Informational Models in Deontic Logic: A Comment on "Ifs and Oughts" by Kolodny and MacFarlane

Karl Pettersson

Abstract

Recently, in their paper "Ifs and Oughts", Niko Kolodny and John MacFarlane have proposed modeling deontic logic on the basis of so-called informational models, with a new interpretation of both the deontic obligation operator and the indicative conditional, and claim that they can use this approach to solve some problems related to the conflict between so-called subjective (e.g. versions of consequentialism according to which the expected value of different alternatives given the beliefs or knowledge of an agent are morally relevant) and objective versions of the moral "ought" (e.g. versions of consequentialism according to which the consequences that actually would be realized by different alternatives are morally relevant). In this paper, I will critically discuss their claims. I claim that at least their indicative conditional, given the normative assumptions they seem to accept, yield problematic results in some contexts, and that its alleged benefits in the cases they discuss can be explained in other ways, viz. by distinguishing between different senses of the normative expressions.

1. The Trapped Miners

Kolodny and MacFarlane begin their recent paper "Ifs and Oughts" by presenting a scenario (Kolodny & MacFarlane 2010, p.115f) with the same structure as a well-known example from Frank Jackson (Jackson 1991, p.462f), where it is certain that the alternative with best actual outcome does not have the greatest expected value, given the available information.

(**Trap**) Ten miners are trapped either in shaft A or B, but we cannot know which. We can block just one shaft from a threatening flood with sandbags, and if we do that, all the water will go into the other shaft and drown any miners inside. If we block neither shaft, both shafts will fill halfway with water, and just one miner will be drowned.

Kolodny and MacFarlane accept the following four sentences as true in (Trap):

- [1] We ought to block neither shaft.
- [2] If the miners are in shaft A, we ought to block shaft A.
- [3] If the miners are in shaft *B*, we ought to block shaft *B*.
- [4] Either the miners are in shaft A, or the miners are in shaft B.

They take it as "obvious" that the outcome of the deliberation should be ([1]). (Even if they do not explicitly say so, we may assume that the probability, given our evidence, that they are in A is approximately 0.5, and the same, of course, for the probability that they are in B.) At the same time, "in deliberating what to do", ([2]) and ([3]) seem "natural to accept". But the sentences ([2])–([4]) seem to entail:

[5] Either we ought to block shaft *A*, or we ought to block shaft *B*.

This is inconsistent, or at least yields a dilemma (viz. that we both ought and ought not to block one of the shafts), with ([1]). They show how the argument from ([2])–([4]) to ([5]) can be shown to be valid, with standard disjunction and e.g. material conditional (they do not explicitly assume any specific conditional: any connective validating modus ponens would do) (Kolodny & MacFarlane 2010, p.127). In the argument, they just use three well-known rules of classical logic to derive the problematic conclusion: modus ponens, and introduction and elimination of disjunction. Rejecting any of them does not seem *prima facie* plausible. Nevertheless, Kolodny's and MacFarlane's solution partly consists in proposing a new type of indicative conditional, for which modus ponens is severely restricted, after having discussed and rejected various maybe *prima facie* plausible solutions of the puzzle.

Some of their arguments against alternative solutions are of special interest for what follows. They consider an "objectivist" conception of ought, which would imply rejecting ([1]).

Objectivism "*S* ought (at *t*) to do ϕ " is true iff ϕ -ing is the best choice available to *S* in light of all the fact, known and unknown (Kolodny & MacFarlane 2010, p.117).

They accept a common line argument against moral views of this type, viz. that the objectivist "ought" becomes useless in deliberation under limited information. They also argue against the opposite "subjectivist" conception, which would imply rejecting ([2]) and ([3]).

Subjectivism "S ought (at t) to do ϕ " is true iff ϕ -ing is the best choice available to S in light of what S knows at t (Kolodny & MacFarlane 2010, p.118).

They claim that (Subjectivism) "cannot make good sense of the use of "ought in *advice*" (Kolodny & MacFarlane 2010, p.119), and they give an example with their Dialogue 1, where the deliberating Agent in the (Trap) is confronted with an Adviser, who knows that the miners are in fact in shaft A, and then would disagree with Agent's judgment that Agent ought to block neither shaft, and instead give the advice that A ought to block shaft A, which would be false according to (Subjectivism).

One might try to remove the conflict between the premises in (Trap) by interpreting the "ought" in ([1]) according to (Subjectivism), and the "ought" in ([2]), ([3]) and ([5]) according to (Objectivism) (Kolodny & MacFarlane 2010, p.120). Against this, they argue that we would have no genuine disagreement between Agent and Adviser in Dialogue 1, if we try to secure the truth of Adviser's advice statement that A ought to be blocked by interpreting this advice according to (Objectivism), while Agent's statement is interpreted according to (Subjectivism). They also give further examples, with their Dialogue 2, where Adivser's advice cannot be interpreted according to (Objectivism), but rather seems to refer to what is best in light of some limited information, which, however, is more inclusive than Agent's initial information (Kolodny & MacFarlane 2010, p.121).

2. An Alleged Solution

In order to solve the paradox, Kolodny and MacFarlane, present their own modal semantics, with a new interpretation of the epistemic and deontic modals, and an indicative conditional, for which modus ponens is only restrictedly valid (Kolodny & MacFarlane 2010, p.130–136).

I will not recapitulate the details of their semantics in this paper. The truth of a sentence ϕ is assigned relative to a world w and an information-state (set of worlds) *i*. For evaluation of the deontic and epistemic modalities, they use an operator, \Box_f indexed to a selection function.

 $\Box_{f}\phi^{\dagger}$ is true at $\langle w, i \rangle$ iff for all $w' \in f(i)$, ϕ is true at w' (Kolodny & MacFarlane 2010, p.131).

The selection function may be an epistemic function e that selects a set of world-states that might, as far as this state knows, be actual, or a deontic function d, that selects a set of, in some normative sense, ideal world-states, given an information-state.

Kolodny and MacFarlane also define an indicative conditional, after first defining the notions of truth throughout an information state, and a maximal ϕ -subset (Kolodny & MacFarlane 2010, p.135).

True Throughout ϕ is true throughout an information state *i*, iff for all $w \in i$, ϕ is true at $\langle w, i \rangle$.

Maximal ϕ -subset *i'* is a maximal ϕ -subset of *i* iff (a) for all $w \in i'$, ϕ is true throughout *i'*, and (b) there is no *i''* such that $i' \subset i'' \subseteq i$ and ϕ is true throughout *i''*.

if $[\text{if } \phi]\psi^{\uparrow}$ is true at $\langle w, i \rangle$ iff ψ is true at $\langle w, i' \rangle$ for every maximal ϕ -subset *i'* of *i*.

Kolodny and MacFarlane say that the normative judgment ([1]) would be ratified by a version of consequentialism telling us to maximize expected utility, but also by "most reasonable deontological and virtue theories" (Kolodny & MacFarlane 2010, p.115). In the situation where we do not know where the miners are, we can assume that we have an information state containing both worlds where they are in shaft A and worlds where they are in shaft B. Let *inX* stand for a proposition stating that the miners are in shaft A, and *blX* stand for a proposition that shaft X is blocked. In that case, the deontic selection function may select worlds where we do not block any of the shafts, i.e. $\Box_d \neg (blA \lor blB)$ will come out as true. On the other hand, when we evaluate a conditional norm, e.g. ([2]), we only use sets of worlds where they are in shaft A as arguments in the deontic selection function, and it then gives the result that we ought to block shaft A, i.e. $[if inA] \Box_d blA$, will come out as true.

But how are we to avoid reaching the unwanted conclusion ([5]), that we ought to block either A or B, by means of hypothetical modus ponens? Well,

if we assume that it is true at $\langle w, i \rangle$ that the miners are in shaft *A*, the information state *i* will still contain worlds where they are in shaft *B*, and we cannot use modus ponens to derive that we ought to block shaft *A*, because the deontic operator is sensitive to the whole information state.

However, if it is assumed that it is known that they are in shaft A, and that we, if so, ought to block shaft A, we can derive that we ought to block shaft A.

We can also make hypothetical derivations with apparently the same structure as the, in their system, invalid derivation from ([2])-([4]) to ([5]). They have the following example (Kolodny & MacFarlane 2010, p.141):

[6] If the miners are in shaft A, they have a jackhammer.

- [7] If the miners are in shaft *B*, they have a blowtorch.
- [8] Either the miners are in shaft A, or the miners are in shaft B.

[9] So, either they have a jackhammer or they have a blowtorch.

Kolodny and MacFarlane show that an argument like that from ([6])-([8]) to ([9]), where the consequents in ([6]) and ([7]), unlike the deontic consequents in ([2]) and ([3]), are not relative to the information state, is valid in their semantics (given that the antecedent is also invariant among either worlds or information-states) (Kolodny & MacFarlane 2010, p.141f).

One might ask how Kolodny's and MacFarlane's model is able to handle the problems with explaining moral disagreement that they claim create serious problem for the subjectivist and disambiguation solution. We must recall that there is a "relevant" information-state in a context where a sentence is assessed (Kolodny & MacFarlane 2010, p.141f). Agent and Adviser in the dialogues would thus be disagreeing about what is true given a world-state and a certain relevant information-state. They are unclear about what this relevant state would be in this situation, but it could presumably be something like the set of state-descriptions that might, given the maximal information available to Adviser and Agent (with the help of Adviser), depict the actual world.

3. Problems for the Model

3.1. A Problematic Case

As we have seen, Kolodny and MacFarlane intend their model to reconcile some, in their view, appealing features of subjectivist and objetivist consequentialist principles by making the deontic operator relative to

information states. But how helpful is it for that purpose? Let me describe a case where I claim they get into trouble.

(Fever) You wake up one morning, feeling feverish. If you are really running a fever, going to school will have worse consequences than staying at home, because you will then also infect some of your fellows. But if you are not running a fever, going to school will have better consequences than staying at home, because you are responsible for a seminar and will cause some disruption among your fellows if you do not go to school. Once you have decided about whether or not to go to school, you will not worry about your decision afterwards. Checking whether or not you are really running a fever is also somewhat tedious. So, if you in fact do not have a fever, it will have the best consequences if you go to school directly without taking your temperature. On the other hand, if you really have a fever, it will have the best consequences if you stay at home, again without taking your temperature. At waking up, you have evidence for all this.

Let us make the following value assumptions, given your evidence at waking up, where s: you go to school, t: you check your temperature, and f: you have a fever. (We ignore the alternatives where you go to school, even though you know that you have a fever, and those where you stay at home even though you know you do not.)

10. $V(f \land t \land \neg s) = -10$ 11. $V(f \land \neg t \land \neg s) = -9$ 12. $V(f \land \neg t \land s) = -50$ 13. $V(\neg f \land t \land s) = 9$ 14. $V(\neg f \land \neg t \land s) = 10$ 15. $V(\neg f \land \neg t \land \neg s) = 0$ 16. P(f) = 0.5

You can choose between the alternatives $t \land (s \leftrightarrow \neg f), \neg t \land s$ and $\neg t \land \neg s$. Given the value and probability assumptions above, you may calculate the following expected utilities (where EXV(ϕ) gives the expected value of state ϕ):

17. EXV
$$(t \land (s \leftrightarrow \neg f)) = -0.5$$

Informational Models in Deontic Logic

18. EXV($\neg t \land s$) = -20 19. EXV($\neg t \land \neg s$) = -4.5

So, expected utility consequentialism would tell you to check your temperature, and then go to school iff you have no fever. Kolodny's and MacFarlane's model would also give this result, with an expected utility-maximizing deontic selection function. But if we let such a function select the ideal worlds, given an input consisting of the worlds where *f* is true, it will select a subset of the $(\neg t \land \neg s)$ -worlds, and the conditional $[\text{if } f] \square_d (\neg t \land \neg s)$ will then come out as true, given how Koldony and MacFarlane have defined the indicative conditional. In the same way, the conditional $[\text{if } \neg f] \square_d (\neg t \land s)$ will come out as true.

The natural English equivalents of these would be something like:

[20] If you in fact have a fever, you should just avoid going to school, without bothering to take your temperature.

[21] If you in fact do not have a fever, you should just go to school, without bothering to take your temperature.

An objectivist consequentialist would, of course, welcome these as true in the situation (Fever). How could Kolodny and MacFarlane avoid them, given the way they argue for the truth of ([2]) and ([3]) in (Trap)? They cannot point to some constraint against taking risks, because this would also invalidate ([2]) and ([3]). But are ([20]) and ([21]) really something that "naturally occur in the course of deliberation", as Kolodny and MacFarlane say that the conditionals ([2]) and ([3]) in case (Trap) do? I do not think so.

Given another deontic selection function, where the negative value of t is disregarded, because (as many would say) negative effects of an action on the agent's own welfare cannot themselves bring about a moral prohibition, we can still say e.g. that if you have a fever, you *may* stay at home without checking your temperature. And we may construct a similar case when you are evaluating the health status of someone else, e.g. your child, who then would suffer the negative effects of the measuring.

Could we claim that ([20]) and ([21]) are not really problematic, because the relevant information-state in the evaluation of these sentences must be one where we have gathered all available information, and then, there would be no need to perform any more temperature-checking? But how are we, then, to evaluate a claim like $\Box_d t$, which is about an act of information-

gathering? Are we to say that any such claim is automatically true, that we always ought to gather as much information as we can? That is obviously not a viable option. Sometimes, an act of information-gathering is worth the costs, sometimes not, and any plausible selection function for moral norms must be sensitive to this.

However, I think that ([2]) and ([3]) may seem a bit more natural to assert in the case (Trap) than ([20]) and ([21]) in case (Fever), but I also think I can explain why.

When we try to calculate expected value, the first thing to do is to estimate the values of the different possible outcomes, and we may start by making some sort of ordinal scale of them. In such a case, we may use "ought" as interchangeable with "having the best outcome", e.g. instead of ([2]), we could say "It will have the best consequences to block shaft A, given that they are there." ([20]) or ([21]) do not really seem useful in deliberation, because the truth of their consequents exclude our knowing the antecedents, and because we know beforehand that the difference between t and $\neg t$ do not make a very great a difference in value, given that f and s are held constant. If t were something very bad, so that it would outweigh the expected benefits of gaining certain knowledge, and we could gain (less certain) evidence about f in some other way, we would perhaps be prepared to assert ([20]) or ([21]). We may in a situation like (Fever) say that ([20]) and ([21]) are, in some sense, true, but that they are not the kind of norm we are interested in, in such a situation.

([20]) and ([21]) may seem strange to assert for pragmatic reasons, because the conjunct $\neg t$ in the scope of the operator consequent does not change if we change the antecedent. Why not just assert that you should go to school if you do not have a fever, and stay home if you have a fever? Is this inassertability everything my example shows? Let us consider a slightly revised version of (Fever).

(Fever*) The situation is as in (Fever), with the exception that you know that if you stay at home without checking your temperature, you will feel worried about missing school, since you might have been able to go to school without infecting your fellows. On the other hand, if you go to school, you will not worry about your fellows. Assume that, in (Fever*), $V(f \land \neg t \land \neg s) = -15$.

Informational Models in Deontic Logic

The particular empirical assumptions in (Fever) may of course be changed indefinitely, but the important thing about (Fever*) is that Koldony's and MacFarlane's model runs into trouble, even though the pragmatic reasons that may be used to try to explain the apparent oddity of ([20]) and ([21]) in (Fever) do not apply in (Fever*). (Fever*) differs from (Trap) and (Fever) also in that the alternative with highest expected value, viz. to check your temperature and go to school iff you have no fever, might also be the action with highest actual value. ([21]) will come out as true, even in (Fever*), but ([20]) will come out as false. Instead, we have

[22] If you in fact have a fever, you should take your temperature and avoid going to school.

Making the question whether or not you should check your temperature depend on whether or not you in fact have a fever does not seem very plausible in a deliberative context, and the reply "It is not the right kind of should!" seems appropriate. Even in (Trap), the conjunction of any of, or both, conditionals ([2]) and ([3]) with ([1]) sound strange: "We ought to block neither shaft, but if they are in shaft A, we ought to block that shaft, and if they are in shaft B, we ought to block that shaft."

Note that, if we use an indicative conditional, \rightarrow , validating modus ponens, we must, if we hold $\Box_d t$ true in (Fever), also hold $f \rightarrow \Box_d t$ and $\neg f \rightarrow \Box_d t$ true. Would not they also be strange to assert, and unnatural to think of in deliberation, just like ([20]) and ([21])? Yes, they would be unnatural to assert in isolation, but they may nevertheless be embedded in an emphasizing context, like

[23] Maybe you do not have a fever, but even then, you should check, so that you can avoid taking unnecessary risks.

What happens if we just evaluate a conditional like $[if f]\Box_d(t)$, where the truth-value *s* is not specified, in the scope of the deontic operator, in Kolodny's and MacFarlane's model? It should come out as false, because what you ought to do, given *f*, involves $\neg t$. But it might be argued that if you do not check your temperature, you might end up going to school and infecting your fellows, so that *t* has greater expected utility than $\neg t$, given *f*. This seems to be a case of the well-known Chisholm Paradox (Chisholm 1963) that gives rise to problems in many (if not all) systems of deontic logic,

where entering into an optimal action-pattern that is not fulfilled may be worse than entering into a sub-optimal action-pattern. I will not discuss this issue any further here.

So, the above considerations suggest that Kolodny and MacFarlane are wrong when they claim that there is a unified information-sensitive sense of "ought", that differs both from traditional objectivist and subjectivist conceptions that interacts with an information-sensitive indicative conditional the way they argue. Instead, we should opt for something like a disambiguation solution of their paradox, i.e. that "ought" in some contexts is ambiguous between an objective and a subjective sense, so that ([1]) in (Trap) should be interpreted as referring to a subjective "ought", and that ([2]) and ([3]) should be interpreted as referring to an objective "ought".

3.2. Possible Problems with my Explanation

As we have seen, Kolodny and MacFarlane have argued against disambiguation solutions of the type I propose. They claim that neither subjectivism nor disambiguation can adequately handle advice situations like their Dialogue 1 or Dialogue 2, where the agent who thinks that both shafts ought to be left open is confronted with an adviser, who has more information about the location of the miners or other features of the context, and seems to disagree with the "ought"-statements made by the agent (Kolodny & MacFarlane 2010, p.119–121). What am I to say about this?

First, we should note that what I have claimed is that the consequents in ([2]) and ([3]) in (Trap) may be best explained as referring to some objective "ought". I am not sure that Adviser's statement in Dialogue 1, that the agent ought to block shaft A, where Adviser knows where the miners are, should be explained that way. Rather, Adviser's claim might be that given Adviser's evidence, blocking shaft A is the best thing to do. In this way, we could also, I think, make sense of Dialogue 2, where Adviser does not seem to refer to some objective "ought".

We may recall Kolodny's and MacFarlane's agent-relative definition of subjectivism:

Subjectivism "S ought (at t) to do ϕ " is true iff ϕ -ing is the best choice available to S in light of what S knows at t (Kolodny & MacFarlane 2010, p.118).

Another proposal would be that when we assert what S ought to do, we just make a claim about what would be the best choice (e.g. in the expected utility-sense) for S to do, given *our* current body of knowledge (or evidence). Such a variant of subjectivism seems more suitable to an advice context. It can make sense of the claim that Adviser really believes that Agent ought to do what would not be best in light of Agent's initial knowledge (evidence) in these situation, which would solve one problem Kolodny and MacFarlane point out with subjectivism, according to their definition, viz. that it cannot explain why Adviser really believes that Agent ought to block shaft A in Dialogue 1. In some contexts, e.g. when we try to judge afterwards whether or not an agent has acted wisely, it may be more appropriate to use a version of subjectivism more like Kolodny's and MacFarlane's.

However, Kolodny and MacFarlane would probably complain against my "speaker-relative" subjectivism, that we would have no genuine disagreement in Dialogue 1 or Dialogue 2, in the sense that Agent and Adviser would be asserting incompatible claims about what action Agent should perform: instead they would make claims about what action would be best relative to their different bodies of evidence.

To this I answer that, in both Dialogue 1 and Dialogue 2, Adviser and Agent would have initially incompatible (in the sense of not jointly satisfiable) conceptions on what line of action will maximize expected utility (relative to their belief systems), and listening to Adviser may also have greater expected utility, given Agent's belief system, than not listening. Do we really need more than that to make sense of such advice situations as they describe? They point out that Adviser would not have reason to give advice if Adviser knows that Agent do not trust Adviser. But I do not really understand why this should be a problem. We do not give advice to people that we are sure do not listen to us.

Kolodny's and MacFarlane's main argument against the objectivist type of ought is that it seems useless in deliberation, because we cannot know the actual outcomes. But does their own theory, where the deontic status of propositions is related to a certain "relevant" information-state, fare much better in that respect? They do not say much about how to make sense of the idea of an information-state being "relevant", but they will probably get into problems no matter how they interpret this notion. As e.g. Fred Feldman has noted (Feldman 2006), determining the expected utilities of all one's alternatives is also practically impossible in many situations. Given just our current evidence, or some evidence we could easily acquire, we may not be

able to form any determined conception at all of the expected utilities of our different alternatives, and even if we could do that given some larger body of evidence, gathering all that information and performing all the necessary calculations may be extremely time- and energy-consuming. Moving to expected utility consequentialism is thus no solution to the general problem that objective consequentialism often does not give any action-guidance.

My "speaker-relative" subjectivism, and their "agent-relative" subjectivism are also vulnerable to these problems. If we hold that both these oughts are relevant in different contexts, this may also seem to give rise to a disturbing multitude of different oughts. The most plausible alternative may be to say that the objective ought after all is the primary criterion of rightness for actions, and that the subjective oughts are secondary to that e.g. as to be used in some, rather special, contexts as decision procedures aiming at maximizing actual utility. In certain situations, like Kolodny's and MacFarlane's (Trap), where we have a rather limited set of alternatives with not very many different likely outcomes (in relation to the time and resources we have left to decide), it may after all be appropriate to perform expected utility calculations on those alternatives. If the subjective oughts are just decision procedures, and not criterions of rightness, it may also seem less disturbing that we do not have any genuine disagreement in advice situations such as those described in Kolodny's and MacFarlane's dialogues. It is noteworthy that Koldony and MacFarlane never mention this distinction betweeen decision procedures and criterions of rightness, which is commonly made, especially in the debate about consequentialist ethics, and often has been invoked in order to meet the challenge against consequentialism that it fails to give action-guidance.

We should note that Kolodny and MacFarlane argue against subjectivism both that it cannot make sense of disagreement, and that it will invalidate the conditionals ([2]) and ([3]). If we want to keep genuine disagreement in situations like their dialogues, but reject their indicative conditional, perhaps because it, as I have argued, leads to unacceptable results in situations like (Fever), it should be possible to combine a deontic operator that is information-variant the way they argue with an indicative conditional, such as the ordinary material conditional, that is not directly information-variant (other than through the evaluation of its antecedent or consequent).

We may also note that Kolodny's and MacFarlane's information-variant deontic operator do not seem to rule out objective consequentialism a priori. If we want to have objective consequentialism, it seems that we can specify that the relevant information-state for the evaluation of a moral sentence should just contain one world for each alternative, viz. the world that (given some counterfactual determinism) would be realized by that alternative.

4. Concluding remarks

I conclude that Kolodny and MacFarlane have given no strong reasons in support of their claim that adopting their informational model semantics as a whole would be beneficial to deontic logic, because it would be better suited than standard semantics to capture some common forms of moral reasoning. I have tried to show that their indicative conditional leads to implausible results in some contexts of moral reasoning, and that the examples where it seems to give plausible results can be explained without adopting it. I have proposed an explanation of the apparent paradox in cases like (Trap) consisting in disambiguating between different sense of "ought", and argued that Kolodny and MacFarlane have dismissed explanations of this type prematurely.

Acknowledgements

Drafts of this text have been presented at philosophy seminars at Stockholm University and the Royal Institute of Technology. I would like to thank the participants at these seminars, in particular Erik Carlson, Karin Enflo, Sven Ove Hansson, Sofia Jeppsson, Jonas Olson, Tor Sandqvist and Torbjörn Tännsjö for useful comments.

References

- Chisholm, R. M. (1963). Contrary-to-Duty imperatives and deontic logic. *Analysis*. 24. pp. 33–36.
- Feldman, F. (2006). Actual utility, the objection from impracticality, and the move to expected utility. *Philosophical Studies*. 129. pp. 49–79.
- Jackson, F. (1991). Decision-Theoretic consequentialism and the nearest and dearest objection. *Ethics*. 101. pp. 461–482.
- Kolodny, N. & MacFarlane, J. (2010). Ifs and oughts. *Journal of Philosophy*. 107. pp. 115–143.

Karl Pettersson Department of Philosophy Uppsala University karl.pettersson@filosofi.uu.se