their empirical motivation, these accounts don't in fact preserve stability but something much weaker.

In a nutshell, my argument is this. Keeping certain semantic parameters fixed—as the advocates of stability do—puts a heavy restriction on the types of options and alternatives that might become relevant to deliberation, and ultimately puts a heavy restriction on what the deliberating agents can learn (under the strictures of a stable preference). These strong restrictions on stability make it the case that the principle remains silent on the topic of preference reversal. The stability-accounts fail to address the phenomena consisting of deontic preference reversals which are the staple of Instability. They are not doing so much as denying Instability, but instead they are simply looking elsewhere.

To see this, we need to recall our discussion of kinds of (in)stability. Let's start with the very idea of stability. We have a theoretically neutral picture of the way stability constrains the decision maker's choice. When the decision maker faces a choice and she prefers a certain option to one of its alternatives, then, according to stability, this should be her preference no matter what other options disappear or become available. It is for this reason that the principle is sometimes called Independence of Irrelevant Alternatives.

This principle is consistent with—if a bit more general than—other glosses on stability throughout the literature (cf. e.g. Kolodny and MacFarlane (2010), Cariani et al. (2013), Charlow (2013)). These glosses are stated in terms of possible worlds, and refer to downward stability of deontic modals. They have it that a possible world's ideality or desirability status does not depend on what its competitors happen to be. Thus, focusing on the downward stability, the worlds that are ideal in a given domain remain ideal when we restrict the domain (given that we do not get rid of those worlds altogether). This possible worlds perspective can be translated back into our picture of stability. Restricting the set of worlds is equivalent to removing an option from the menu, since by removing an option we eliminate from the information state the epistemic possibilities which are compatible with the choice

context of values, expectations, and aims underlying the choice. This is not to imply, however, that these principles or internal constraints should be defended at all costs.

of that option.

Now, compare this idea of stability with what I have called stepwise stability. Stepwise stability is stability relative to certain propositions: propositions whose learning does not exploit the desirability of the different outcomes available in the context. Any proposition ϕ such that act a_i entails outcome o_i for $i = 1, 2, \ldots$ in both F and $F \pm \phi$ is a proposition for which the deontic function is stepwise stable. (F is an information state or an epistemic modal base.) To illustrate, consider the diagrammatic representations of stepwise stable and stepwise unstable updates in figure 4.

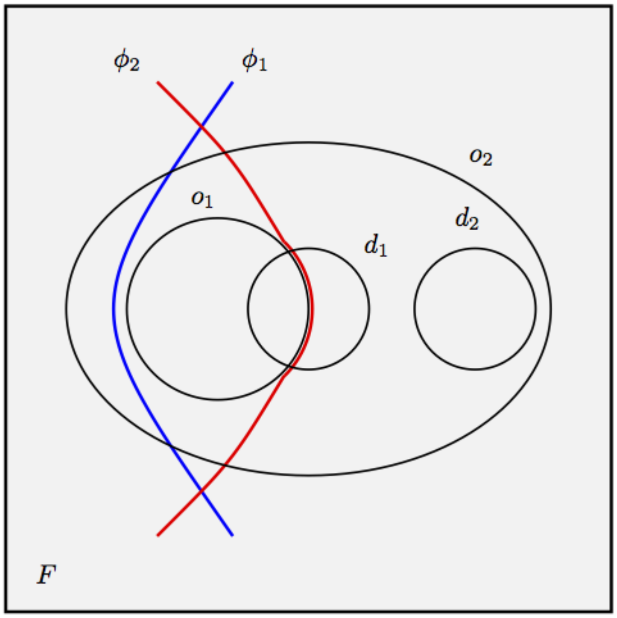

Figure 5: Stepwise stable and stepwise unstable updates

The stepwise stable update with ϕ_1 does not rule out any of the candidates for ideal worlds in d_1 and d_2, and thus it cannot change the decision maker's preferences. In contrast, the update with ϕ_2 does make a difference, because it rules out d_2 and the part of d_1 which does not overlap with o_1. In this case, o_1 can be realised, and since it is stronger than o_2 and thus more desirable (because $o_2 \subset o_1$), the update induces a switch from d_1-worlds to d_2-worlds. The latter are the best worlds, since there is an action (represented by a proposition true in those worlds) that guarantees the preferred outcome o_2.

A concrete instance of the two types of updates is found in the miners' context, as represented in figures 5 and 6.

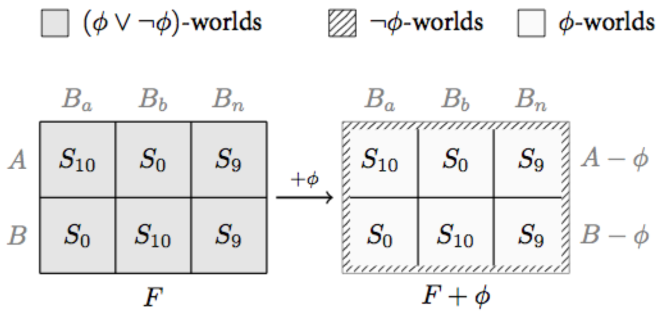

Figure 6: Stable update of the miners' context

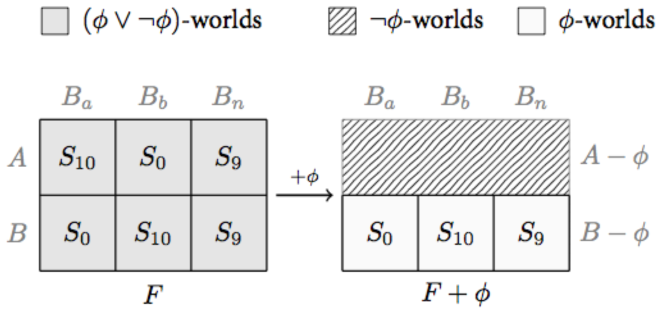

Figure 7: Unstable update of the miners' context

The updates which are stepwise stable (relative to ϕ) leave all the options untouched. As the first diagram shows, all the cells with which we have begun in F remain there, and represent live options in $F + \phi$. In contrast, the stepwise unstable updates remove cells, as shown in the second diagram. After learning ϕ, the outcome S_{10} can be guaranteed by the act B_a of blocking shaft A, and thus our preferences change. This is not the case with the stepwise stable updates, because, as shown in the first figure, blocking shaft B leaves open the undesirable possibility of loosing all the miners (S_0).

My point is that the stepwise downward stable updates include the updates that satisfy the assumptions meant to save stability, that is, the updates whereby the set of practical ends is kept fixed and the updates whereby the modal base is kept fixed. One of these points—that I derived from Silk (2014)—is not too informative: the updates whereby the modal base is kept fixed are just null updates or updates with information we already know. So the restriction to a fixed modal base is so strong that it trivialises the stability thesis. On the other hand, Charlow (2013) 's suggestion of relativising the deontic function to a fixed set of practical ends is weaker and determines a less trivial definition of stability. It is less trivial because it allows for transformations of the information state that does not change worlds with respect to the outcomes that they guarantee. (This is equivalent to forcing the modal bases to preserve the same guaranteed outcomes.)

Indeed, fixing the practical ends in the miners context does the job. In figure 6, the partition on the right is stable relative to further updates. It represents a set of practical ends, $\{S_0, S_{10}, S_9\}$, and a modal base, $F + \phi$. Relative to these parameters any further strengthenings will keep the preference for blocking shaft B unchanged. However, as we have seen, the miners' context is a relatively narrow decision problem.

A more general statement suggests itself, and this is our final formulation of stepwise stability. The decision makers can learn anything as long as they don't end up with an information state with no worlds realising the best outcome.[28]

(20) **Stepwise Stability** *If the modal base is partitioned such that there is a cell c which is best in the sense of representing the best outcome and this outcome is guaranteed or most reliable, then (and only then) any further strengthening that does not completely remove cell c is stepwise stable.*[29]

A cell of the partition is a set of worlds that represents the conditions (or state of the world) in which the agent acts in a certain way and produces a certain outcome. (See

[28] This claim defines stepwise downward stability in terms of a restricted version of stability. In order to use the less restrictive version, we have to substitute "less worlds" for "no worlds" in the above gloss, and "does not remove any world from cell *c*" for "does not completely remove cell *c*" in (20).

[29] Cariani et al. (2013, pp. 251-2) argue that (downward) stability—what they call *persistence*—holds for (i) the circumstantial modal base, (ii) situations in which all the relevant facts about the decision problem have been learned, and (iii) all the ordering source propositions are entailed by all the acts available. Likewise, Kolodny and MacFarlane (2010, 139ff.) argue that modus ponens can be saved by restricting its domain of application to (a) known antecedents, or (b) information-insensitive consequents. All these restrictions, except for (b), can be subsumed under stepwise downward stability.

figures 2, 5, and 6 for relevant examples.) Then, according to (20), stepwise stability requires that the learning (*qua* strengthening) leave the best outcome untouched, and that this outcome be guaranteed before and after learning.[30] The ensuing restriction is a bit weaker than the fixed-practical-ends strategy because it covers cases where the set of guaranteed outcomes is only partially fixed, and where—in line with Reliable Information—uncertain outcomes can be best. So (20) is the weakest condition that preserves stability and thus is entailed by any other condition suggested in the literature.

It follows that there is no real disagreement between the stability- and the instability-accounts. Fixing the set of practical ends or the modal base preserves stability because these strategies ensure that the cell c (representing the best outcome) never gets to be completely removed. Moreover, (20) yields a necessary and sufficient condition for the constituency of the type of epistemic modal base (or state of information) that guarantees downward stability under any number of learning-steps. This condition is weaker than null updates or fixed outcome restrictions. At the same time, the condition expressed by (20) heavily restricts the things the decision makers can learn, since it removes in one stroke preference reversals. On thing is clear: neither (20) nor the stronger alternative restrictions will make preference reversals disappear. Preference reversals do exist, and are as real as stable preferences. So Instability and stepwise stability, rather than being opposed, characterise different phenomena.

I will not make pronouncements on which of these stability-related phenomena is more common or robust. There is a clear way in which getting stability requires specific conditions, whilst instability appears in the absence of such conditions. Seen in this light, preferences and deontic modals are generally unstable, although they can be locally stable in restricted conditions.

[30] It is important to appreciate the role of outcomes in our argument. The fixation of parameters (specific to the stability-views) can leave open several warranted outcomes, as it happens in (M) after learning where the miners are. So one may think that we can argue *via* Shifting Outcomes that the preference and deontic modal commitments in that situation is unstable. While this is true according to (9), note that the preference is unstable in virtue of its being *upward* unstable. (Instability is a disjunction of upward and downward instability.) On the other hand, the concern of the stability-views was to defend downward stability.

6 Conclusions

I argued that Instability is both formally correct and empirically plausible. We have the capacity to revise, in light of new information, not only our beliefs but also our *deontic modal* beliefs, the beliefs about what we need to do. The revisions of the deontic modal beliefs are unstable because they can generate unstable preferences. I forged a connection between the instability of deontic modal beliefs and the contextual presence of distinct outcomes. Updated contexts *can* make available information that guarantees new outcomes, and insofar as these outcomes are differently valued than the ones guaranteed by our initially preferred choice, it becomes appealing to switch to the new choices (that give rise to the new outcomes).

I take Instability to be a substantial top-down constraint on a linguistic theory of deontic modality. Unlike quantifiers and other natural language operators, deontic modals are both downward and upward unstable. I remained neutral on the question of where to locate Instability, on whether it is a semantic or a pragmatic phenomenon. In a straightforward sense, Instability is a pragmatic phenomenon. This is because reversals in preference and changes in deontic commitments appear with shifts in context—where the context is taken to encode, along the standard semantic lines, the available information *qua* set of epistemic possibilities and the principles ordering the desirability of these possibilities. In light of this observation, arguing that stability should be preserved at the level of semantics still remains an intriguing theoretical possibility.

Appendix

(21) **Theorem**. *Shifting Outcomes and Reliable Information imply Instability.*

Proof. Let o_1 and o_2 be distinct, but comparable outcomes in accord with Shifting Outcomes. We can then pick an information state F fully compatible with these outcomes (viz. $o_1, o_2 \subset F$). We can further pick three sets in F, namely two disjoint acts a_1, a_2, and proposition ϕ such that $a_1 \cap F \subset o_1$ and $a_2 \cap (F \pm \phi) \subset o_2$. In such a situation, it is epistemically necessary that $(a_i \subset o_i)$ is true for $i = 1$ in F, but false for $i = 2$ in $F \pm \phi$. That we can pick sets with such properties is guaranteed by Reliable Information. Note, in particular, that in line with Reliable Information, the entailments between acts and outcomes are known to hold and thus are epistemically necessary given the state of information. For simplicity, I shall leave this assumption implicit.

By the Shifting Outcomes hypothesis, either $o_1 \succ o_2$ or $o_2 \succ o_1$. For present purposes, we understand $\succ$ in terms of $\subset$, i.e., $o_i \succ o_j$ iff $o_j \subset o_j$. (As I will show below, a more general understanding of the relation $\succ$ is possible, in terms of estimated desirability.) It then follows that the desirabilities attached to acts can be changed by exploiting the different desirabilities of outcomes obtained as a result (viz. entailment) of these acts when expanding or contracting F with ϕ. According to the relative ranking of the outcomes, there are two possibilities.

- If $o_2 \succ o_1$, the addition or retraction of ϕ to F, $\pm\phi$, will generate a better, more desirable outcome o_2, and thus $d(F \pm \phi) \subseteq a_2$ and $\Box_d a_2$ in $F \pm \phi$, as a_2 entails o_2 in $F \pm \phi$. Thus, the desirability of acts has shifted along with the desirability of the outcomes guaranteed (or entailed) by these acts, since o_1 is guaranteed by an act in F, whilst o_2 is guaranteed by a different act in $F \pm \phi$. (Note that Reliable Information also guarantees that when we update the information state with $\pm\phi$ we know ϕ or $\neg\phi$, as the case may require.)
- If $o_1 \succ o_2$, then we can proceed backwards, from $F \pm \phi$, to show that by the addition or retraction of ϕ, we can reach a state of information F where the preference between o_1 and o_2 is reversed. This is straightforward, since F and $F \pm \phi$ can be seen as modifications of each other. Note that if $F \pm \phi$ is an expansion relative to F, F is a contraction relative to $F \pm \phi$, and, likewise, if $F \pm \phi$ is a contraction relative to F, F is an expansion relative to $F \pm \phi$. Either way, by expanding or contracting an information state where o_2 is best (namely, $F \pm \phi$), we reach another information state where o_1 is best (namely, F). And thus since a_1 entails o_1, it will become the desirable action in the modified information state: $\Box_d a_1$ and $d(F) \subseteq a_1$.

In sum, if $o_1 \neq o_2$, there are sets F, a_1, a_2 and ϕ such that a_1 is preferred in one information state, and a_2 is preferred in another information state obtained by contracting or expanding the former. Therefore, an information-sensitive deontic function d is unstable iff there are at least two different outcomes, and we have reliable information about when these outcomes obtain. ■

References

Alonso-Ovalle, Luis (2009). "Counterfactuals, correlatives, and disjunction". *Linguistics and Philosophy* 32, pp. 207–244. doi: 10.1007/s10988-009-9059-0 (cit. on p. 29).

Arlo-Costa, Horacio (2014). "The logic of conditionals". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2014. url: http://plato.stanford.edu/archives/sum2014/entries/logic-conditionals/ (cit. on p. 4).

Cariani, Fabrizio (2011). "'Ought' and Resolution Semantics". *Noûs* 47.3, pp. 534–558. doi: 10.1111/j.1468-0068.2011.00839.x (cit. on pp. 2, 15).

— (forthcoming). "Deontic modals and probabilities: one theory to rule them all?" In: *Deontic modality*. Ed. by Nate Charlow and Matthew Chrisman. Oxford University Press (cit. on pp. 16, 18).

Cariani, Fabrizio, Stefan Kaufmann, and Magdalena Kaufmann (2013). "Deliberative modality under epistemic uncertainty". *Linguistics and Philosophy*. doi: 10.1007/s10988-013-9134-4 (cit. on pp. 15, 23, 31, 34).

Carr, Jeniffer (2012). "Subjective Ought". ms. MIT (cit. on p. 16).

Charlow, Nate (2013). "What we know and what to do". *Synthese* 190, pp. 2291–2323. doi: 10.1007/s11229-011-9974-9 (cit. on pp. 2, 19, 21, 25, 30, 31, 34).

Chierchia, Gennaro (2013). *Logic in grammar. Polarity, free choice, and intervention*. Oxford: Oxford University Press (cit. on p. 2).

Ellsberg, Daniel (1961). "Risk, ambiguity, and the Savage axioms". *The Quarterly Journal of Economics*, pp. 643–669 (cit. on p. 6).

Goble, Lou (1996). "Utilitarian deontic logic". *Philosophical Studies* 82.3, pp. 317–257 (cit. on p. 18).

Hawthorne, James and David Makinson (2007). "The quantitative/qualitative watershed for rules of uncertain inference". *Studia Logica* 86.2, pp. 247–297 (cit. on p. 29).

Horty, John F. (1997). "Nonmonotonic foundations for deontic logic". In: *Defeasible deontic logic*. Ed. by Donald E. Nute. Dordrecht: Kluwer. doi: 10.1007/978-94-015-8851-5_2 (cit. on p. 29).

Kolodny, Niko and John MacFarlane (2010). "Ifs and Oughts". *Journal of Philosophy* 107.3, pp. 115–143 (cit. on pp. 6, 12, 14, 15, 18–21, 23, 31, 34).

Kratzer, Angelika (2012). *Modals and Conditionals*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199234684.001.0001 (cit. on p. 4).

Lassiter, Daniel (2011). "Measurement and Modality: the Scalar Basis of Modal Semantics". PhD thesis. New York University. url: http://semanticsarchive.net/Archive/WMzOWU2O/ (cit. on pp. 16, 18, 25).

Lewis, David K. (2001). *Counterfactuals*. Oxford: Blackwell Publishers (cit. on p. 29).

Machina, Mark J. (1982). ""Expected utility" analysis without the independence axiom". *Econometrica* 50.277–323. url: http://www.jstor.org/stable/1912631 (cit. on p. 26).

Moss, Sarah (2012). "On the Pragmatics of Counterfactuals". *Noûs* 46.3, pp. 561–586 (cit. on p. 29).

— (2013). "Epistemology Formalized". *The Philosophical Review* 122.1, pp. 1–43. doi: 10.1215/00318108-1728705 (cit. on p. 16).

— (2015). "On the semantics and pragmatics of epistemic vocabulary". *Semantics & Pragmatics* 8.Article 5, pp. 1–81. doi: http://dx.doi.org/10.3765/sp.8.5 (cit. on p. 18).

Peters, Stanley and Dag Westerstahl (2006). *Quantifiers in Langauge and Logic*. Oxford: Oxford University Press (cit. on p. 2).

Ramsey, F. P. (1929). "General propositions and causality". In: *Philosophical Papers*. 1990th ed. Cambridge: Cambridge University Press (cit. on p. 4).

Savage, Leonard J. (1972). *The Foundations of Statistics*. Wiley, New York: Dover edition (cit. on p. 26).

Sen, Amartya (1969). "Quasi-transitivity, rational choice and collective decisions". *The Review of Economic Studies* 36.3, pp. 381–393. url: http://www.jstor.org/stable/2296434 (cit. on pp. 25, 26).

— (1985). "Rationality and uncertainty". *Theory and Decision* 18, pp. 109–127. doi: 10.1007/BF00134068 (cit. on p. 26).

— (1993). "Internal consistency of choice". *Econometrica* 61.3, pp. 495–521. url: http://www.jstor.org/stable/2951715 (cit. on pp. 4, 26, 27, 30).

Shafer, Glenn (1986). "Savage Revisited". *Statistical Science* 1.4, pp. 463–485. url: http://www.jstor.org/stable/2245794 (cit. on p. 26).

Silk, Alex (2014). "Evidence Sensitivity in Weak Necessity Deontic Modals". *Journal of Philosophical Logic* 43, pp. 691–723. doi: 10.1007/s10992-013-9286-2 (cit. on pp. 25, 30, 34).

Stalnaker, Robert C. (1968). "A theory of conditionals". In: *Studies in Logical Theory*. American Philosophical Quarterly Monograph Series 2, pp. 98–112 (cit. on p. 4).

von Fintel, Kai (2001). "Counterfactuals in dynamic context". In: *Ken Hale: A life in language*. Ed. by Michael Kenstowicz. Cambridge, Mass.: MIT Press (cit. on p. 29).

von Fintel, Kai (2012). "The best we can (expect to) get? Challenges to the classic semantics for deontic modals". In: *85th Annual Meeting of the American Philosophical Association, Chicago, IL*. url: http://web.mit.edu/fintel/fintel-2012-apa-ought.pdf (cit. on pp. 2, 25, 26, 30).

von Fintel, Kai and Irene Heim (2011). "Intensional Semantics". manuscript MIT. url: http://mit.edu/fintel/fintel-heim-intensional.pdf (cit. on p. 4).

von Neuman, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behaviour*. Princeton NJ: Princeton University Press (cit. on p. 26).

Willer, Malte (2012). "A remark on iffy oughts". *Journal of Philosophy* 109.7, pp. 449–461 (cit. on pp. 19, 21).

— (2014). "Dynamic thoughts on ifs and oughts". *Philosophers' Imprint* 14.28, pp. 1–30. doi: 2027/spo.3521354.0014.028 (cit. on pp. 12, 15, 23).

Yalcin, Seth (2007). "Epistemic Modals". *Mind* 116.464, pp. 983–1026. doi: 10.1093/mind/fzm983 (cit. on p. 18).