Introduction

This collection reprints all my previously published papers in ethics and social philosophy, except for those that were previously reprinted in another collection, *Philosophical Papers*.[1] I have taken the opportunity to correct typographical errors and editorial alterations. But I have left the philosophical content as it originally was, rather than trying to rewrite the papers as I would write them today.

The first three papers deal with the deontic logic of obligation and permission. Such a system of logic, in which operators of obligation and permission are taken to be dual modal operators analogous to operators of necessity and possibility, can be extended to what is obligatory or permissible given some condition. 'Semantic Analyses for Dyadic Deontic Logic' surveys a number of published treatments of conditional obligation and permission with a view to separating substantive differences – different degrees of generality, as it turns out – from mere differences between equivalent styles of bookkeeping.

The deontic logic of permission (whether conditional or unconditional) ignores the performative character of permission. By saying that something is or isn't permitted (unconditionally or conditionally) we can make it so. But there's a complication. If I say that *some* of the courses of action in which so-and-so happens are permissible, saying so makes it so. But *which* of those courses of action do I thereby

1 David Lewis, *Philosophical Papers*, volumes I and II (Oxford University Press, 1983 and 1986).

bring into permissibility? 'A Problem about Permission' surveys various possible answers.

In 'Reply to McMichael', I insist that deontic logic, conditional or otherwise, characterizes only the formalities of moral thinking. What substantive conclusions come out will depend on what substantive assumptions went in.

The next three papers concern belief, desire, and decision. In 'Why Ain'cha Rich?' I examine the embarrassing fact that the choices I endorse as rational in Newcomb's problem are the choices that foreseeably lead to bad outcomes. Those who think as I do explain away this embarrassing fact thus: if a powerful predictor sets out to reward predicted irrationality, then it is only to be expected that the rewards will go to the irrational. I ask whether this remark is common ground between both sides of the dispute about what is rational, and I conclude with regret that it is not.

In 'Desire as Belief' and its sequel, I explore the consequence of supplementing standard decision theory with various versions of the assumption that desires are identical to, or are necessarily correlated with, beliefs about what would be good. For some versions, we get a collapse into triviality; for other versions, a collapse into the view that there are some things – what? – that we desire by necessity.

The next paper, 'Dispositional Theories of Value', defends a subjectivist position in meta-ethics: a form of naturalism according to which values are defined as those properties that we are disposed, given a certain degree of empathetic understanding, to desire to desire. The position defended is similar to one that G. E. Moore chose as a target for his 'naturalistic fallacy' argument.

Moral dilemmas in consequentialist and deontological ethics are much discussed; but similar dilemmas can arise also in virtue ethics. In 'The Trap's Dilemma' I discuss Ned Kelly's proof that a policeman cannot be an honest man: because if a policeman has sworn an oath to obtain a conviction if possible, and if the only effective way to do so is to swear a lie, then the policeman is dishonest whether he keeps his oath or whether he breaks it.

'Evil for Freedom's Sake?' explores free-will theodicy as a reply to the problem of evil. It turns out that after we grant several points to

the proponent of free-will theodicy for the sake of the argument, we end up bogged down inconclusively in some complicated double counterfactuals. The deadlocked issues bear a striking resemblance to the well-known deadlock we encounter in Newcomb's problem.

'Do We Believe in Penal Substitution?' suggests that we are of two minds on the question of penal substitution. Mostly, we would think it absurd to let an offender go unpunished just because some innocent substitute has volunteered to be punished in his place. So when some Christians explain the Atonement as a case of penal substitution, that seems much at odds with our ordinary thinking. But our ordinary thinking is ambivalent. Though we would think it absurd to allow penal substitution when the punishment is a prison sentence, we think it not amiss if a generous volunteer pays someone else's fine – even if the fine is big enough to be no less onerous than a prison sentence.

In the next two papers, rejoinders respectively to Dale Jamieson and John Hawthorne, I defend and elaborate my account of conventions generally, and conventions of language in particular.[2]

The next paper, 'Illusory Innocence?' seeks to avoid Peter Unger's incredible conclusion that, so long as there is an inexhaustible supply of distant strangers whom we could rescue from urgent need at small cost to ourselves, we are obliged to give almost all we have – and all we can beg, borrow, or steal – to assist them.

'Mill and Milquetoast' argues that liberal customs of toleration may be seen as a tacit treaty to limit warfare between factions: if, for each side, the fear of defeat outweighs the hope of victory, it may be best for all concerned to settle for a stalemate. 'Academic Appointments: Why Ignore The Advantage of Being Right?' asks why we seem to think it wrong to deny academic appointments to candidates simply on the grounds that their views are false. I answer that this custom, again, may be seen as a tacit treaty to limit intellectual warfare. The banned weapon, denial of employment, would make wars more

2 David Lewis, *Convention: A Philosophical Study* (Harvard University Press, 1969); 'Languages and Language' in *Minnesota Studies in the Philosophy of Science*, Volume VII, ed. by Keith Gunderson (University of Minnesota Press, 1975).

costly for all concerned, yet would not give advantage to any one side – hence would not advance the side of truth, whichever side that may be.

'Devil's Bargains and the Real World' and 'Buy Like a MADman, Use Like a NUT' argue against 'paradoxical' nuclear deterrence: the idea that the only effective way – or the most benign way! – to conduct nuclear deterrence is to cultivate an irrational disposition to respond to attack by inflicting vast and useless harm. Although these papers were originally written with reference to the nuclear confrontation between the United States and the former Soviet Union, I fear they are not obsolete. It is all too likely that future history will contain other similar confrontations.

'The Punishment that Leaves Something to Chance' examines a puzzle about punishment. Why do we punish failed attempts at murder more leniently than successful attempts, although it would seem that they are no less wicked? My answer is that the principal punishment is probabilistic: he who subjects his victim to a risk of death is punished by a like risk – the risk that the victim will die, and the perpetrator will thereby earn the full punishment for a successful attempt. Whether probabilistic punishment is just is, however, open to question.

In the final paper, 'Scriven on Human Unpredictability', Jane S. Richardson and I reply to Michael Scriven's argument that, even in a deterministic world, those who so wish have a sure-fire strategy to avoid being predicted: they can replicate the predictions others might make about them, and then do the opposite. We object that Scriven cannot consistently combine all the assumptions he needs in order to argue both that this method of avoiding prediction will work, and that it shows something more interesting than just that prediction will fail if the would-be predictor runs out of time to finish his calculations.

David Lewis
Princeton, May 1998

1

Semantic analyses for dyadic deontic logic

1. INTRODUCTION

It ought not to be that you are robbed. *A fortiori*, it ought not to be that you are robbed and then helped. But you ought to be helped, given that you have been robbed. The robbing excludes the best possibilities that might otherwise have been actualized, and the helping is needed in order to actualize the best of those that remain. Among the possible worlds marred by the robbing, the best of a bad lot are some of those where the robbing is followed by helping.

In this paper, I am concerned with semantic analyses for dyadic deontic logic that embody the idea just sketched. Four such are known to me: the treatments in Bengt Hansson [4], Sections 10–15; in Dagfinn Føllesdal and Risto Hilpinen [2], Section 9; in Bas van Fraassen [9]; and in my own [8], Section 5.1.[1] My purpose here is to place these four treatments within a systematic array of alternatives,

First published in Sören Stenlund, ed., *Logical Theory and Semantic Analysis: Essays Dedicated to Stig Kanger on His Fiftieth Birthday* (Dordrecht, Reidel, 1974).

This research was supported by a fellowship from the American Council of Learned Societies.

1 Some other treatments of dyadic deontic logic fall outside the scope of this paper because they seem, on examination, to be based on ideas quite unlike the one I wish to consider. In particular, see the discussion in [4], [2], and [9] of several systems proposed by von Wright and by Rescher.

and thereby to facilitate comparison. There are superficial differences galore; there are also some serious differences.

My results here are mostly implicit in [8], and to some extent also in [7]. But those works are devoted primarily to the study of counterfactual conditionals. The results about dyadic deontic logic that can be extracted thence *via* an imperfect formal analogy between the two subjects are here isolated, consolidated, and restated in more customary terms.

II. LANGUAGE

The language of dyadic deontic logic is built up from the following vocabulary: (1) a fixed set of sentence letters; (2) the usual truth-functional connectives $\top$, $\bot$, $\sim$, $\&$, $\vee$, $\supset$, and $\equiv$ (the first two being zero-adic 'connectives'); and (3) the two dyadic deontic operators $O(-/-)$ and $P(-/-)$, which we may read as *'It ought to be that . . . , given that . . . '* and *'It is permissible that . . . , given that . . . '*, respectively. They are meant to be interdefinable as follows: either $P(A/B) =^{df} \sim O(\sim A/B)$ or else $O(A/B) =^{df} \sim P(\sim A/B)$. Any sentence in which $O(-/-)$ or $P(-/-)$ occurs is a *deontic sentence*; a sentence is *iterative* iff it has a subsentence of the form $O(A/B)$ or $P(A/B)$ where A or B is already a deontic sentence. (We regard a sentence as one of its own subsentences.) In metalinguistic discourse, as exemplified above, vocabulary items are used to name themselves; the letters early in the alphabet, perhaps subscripted, are used as variables over sentences; and concatenation is represented by concatenation.

III. INTERPRETATIONS

$[\![\ ]\!]$ is an *interpretation* of this language *over* a set I iff (1) $[\![\ ]\!]$ is a function that assigns to each sentence A a subset $[\![A]\!]$ of I, and (2) $[\![\ ]\!]$ obeys the following conditions of standardness:

(2.1) $[\![\top]\!] = I$,
(2.2) $[\![\bot]\!] = \varnothing$,
(2.3) $[\![\sim A]\!] = I - [\![A]\!]$,
(2.4) $[\![A \,\&\, B]\!] = [\![A]\!] \cap [\![B]\!]$,

(2.5) $[\![A \vee B]\!] = [\![A]\!] \cup [\![B]\!]$,
(2.6) $[\![A \supset B]\!] = [\![\sim A \vee B]\!]$,
(2.7) $[\![A \equiv B]\!] = [\![(A \supset B) \,\&\, (B \supset A)]\!]$
(2.8) $[\![P(A/B)]\!] = [\![\sim O(\sim A/B)]\!]$.

We call $[\![A]\!]$ the *truth* set of a sentence A, and we say that A is *true* or *false* at a member i of I (*under* the interpretation $[\![\ ]\!]$) according as i does or does not belong to the truth set $[\![A]\!]$.

We have foremost in mind the case that I is the set of all possible worlds (and we shall take the liberty of calling the members of I *worlds* whether they are or not). Then we can think of $[\![A]\!]$ also as the proposition expressed by the sentence A (under $[\![\ ]\!]$): an interpretation pairs sentences with propositions, a proposition is identified with the set of worlds where it is true, and a sentence is true or false according as it expresses a true or false proposition.

The sentences of the language are built up from the sentence letters by means of the truth-functional connectives and the deontic operators. Likewise an interpretation is determined stepwise from the truth sets of the sentence letters by means of the truth conditions for those connectives and operators. (2.1–7) impose the standard truth conditions for the former. (2.8) transforms truth conditions for $O(-/-)$ into truth conditions for $P(-/-)$, making the two interdefinable as we intended. The truth conditions for $O(-/-)$ have so far been left entirely unconstrained.

IV. VALUE STRUCTURES

Our intended truth conditions for $O(-/-)$ are to depend on a posited structure of evaluations of possible worlds. We seek generality, wherefore we say nothing in particular about the nature, source, or justifiability of these evaluations. Rather, our concern is with their structure. A mere division of worlds into the ideal and the less-than-ideal will not meet our needs. We must use more complicated value structures that somehow bear information about comparisons or gradations of value.

An interpretation is *based*, *at* a particular world, *on* a value structure iff the truth or falsity of every sentence of the form $O(A/B)$, at that

world and under that interpretation, depends in the proper way on the evaluations represented by the value structure.

Let $[\![\ ]\!]$ be an interpretation over a set I, and let i be some particular world in I. In the case we have foremost in mind, I really is the set of all possible worlds; and i is our actual world, so that truth at i is actual truth, or truth *simpliciter*. We consider value structures of four kinds.

First, a *choice function f over I* is a function that assigns to each subset X of I a subset fX of X, subject to two conditions: (1) if X is a subset of Y and fX is nonempty, then fY also is nonempty; and (2) if X is a subset of Y and X overlaps fY, then $fX = X \cap fY$. $[\![\ ]\!]$ is *based, at i, on* a choice function f over I iff any sentence of the form $O(A/B)$ is true at i under $[\![\ ]\!]$ iff $f[\![B]\!]$ is a nonempty subset of $[\![A]\!]$. Motivation: fX is to be the set of the best worlds in X. Then $O(A/B)$ is true iff, non-vacuously, A holds throughout the B-worlds chosen as best.

Second, a *ranking $\langle K, R\rangle$ over I* is a pair such that (1) K is a subset of I; and (2) R is a weak ordering of K. R is a *weak ordering*, also called a *total preordering*, of a set K iff (1) R is a dyadic relation among members of K; (2) R is transitive; and (3) for any j and k in K, either jRk or kRj – that is, R is *strongly connected* on K. $[\![\ ]\!]$ is *based, at i, on* a ranking $\langle K, R\rangle$ over I iff any sentence of the form $O(A/B)$ is true at i under $[\![\ ]\!]$ iff, for some j in $[\![A \,\&\, B]\!] \cap K$, there is no k in $[\![\sim A \,\&\, B]\!] \cap K$ such that kRj. Motivation: K is to be the set of worlds that can be evaluated – perhaps some cannot be – and kRj is to mean that k is at least as good as j. Then $O(A/B)$ is true iff some B-world where A holds is ranked above all B-worlds where A does not hold.

Third, a *nesting $\$$ over I* is a set of subsets of I such that, whenever S and T both belong to $\$$, either S is a subset of T or T is a subset of S. $[\![\ ]\!]$ is *based, at i, on* a nesting $\$$ over I iff any sentence of the form $O(A/B)$ is true at i under $[\![\ ]\!]$ iff, for some S in $\$$, $S \cap [\![B]\!]$ is a nonempty subset of $[\![A]\!]$. Motivation: each S in $\$$ is to represent one permissible way to divide the worlds into the ideal ones (those in S) and the non-ideal ones. Different members of $\$$ represent more or less stringent ways to draw the line. Then $O(A/B)$ is true iff there is some permissible way to divide the worlds on which, non-vacuously, A holds at all ideal B-worlds.

Fourth, an *indirect ranking $\langle V, R, \dagger\rangle$ over I* is a triple such that (1) V

is a set; (2) R is a weak ordering of V (defined as before); and (3) $\mathfrak{f}$ is a function that assigns to each j in I a subset $\mathfrak{f}(j)$ of V. $[\![\ ]\!]$ is *based, at i, on* an indirect ranking $\langle V, R, \mathfrak{f}\rangle$ iff any sentence of the form $O(A/B)$ is true at i under $[\![\ ]\!]$ iff, for some v in some $\mathfrak{f}(j)$ such that j belongs to $[\![A \,\&\, B]\!]$, there is no w, in any $\mathfrak{f}(k)$ such that k belongs to $[\![\sim A \,\&\, B]\!]$, such that wRv. Motivation (first version): V is to be a set of 'values' realizable at worlds; wRv is to mean that w is at least as good as v; and $\mathfrak{f}(j)$ is to be the set of values realized at the world j. Then $O(A/B)$ is true iff some value realized at some B-world where A holds is ranked higher than any value realized at any B-world where A does not hold. Motivation (second version): we want a ranking of worlds in which a single world can recur at more than one position – much as Grover Cleveland has two positions in the list of American presidents, being the 22nd and also the 24th. Such a 'multipositional' ranking cannot be a genuine ordering in the usual mathematical sense, but we can represent it by taking a genuine ordering R of an arbitrarily chosen set V of 'positions' and providing a function $\mathfrak{f}$ to assign a set of positions – one, many, or none – to each of the objects being ranked. Then $O(A/B)$ is true iff some B-world where A holds, in some one of its positions, is ranked above all B-worlds where A does not hold, in all of their positions.

The *value structures over I* comprise all four kinds: all choice functions, rankings, nestings, and indirect rankings over I. Note that (unless I is empty) nothing is a value structure of two different kinds over I.

An arbitrary element in our truth conditions must be noted. A value structure may ignore certain *inevaluable* worlds: for a choice function f, the worlds that belong to no fX; for a ranking $\langle K, R\rangle$, the worlds left out of K; for a nesting $\$$, the worlds that belong to no S in $\$$; and for an indirect ranking $\langle V, R, \mathfrak{f}\rangle$, the worlds j such that $\mathfrak{f}(j)$ is empty. Suppose now that B is true only at some of these inevaluable worlds, or that B is impossible and true at no worlds at all. Then $O(\text{-}/B)$ and $P(\text{-}/B)$ are *vacuous*. We have chosen always to make $O(A/B)$ false and $P(A/B)$ true in case of vacuity, but we could just as well have made $O(A/B)$ true and $P(A/B)$ false. Which is right? Given that $0 = 1$, ought nothing or everything to be the case? Is everything or nothing permissible? The mind boggles. As for formal elegance,

either choice makes complications that the other avoids. As for precedent, van Fraassen has gone our way but Hansson and Føllesdal and Hilpinen have gone the other way. In any case, the choice is not irrevocable either way. Let $O'(-/-)$ and $P'(-/-)$ be just like our pair $O(-/-)$ and $P(-/-)$ except that they take the opposite truth values in case of vacuity. The pairs are interdefinable: either let $O'(A/B) =^{df} O(\top/B) \supset O(A/B)$ or else let $O(A/B) =^{df} \sim O'(\bot/B)\ \&\ O'(A/B)$.

V. TRIVIAL, NORMAL, AND UNIVERSAL VALUE STRUCTURES

There exist *trivial* value structures, of all four kinds, in which every world is inevaluable. We might wish to ignore these, and use only the remaining non-trivial, or *normal*, value structures. Or we might go further and use only the *universal* value structures with no inevaluable worlds at all. It is easily shown that a value structure is normal iff, under any interpretation based on it at any world i, some sentence of the form $O(\top/B)$ is true at i. (And if so, then in particular $O(\top/\top)$ is true at i.) Likewise, a value structure is universal iff, under any interpretation based on it at any world i, any $O(\top/B)$ is true at i except when B is false at all worlds.

VI. LIMITED AND SEPARATIVE VALUE STRUCTURES

The *limited* value structures are, informally, those with no infinitely ascending sequences of better and better and better worlds. More precisely, they are: (1) all choice functions; (2) all rankings $\langle K, R\rangle$ such that every nonempty subset X of K has at least one *R-maximal element*, that being a world j in X such that jRk for any k in X; (3) all nestings $\$$ such that, for any nonempty subset $\mathbf{S}$ of $\$$, the intersection $\cap\mathbf{S}$ of all sets in $\mathbf{S}$ is itself a member – the smallest one – of $\mathbf{S}$; and (4) all indirect rankings $\langle V, R, \mathfrak{f}\rangle$ such that, if we define the *supersphere* of any v in V as the set of all worlds j such that wRv for some w in $\mathfrak{f}(j)$, then for any nonempty set $\mathbf{S}$ of superspheres, the intersection $\cap\mathbf{S}$ of all sets in $\mathbf{S}$ is itself a member of $\mathbf{S}$. Clearly some but not all rankings, some but not all nestings, and some but not all indirect