9. An Expected-Value Approach to the Dual-Use Problem

Thomas Douglas

In this chapter I examine how expected-value theory might inform responses to what I call the dual-use problem. I begin by defining that problem. I then outline a procedure, which invokes expected-value theory, for tackling it. I first illustrate the procedure with the aid of a simplified schematic example of a dual-use problem, and then describe how it might also guide responses to more complex real-world cases. I outline some attractive features of the procedure. Finally, I consider whether and how the procedure might be amended to accommodate various criticisms of it. The aim is not to defend the procedure in its original form, but to consider how far we must deviate from it to evade the criticisms that I consider. I seek to show that, though it is necessary to make substantial ammendments to the procedure, we need not eschew a role for expected-value theory altogether. Even if one accepts the criticisms I discuss, it is possible to defend a role for expected-value theory in responding to dual-use problems.

The dual-use problem

Scientific work can typically be used in both good and bad ways, and sometimes it is difficult or impossible for a scientist to prevent the bad applications without also forestalling the good ones. For example, if a scientist publishes her findings, thus enabling good applications of those findings, she may then have little power to prevent bad applications of the same results. Scientists may thus face a trade-off between enabling certain good applications of their work and preventing certain bad ones. In a few cases, the risk of bad applications may be so high that there is a genuine dilemma about how to resolve this trade-off. Consider the following hypothetical scenario:

> It's 2030. Designer viruses are now used to treat some cancers and infectious diseases. But they're expensive and difficult to manufacture. You discover a new, cheap way to produce synthetic viruses using out-of-date benchtop DNA synthesisers that are now ubiquitous, even in developing countries. You're excited about the discovery and hoping to publish it in *Nature*. You think it could bring a wide range of medical treatments, not to mention research tools, within the grasp of the

> developing world. However, there's a catch: every major military and terrorist group in the world has access to these obsolete synthesisers. It would take only one malevolent agent and one such machine to produce enough vaccine-resistant smallpox virions to devastate humanity.[1]

Arguably, it is genuinely unclear in this case whether you ought to publish your discovery. If this is so, you may be faced with what has been termed the *dual-use dilemma*.

As a tentative starting point, we might define the dual-use dilemma as the quandary that arises whenever: 1) a scientist faces a choice about whether or not to produce or disseminate some scientific output, such as a piece of scientific knowledge, or a physical technology; 2) that output could be used in both good and bad ways, and the agent is not in a position to altogether eliminate the risk of bad use without also reducing the likelihood of good use; 3) there is therefore a trade-off between enabling the good application(s) and preventing the bad one(s); and 4) it is consequently unclear whether producing the scientific output is the right thing to do, morally speaking.

This definition captures some commonplace ideas about dual-use dilemmas; however, there are, I think, good reasons to broaden the definition somewhat before embarking on any ethical analysis.

First, though the dual-use dilemma is typically presented as a problem that arises *for scientists* and *in relation to particular scientific projects*, similar problems may arise for others in a position to support or impede particular projects, and for those faced with decisions about what general stance or policy to take on scientific work, or some particular class of scientific work. For example, a policymaker considering whether to institute censorship of scientific journals may also face a trade-off between enabling good applications of scientific findings and preventing bad ones, though in this case the choice does not relate to any particular scientific project. Given the similar structure of this kind of ethical problem to paradigmatic examples of the dual-use dilemma faced by individual scientists, it seems sensible to consider them together.

Second, as defined above, the dual-use dilemma arises because of a tension between *two* ethical considerations: reasons to enable good applications of some scientific output, and reasons to prevent bad ones. But in real-world cases, there will be a number of other ethical considerations bearing on decisions relating to the production or dissemination of scientific outputs.[2] For example, a scientist

[1] Adapted from Douglas, T. and Savulescu, J. 2010, 'Synthetic biology and the ethics of knowledge', *Journal of Medical Ethics*, vol. 36, no. 11, pp. 687–93.

[2] Buchanan, A. and Kelley, M. 2013, 'Biodefense and the production of knowledge: rethinking the dual-use problem', *Journal of Medical Ethics*, vol. 39, pp. 195–204.

could have a reason to pursue scientific knowledge about some phenomenon regardless of how that knowledge is likely to be used; she could have a reason to pursue the knowledge because it would be intrinsically valuable, or simply because she previously promised to do so. Similarly, a policymaker could have reasons (not) to censor scientific publications that are independent of how the censored information would have been used—for example, she might have reasons not to censor the information in order to respect freedom of speech. An adequate formulation of the dual-use dilemma should accommodate the possible existence of reasons besides those normally taken to generate the dilemma.

Third, according to the above definition, the dual-use dilemma involves making a choice between exactly two alternatives. This feature is suggested by the term *dilemma*. But it might be thought that such binary choices are in fact rather rare. Often when we think there are only two alternatives on the table there are actually more. For example, a scientist may think she is faced with a choice between either publishing her research in full or not, when in fact there may be several other options—for example, publishing elements of the research, or publishing the research conditional upon its distribution being restricted. Similarly, a policymaker may believe that there is a simple choice between censoring scientific journals and not, when in fact there are many different types of censorship (for example, self-censorship by journal editors, censorship by peer reviewers, censorship by external agencies). It is tempting to say that, when presented with a dual-use dilemma, an agent should in the first instance try to find a way to *escape* the dilemma by finding some third, superior course of action. Only if this is impossible should he grasp one horn of the dilemma.

Recognising the existence of further alternatives, however, does not necessarily resolve the problem. Strictly speaking, when there are more than two alternatives available, there can be no *dilemma*. However, there may still be a quandary, since alternatives more likely to enable good applications may also be more conducive to bad applications or be more problematic in other ways (for example, they may be more resource intensive). We can imagine a case in which the more likely an option is to allow good applications relative to alternative options, the more likely it is to also enable bad applications. It may thus remain unclear which option to select.

I wish to include within the scope of the discussion that follows ethical problems that are similar to paradigmatic dual-use dilemmas but are not faced by individual scientists, do not relate to specific scientific projects, involve choices between more than two alternatives, and involve ethical considerations besides reasons to enable good and prevent bad applications of scientific outputs. I will do this by introducing the broader idea of a *dual-use problem*. I define this as the quandary that arises whenever: 1) an agent faces a choice between two or more alternatives that will influence the production and/or dissemination of some scientific output, such as a piece of knowledge or a physical technology; 2) that scientific output could be used in both good and bad ways; 3) there is a trade-off between enabling the good application(s) and preventing the bad one(s), since any alternative that increases the likelihood of the good use, relative to some alternative, also increases the likelihood of the bad use; and 4) given this trade-off and other normative considerations bearing on the choice, it is unclear which alternative(s) it would be morally right for the agent to take.

Introducing an expected-value approach

Having just suggested that the dual-use problem should be understood broadly, I am now, temporarily, going to narrow down on a simplified example of a dual-use problem, so that we have something more tractable to work with. The example is a schematic version of a putative dual-use problem faced recently by scientists working on H1N1 avian influenza.[3]

Suppose that a scientist is deciding whether to embark, with her assistants, on a research project that will investigate whether it is possible to mutate the H1N1 virus so as to make it transmissible by air between humans. (Assume that existing variants of the virus cannot be transmitted in this way.) A likely outcome of the project is that the scientists will indeed discover a method for rendering the virus air-transmissible—that is, that will acquire knowledge that would enable others to create such a virus. This knowledge could, let us suppose, be used to develop a vaccine against the air-transmissible virus that will save 5000 lives, or to facilitate a highly lethal act of biological terrorism in which an air-transmissible variant of H1N1 is released and kills 100 000 people. Suppose (implausibly) that these are the only two possible applications of the knowledge, and that there are no other moral considerations bearing on the scientist's decision whether to embark on the project. To further simplify the case, suppose that the development of the vaccine and the lethal biological attack are events that can occur at most once each. Finally, suppose also that knowledge about how to render H1N1 air-transmissible will certainly not be produced by anyone other than the scientist and her team, so that if the scientist decides not to pursue the project, the lethal H1N1 attack will certainly not occur, but nor will a vaccine against the air-transmissible variant of H1N1 be developed.

The scientist is faced with the choice of taking a gamble. She can gamble on bringing about the good outcome—the development of the vaccine—but it is possible that the bad outcome, the lethal attack, will occur instead or as well. Taking the gamble generates four possible future states of the world

[3] See, for a brief description of the case, Evans, N. G. 2013, 'Great expectations—ethics, avian flu and the value of progress', *Journal of Medical Ethics*, vol. 39, no. 4, pp. 209–13.

- state I: vaccine developed; no bioterrorist attack
- state II: bioterrorist attack; no vaccine developed
- state III: vaccine developed and bioterrorist attack
- state IV: no vaccine developed; no bioterrorist attack.

If the scientist does not take the gamble, state IV is guaranteed.

Assuming (again, implausibly) that the values of these states of the world can be assessed simply by tallying up the number of lives saved by the vaccine and subtracting the number of lives lost in a bioterrorist attack, these states can be assigned the following values

- state I: 5000 lives saved – 0 lives lost = 5000
- state II: 0 lives saved – 100 000 lives lost = –100 000
- state III: 5000 lives saved – 100 000 lives lost = –95 000
- state IV: 0 lives saved – 0 lives lost = 0.

Suppose that the probability of developing a vaccine is rather high if the project goes ahead—say, 0.5 (50 per cent). On the other hand, suppose that the probability of the devastating bioterrorist attack, if the project goes ahead, is low—say, 0.05 (5 per cent). Then the probabilities of each of these states coming about will be

- state I: 0.5 x 0.95 = 0.475 (47.5 per cent)
- state II: 0.5 x 0.05 = 0.025 (2.5 per cent)
- state III: 0.5 x 0.05 = 0.025 (2.5 per cent)
- state IV: 0.5 x 0.05 = 0.475 (47.5 per cent).

States II and III are both very bad. In terms of lives saved and lost, they are more bad than states I and II are good. This counts in favour of abstaining from the scientific project; however, given the probabilities specified above, states I and IV are much more likely to come about than states II and III. This militates in the opposite direction. How should *A* decide whether to take the gamble? One approach would be to calculate the *expected value* of the gamble, and proceed only if it exceeds the expected value of the alternative—namely, obtaining state IV with certainty. (Readers familiar with expected value theory may which to skip to the next section, 'Complicating the Picture'.)

The expected value of an individual state is given by multiplying the value of that state by its probability of occurring. The expected value of a gamble over multiple possible states is given by summing the expected values of the individual states. We can think of the expected value of the gamble as a weighted average of the values of the individual states that may result, with the weights given by the probabilities of those states coming about.[4]

To illustrate, suppose that you are considering whether to engage in a gamble that involves tossing a coin. If the coin turns up heads, you win \$10, if it turns up tails, you lose \$5. The probability of a heads is 0.5, and likewise for tails. If you don't take the gamble, you win or lose nothing. Thus, the expected value of the gamble will be given by

$$\text{probability(heads).value(heads)} + \text{probability(tails).value(tails)}$$

$$= 0.5 \text{ x } \$10 + 0.5 \text{ x } -\$5$$

$$= \$2.50.$$

The expected value of this gamble is \$2.50. On the other hand, the expected value associated with declining the gamble is \$0, as, with certainty, one wins nothing and loses nothing. Since the expected value of the gamble exceeds that of the alternative, the expected-value approach under consideration will advise you to take the gamble.

Now let's return to our schematic dual-use problem. The gamble faced by the scientist has four possible outcomes: I, II, III and IV. So the expected value of the gamble, V_{gamble}, will be given by

$$V_{\text{gamble}} = p(\text{I}).v(\text{I}) + p.(\text{II}).v(\text{II}) + p(\text{III}).v(\text{III}) + p(\text{IV}).v(\text{IV})$$

where p(X) is the probability of state X and v(X) is the value of state X. (A difference with the coin toss example, however, is that in this case, the values of the states are supposed to reflect their overall value, not simply their *value to the agent*, in this case, the scientist. Our question is not whether the scientist ought, from a self-interested point of view, to undertake the research, but whether it would be morally right for her to do so. Insofar as the value of an outcome bears on the moral rightness of the action that brings it about, it is usually thought to be the overall value that matters.)

Plugging in the values and probabilities that we assigned to these gambles, we get the following result:

$$V_{\text{gamble}} = 0.475 \times 5000 + 0.025 \times -100\,000 + 0.025 \times -95\,000 + 0.475 \times 0$$

$$= -2500.$$

[4] Some use the term 'expected value' in a way that presupposes that the values of individual states are measured in monetary terms. Expected value can then be contrasted with 'expected utility', which is instead determined by the amount of utility contained in each state of the world. I use 'expected value' in a way that is neutral between different metrics of value such as monetary value, utility and health. I thus regard 'expected utility' as a species of expected value.

That is, an expected 2500 lives will be lost, on net. On the other hand, the expected value of the alternative—obtaining state IV with certainty—is given by:

$$\begin{aligned} V_{alt} &= v(\mathrm{IV}) \\ &= 0. \end{aligned}$$

The approach under consideration would advise taking the gamble if and only if V_{gamble} exceeds V_{alt}, and against taking the gamble if and only if V_{alt} exceeds the gamble. Given the probabilites and values that we have used, Valt exceeds V_{gamble} so the approach will recommend against taking the gamble—that is, against pursuing the scientific project.

Complicating the picture

The schematic dual-use problem set out above was highly simplified, but the same basic approach could be taken to more complicated, actual dual-use problems. Some revisions and qualifications will, however, be necessary.

For example, in actual cases, there will typically be some chance that the scientific output in question will come about, and be used in good and/or bad ways, even if the agent in question does not produce it: someone else might do so. In producing the output, then, the agent is not replacing a certain outcome, in which there is no chance of the good and bad applications in question, with a gamble in which these outcomes might occur. Instead, she is replacing one gamble with another. To accommodate this, we will need to amend the expected-value approach so that it compares the expected values of two different gambles. In fact, in most cases, since there will be many different actions open to the agent, we will have to compare the expected values of multiple gambles: there will be a different gamble posed by each alternative.

In addition, each gamble will typically involve many more possible outcomes than in the simplified example set out above. In an actual case where a scientist is considering whether to try to develop an air-transmissible variant of the H1N1 virus, there will be a variety of different ways in which the resulting knowledge, if the project is successful, might be used. In addition to being used to develop a vaccine against the new variant or to facilitate a bioterrorist attack, it might also be used, for example, to publicly demonstrate the ease with which terrorists might create a biological weapon, thus encouraging politicians to take further steps to protect against such an attack, or to advance the understanding of what makes viruses air-tranmissible in a way that allows the development of vaccines for existing air-transmissible diseases. Moreover, there will typically be many different forms that each of these kinds of outcome could take. For example, there are many different levels of severity (in terms of both lives lost and other negative consequences) that a bioterrorist attack could have. In addition, though we assumed above that each outcome would occur either once or not at all, in actual cases many of the important good and bad applications of scientific output could occur multiple times. Finally, the development of a technology or piece of scientific knowledge might have good and bad effects that are unrelated to how it will subsequently be used (for example, one good 'effect' might be that there is now more intrinsically valuable knowledge; one bad 'effect' might be that valuable research funding is consumed).[5] These complications all serve to greatly multiply the number of permutations of different states of the world that could result from the scientist's decision, compared with the simplified dual-use problem outlined above.

Moving from schematic examples to the real world would thus greatly complicate any attempt to apply expected-value theory in order to resolve dual-use problems. However, it does not obviously raise any 'in principle' difficulties. One can still envisage an approach to such problems that would: 1) identify the possible outcomes of each alternative course of action; 2) identify the possible 'states of the world', each consisting of combinations of these outcomes;[6] 3) explicitly assess the values and probabilities of these states of the world; 4) use these probabilities and values to calculate the expected value associated with each alternative; and 5) select the alternative associated with the highest expected value. I will call this approach the expected-value procedure or EVP.

Strengths of the expected-value procedure

The EVP has several attractive features.

First, it captures some commonsense intuitions—for example, that decisions should be made by weighing the upsides and downsides of the various choices, that upsides and downsides should be given equal weight, and that, other things being equal, one should prefer an alternative with more valuable upsides, more likely upsides, less disvaluable downsides, or less likely downsides.

Second, the expected-value approach is consistent with a major school of ethical thought: consequentialism. Consequentialism is often taken to hold that an

[5] The outcomes that I describe here would perhaps not normally be described as 'effects' of the action since their relationship to the action is arguably constitutive rather than causal. There is, however, arguably nothing to prevent us from thinking of them as effects in order to include them within the scope of the expected-value approach.

[6] I leave it open whether, in identifying possible outcomes and states of the world, an agent applying this procedure should seek to list *all possible outcomes/states*, *the most likely outcomes/states*, *all reasonably foreseeable outcomes/states* or some other subset of possible outcomes/states.

act is morally right if and only if its consequences will be at least as good as those of any alternative action. One variant of consequentialism holds that an action's *actual* consequences are what determine its rightness. Others hold that the action's *foreseen*, *foreseeable* or *likely* consequences are what are important. Proponents of these latter variants typically hold that the goodness of a set of foreseen/foreseeable/likely consequences is determined by its *expected value*. Their variants of consequentialism thus imply that an action is right if and only if it is associated with at least as much expected value as any alternative. Thus, if one applies the expected-value procedure outlined above when faced with a dual-use problem, one can think of oneself as explicitly trying to identify and adopt the course of action that is morally right according to some variants of consequentialism.[7]

Third, the expected-value approach has a track record. The methodology used in the EVP is a methodology that has been widely employed, often under the label 'risk analysis', 'risk management' or 'cost–benefit analysis', to inform major corporate and government decisions—for example, decisions between alternative power generation projects or mining operations.[8]

Criticisms of the expected-value procedure

Many criticisms of expected-value theory, of its application to major public and private decisions, and of consequentialism might also be adduced against a proposal to use the EVP to confront dual-use problems. In this section I discuss six such criticisms, in each case considering whether and how it might be possible to ammend the EVP to accommodate the criticism.

My aim is not to defend the EVP tooth and nail, by rebutting all objections that might be raised against it. Indeed, I do not comment on whether the criticisms I discuss are persuasive. Rather, the aim is to explore *how far* we must deviate from the procedure in order to evade these criticisms assuming they are persuasive. My conclusion will be that we have to deviate from it to a rather great degree; the approach I end up defending bears little resemblance to the EVP. However, it does, I will suggest, retain at its heart an important role for expected-value theory.

[7] Note that consequentialists differ on what makes a given consequence good or valuable. The most well-known variety of consequentialism—utilitarianism—holds that the goodness of a set of consequences is determined by the total amount of welfare or wellbeing that it contains, but other variants of consequentialism allow that other things besides wellbeing may be of value. For example, a fairer distribution of wellbeing may be better than a less fair one.

[8] For an early, but somewhat controversial, application of expected-value theory to the assessment of risks from nuclear power generation, see Rasmussen, N. C. (chair) 1975, *Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants*, WASH-1400, US Nuclear Regulatory Commission.

This strategy is motivated by the thought that some authors have, in other areas, been too quick to move from the thought that a straightforward expected-value approach faces fatal problems to the claim, often implicit, that expected-value theory is not relevant at all to the problem at hand. Insofar as my argument succeeds in showing that, at least in relation to the dual-use problem, one can accommodate the most important objections to a straightforward expected-value approach without giving up *entirely* on expected-value thinking, I hope I will have established that such a swift rejection of expected-value theory is not justified in relation to the dual-use problem.

1. No adequate metric

An initial criticism of the EVP focuses on the need, in that procedure, to evaluate various possible states of the world. It denies that there is any adequate metric of value for making such evaluations.

In order to apply the expected-value procedure, we will need a cardinal metric of value that can be used to evaluate the various states of the world that might follow from different responses to a dual-use problem.[9] Moreover, we will want this metric to assign a value to every state of the world that is taken as an input into the EVP, and to capture all of the morally significant features of each state. Arguably, there is no such metric. One possibility would be to find some cardinal metric of individual wellbeing. We could then sum the wellbeing of all individuals in each state of the world and use this measure of aggregate wellbeing to compare different states of the world. A problem, however, is that it is controversial whether there is a common cardinal scale on which we can measure and sum the wellbeing of different people. Some deny that there is any cardinal metric that allows the wellbeing of different people to be measured in a reliable way,[10] others question whether interpersonal comparisons of wellbeing have any meaning at all, implying that, as a matter of principle, there could be no such metric.[11] In addition, an aggregate wellbeing metric might fail to capture some morally significant features of each state of the world. Some would argue that, holding aggregate wellbeing constant, differences in the distribution of wellbeing across individuals are morally significant. For example, some argue that it is worse if a fixed amount of wellbeing is distributed less equally, justly

[9] A cardinal metric is one that preserves orderings uniquely up to linear transformations. In a cardinal metric, the interval between two points on the scale has a consistent meaning: it means the same regardless of where those points fall on the scale.

[10] See, for example, Jevons, W. S. 1970 [1871], *The Theory of Political Economy*, Penguin Books, Harmondsworth, UK, 'Introduction'.

[11] See, for example, Robbins, L. 1935, *An Essay on the Nature and Significance of Economic Science*, 2nd edn, Macmillan, London; Arrow, K. J. 1963 [1951], *Social Choice and Individual Values*, 2nd edn, Yale University Press, New Haven, Conn., p. 9. Robbins allows that interpersonal comparisons of wellbeing have meaning as disguised normative judgments, but denies that they have any descriptive meaning.

or fairly.[12] Similarly, others would argue that, in assessing the value of a given state of the world, it matters not just what distribution of wellbeing it contains, but also how that distribution came about.

2. Empirical and evaluative uncertainty

A second criticism targets both the evaluation of different states of the world and the assignment of probabilities to them.

In actual settings where dual-use problems arise, there will often be significant uncertainty about the probabilities and values of possible outcomes, and indeed about what outcomes are even possible. For example, we may simply lack any good evidence of how likely a project investigating whether to render the H1N1 virus air-transmissible between humans is to succeed in realising that goal. Similarly, we may lack any good evidence of how likely it is that knowledge about such a virus would be used to produce biological weapons, of how likely it is that such weapons would be deployed, and of what the likely effects of their deployment would be. Finally, as an example of *evaluative* uncertainty, we may lack evidence on how much disvalue the effects of a terrorist attack, even if they could be fully specified, would have. Since, when faced with such uncertainty, we will simply have to make imperfect estimates of probabilities and values, our expected-value calculations will often give incorrect results.

Accommodating objections 1 and 2

Though both of these problems are serious and may give us some reason to deviate from the EVP, it can be argued that neither of these concerns gives us a clearly decisive reason to reject the use of expected-value theory altogether.

In response to the first of these concerns, it can be argued that, even if there is no *perfect* value metric for employing in the expected-value procedure, there are nevertheless *adequate* ones. Perhaps there is no cardinal metric of value that captures *all morally significant features* of all of the different states of the world that we wish to compare. But there are cardinal metrics that we could use to evaluate *some* of the morally important features.

One approach, then, would be to simply retain the EVP and employ the best cardinal metric that we have. We might thus fail to capture some morally significant considerations, but we would quantitatively balance as many considerations as it is possible to quantitatively balance.

[12] See, for example, Temkin, L. 1993, *Inequality*, Oxford University Press, New York; Persson, I. 2008, 'The badness of unjust inequality', *Theoria*, vol. 69, nos 1–2, pp. 109–24.

It might be objected that this strategy would give disproportionate weight to outcomes that can be rated on a cardinal scale at the expense of morally important considerations that cannot be rated in this way. It would allow cardinally measurable considerations to fully determine what course of action to take, with other considerations given no influence at all; however, this problem could be avoided by modifying the EVP so that the expected values that one calculates on the basis of one's cardinal metric are treated as only an *imperfect indicator* of what course of action to pursue. We could proceed by calculating expected values using the best cardinal metric that we have, and then use the values produced as an imperfect indicator of which course of action to take, though we might sometimes wish to choose a course of action associated with lower expected value than another because we believe it has virtues that are not captured by the metric we have used.

Thus, suppose we wish to compare the value of a world in which a robust vaccine for air-transmissible H1N1 is discovered, but there is also a major bioterrorist attack, with a world in which neither of these things happens. I suggested above that the scientist could compare the value of these with states of the world by tallying up the number of lives lost and saved. In fact, there are likely to be better cardinal measures available. For example, she could instead look at the total number of quality-adjusted life-years (QALYs) that would be enjoyed by the whole population in each possible state of the world resulting from each of the available courses of action. This would allow her to capture at least some of the possible effects of her actions on morbidity, as well as mortality. Clearly, the QALYs measure would still be imperfect and would not capture many evaluatively important effects of the two alternatives. For example, it would not capture non-health-related outcomes such as the inconvenience and loss of privacy caused by security measures that would be introduced as a result of a major terrorist attack. But a determination of the expected level of QALYs associated with each course of action nevertheless plausibly provides *some indication* of which of these courses of action the scientist should pursue. The scientist could use expected QALYs as an indicator of whether to pursue the H1N1 research, and then consider whether factors not captured by this measure would justify deviating from the course of action that it suggests. Exactly how these other factors would need to be taken into account would of course need to be determined. It is not clear how we should weigh considerations that cannot be quantified using a shared cardinal measure. But this problem is not specific to attempts to use expected-value theory to respond to dual-use problems. If there are factors that cannot be rated on a shared cardinal measure, *any* proposal for how we ought to respond to dual-use problems will need to determine how these should be taken into account.

The concern about uncertainty can, I think, be accommodated in a similar way. As with the concern about metrics, this worry does not take issue with the basic idea underpinning EVP—the idea that our choices when confronted with dual-use problems should reflect the expected value of the available alternatives. Rather, it points out a practical difficulty with trying to implement this idea. But it remains plausible that *trying* to calculate the relevant expected values would be desirable. We could calculate expected values for the available courses of action using our best estimates of the values and probabilities of the relevant outcomes, insofar as it is possible to assign values or probabilities at all. We could then decide what course of action to pursue on the basis of these estimates of expected value.

It might be argued that this approach would yield implausible results in some cases. It is possible that two courses of action could have the same expected value, though the probabilities and values of the outcomes associated with one course of action are highly uncertain, and the probabilities and values associated with the other course of action are certain. Arguably, in this case one should prefer the latter course of action.

But again, this problem could be accommodated by treating the expected values as only an imperfect indicator of what to do. This would retain some role for expected values, but would allow us to favour a course of action with a lower expected value than another because the other course of action could cause outcomes whose probability or value is less certain.

The concerns about the lack of a cardinal metric and uncertainty suggest that the EVP may not be able to capture all of the considerations that are relevant to how one should act when faced with a dual-use problem: it does not capture outcomes that cannot be evaluated on a common cardinal scale, nor does it capture differences in the certainty of the values and probabilities of the possible outcomes. It is, however, still plausible that we should adopt a procedure that incorporates the most important elements of the EVP. More specifically, it remains plausible that we should conscientiously calculate expected values using the best metric available and take this as at least an imperfect indicator of what to do. Call this the *restricted* expected value procedure or rEVP.

There are, however, two further criticisms that suggest it might be unwise even to adopt the rEVP.

3. Bias

In attempting to identify and assign probabilities and values to different outcomes, agents might have systematic tendencies to neglect or exaggerate some considerations. For example, people might overstate the disvalue of negative outcomes relative to the value of positive outcomes, or they might exaggerate the probability of the most extremely valuable or disvaluable events occurring. In some cases, the direction of a bias may depend on the particular circumstances and psychology of the person(s) applying the rEVP. For example, people may be inclined to relatively exaggerate the (dis)value or probability of outcomes that they have experienced in the recent past, and so are associated with readily available mental images.[13] Or they might be prone to exaggerate the (dis)value or probability of outcomes that will primarily affect them rather than others (self-serving bias).[14] They may also be prone to underestimate the disvalue of harms that affect large numbers of people (scope insensitivity)[15] and susceptible to framing effects in which the way an outcome is presented affects the probability or value assigned to it.[16] Given these biases, the application of the rEVP could be expected to give misleading results in many cases—that is, the expected values that it takes as indicators may often be incorrect.

4. Demandingness

Applying the rEVP comprehensively and conscientiously might require substantial time, effort, expertise and financial resources—all resources that could otherwise be spent on other worthwhile activities. It might also be psychologically burdensome in the sense of requiring those who apply it to override their natural inclinations (suppose that a journal editor strongly committed to academic freedom decides, on the basis of the rEVP, that she must heavily censor a submitted article). Applying the rEVP might, due to the presence of such costs, have negative overall consequences even if it leads decision-makers to make the best choices. The benefits of making the best decisions might be outweighed by the costs associated with the means via which those decisions were reached.

[13] Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M. and Combs, B. 1978, 'Judged frequency of lethal events', *Journal of Experimental Psychology: Human Learning and Memory*, vol. 4, no. 6, pp. 551–78; Sunstein, C. 2007, *Worst-Case Scenarios*, Harvard University Press, Cambridge, Mass., p. 6.

[14] Babcock, L., Loewenstein, G., Issacharoff, S. and Camerer, C. 1995, 'Biased judgments of fairness in bargaining', *American Economic Review*, vol. 85, no. 5, pp. 1337–43.

[15] Desvousges, W. H., Johnson, F. R., Dunford, R. W., Boyle, K. J., Hudson, S. P. and Wilson, N. 1993, 'Measuring natural resource damages with contingent valuation: tests of validity and reliability', in J. A. Hausman (ed.), *Contingent Valuation: A Critical Assessment*, North-Holland, Amsterdam, pp. 91–159; Fetherstonhaugh, D., Slovic, P., Johnson, S. and Friedrich, J. 1997, 'Insensitivity to the value of human life: a study of psychophysical numbing', *Journal of Risk and Uncertainty*, vol. 14, pp. 238–300.

[16] Tversky, A. and Kahneman, D. 1981, 'The framing of decisions and the psychology of choice', *Science*, vol. 211, no. 4481, pp. 453–8.

Accommodating objections 3 and 4

The problem of bias is likely to be a problem for many approaches besides the rEVP. Any plausible approach to dual-use problems will require us to identify, evaluate and assess the probability of at least some possible outcomes. Some may permit us to do so in an implicit, subconscious way rather than an explicit, conscious way—for example, rather than adopting the rEVP we could simply have experts make decisions based on their intuitions about the case at hand. This does not, however, necessarily render such approaches less susceptible to bias. Indeed, it may render them more susceptible. The process of assigning values and probabilities may make our biases more obvious and allow us to correct for them. Nevertheless, it may be possible to reduce the risk of bias by deviating from the rEVP. Suppose it were known that people quite generally tend to understate the probability of extremely bad states of the world, even when they are aware of this and attempt to correct for it. We could to some extent compensate for this bias by making some changes to the rEVP. We could require, for example, that the most disvaluable states be given greater weight than other states even if we judge their probability of occurring to be the same.[17]

Similar thoughts apply to the problem of demandingness. Again, this problem is likely to apply to many approaches besides the rEVP; however, it may be possible to reduce the burdens of deciding between alternative strategies by deviating from the rEVP. One approach would be to apply simple heuristics (for example, 'unless there is a clear and present risk of misuse, proceed as if there were no risk'), and resort to the rEVP only in hard cases—for example, cases that are not covered by the simple heuristics, or cases in which heuristics conflict. Alternatively, the problem of demandingness could be mitigated by ensuring that the rEVP is applied by those able to apply it at least cost. For example, it may be less burdensome for a committee of scientific, health and security experts to assess the values and outcomes of synthetic biology research than it is for individual scientists to do the same. Regulatory bodies could be charged with applying rEVP, whilst individual scientists follow a simple heuristic such as 'aggressively pursue your scientific goals unless this violates a regulation'.

Concerns about bias and demandingness may justify deviation from the rEVP. But note that neither justifies a wholesale rejection of the basic ideas underpinning that procedure. In particular, both are consistent with retaining the idea that *the right course of action to take, when faced with a dual-use problem, is (the) one associated with at least as much expected value as any available alternative* (call this the expected-value *criterion* or the EVC). The concerns simply show that the *procedure* by which we make decisions when faced with dual-use problems should not necessarily be to explicitly calculate expected values. Explicitly

[17] This might, for example, involve applying a multiplier (> 1) to the most negatively valued outcomes.

attempting to satisfy the expected-value criterion might not be the best way of in fact satisfying that criterion, or might involve costs that outweigh the benefits of satisfying it.

This distinction between *criteria of rightness* and *decision procedures* has long been emphasised by consequentialists when presented with concerns about demandingness and bias and other related objections. Consequentialists have argued that their theory provides an abstract description of which acts are right and which are not. It does not provide a procedure via which to decide how to act.[18] In fact, consequentialism is consistent with adopting some other ethical theory as a procedure for guiding decisions. In an ideal world where we responded to our evidence in a perfectly rational and costless way, we *could* simply apply consequentialism as a decision procedure. But the actual world is not like that. In the actual world it may, according to consequentialists, be best *not* to make decisions by explicitly calculating which act will have the best consequences.

Drawing a distinction between criteria of rightness and decision procedures allows us to retain the basic idea underpinning EVP (and thus rEVP) while deviating from these procedures. More importantly, it provides us with a higher standard—the EVC—against which we can measure alternative decision procedures. Other things being equal, we should prefer a decision procedure that comes closer than some alternative procedure to recommending the courses of action that are right according to the EVC—the courses of action associated with the highest expected value. In an ideal world, the best decision procedure would be the EVP or rEVP. In the actual world, it may be some amendment of these procedures, or even a wholly different approach. Either way, it is plausible that we should keep the EVC in mind as a standard by which to judge competing decision procedures.

I now turn to consider two criticisms that suggest that even the EVC will need to be rejected.

5. Rational risk aversion

The expected-value procedure, and the criterion that underpins it, is blind to the distribution of value across possible states of the world. Suppose that a

[18] See, for example, Austin, J. 1954 [1832], *The Province of Jurisprudence Determined*, H. L. A. Hart (ed.), Weidenfeld, London, p. 108; Mill, J. S. 1985 [1861], 'Utilitarianism', in *Collected Works. Volume X*, John M. Robson (ed.), University of Toronto Press, Toronto, ch. 2, pp. 203–60; Sidgwick, H. 1907, *The Methods of Ethics*, 7th edn, Macmillan, London, p. 413; Bales, R. E. 1971, 'Act-utilitarianism: account of right-making characteristics or decision making procedure?' *American Philosophical Quarterly*, vol. 8, pp. 257–65; Parfit, D. 1984, *Reasons and Persons*, Clarendon Press, Oxford, pp. 24–9, 31–43; Railton, P. 1984, 'Alienation, consequentialism, and the demands of morality', *Philosophy and Public Affairs*, vol. 13, pp. 134–71, at pp. 165–8. The terminology 'decision procedure' and 'criterion of rightness' is due to Bales, op. cit.

scientist is considering whether to publish a scientific finding that has a 50 per cent chance of being used in a good way and a 50 per cent chance of being used in a bad way. Suppose further that the value of the good outcome is 100, and of the bad outcome is −100. Then there will a 25 per cent chance of realising a state of the world with a value of 100 (good outcome, no bad outcome), a 25 per cent chance of realising a state with a value of −100 (bad outcome, no good outcome), and a 50 per cent chance of realising a world with a value of zero (because both the good outcome and the bad outcome occur or neither occurs). Suppose that the scientist does not publish the result: there is a 100 per cent probability that neither the good nor the bad outcomes will occur. In this case, the expected value associated with publishing the result is zero, as is the expected value associated with not publishing it. Thus, the EVC will hold that both courses of action are right. But, arguably, the right thing to do in this case is to abstain from pursuing the project since this alternative is associated with lower (indeed zero) risk: if the scientist abstains from the project, a world with a value of 0 will obtain with 100 per cent probability, whereas if she pursues the project there is an expected value of zero, but a significant risk of winding up in a world whose value is −100. If we ought to be risk averse, as some have argued,[19] we should reject EVC in favour of an alternative criterion that penalises riskier alternatives.

6. Agent-relativity

The EVC is also blind to the way in which an agent posed with a dual-use problem contributes to a good or a bad outcome; it focuses only on the probability and value of the outcome. But, according to many nonconsequentialist ethical theories, an agent's relation to an outcome is also important. Consider a case in which a scientist performs and publishes some piece of research in synthetic biology that is intended to develop a new cure for cancer but could also be misused in unjustified biowarfare. It might be argued that this possible negative outcome should be discounted relative to at least some of the other outcomes because

1. It is not caused by the scientist, but is merely allowed to occur. In contrast, at least some positive outcomes of the research—for example, the production of intrinsically valuable knowledge—might be said to be *caused* by the scientist.[20]
2. It is not intended, but is merely foreseen, by the scientist. By contrast, the possible positive outcome of developing a cure for cancer is intended.[21]

[19] See, for example, Hansson, S. O. 2003, 'Ethical criteria of risk acceptance', *Erkenntnis*, vol. 59, no. 3, pp. 291–309.

[20] This claim could be grounded on the doctrine of doing and allowing.

[21] This claim could be grounded on the doctrine of double effect. See Suzanne Uniacke's Chapter 10, in this volume.

3. The occurrence of the bad outcome depends on the subsequent immoral actions of another agent: the person who actually engages in unjustified biowarfare. Thus, even if the scientist's actions resulted in unjustified biowarfare, the scientist would not be the primary wrongdoer, but would at most be an *accomplice* to the wrongdoing. By contrast, some other outcomes of the work, such as the production of intrinsically valuable knowledge, may occur without requiring the intervention of other moral agents. The scientist might be said to be the principal agent implicated in bringing about these outcomes.

Accommodating objections 5 and 6

If these criticisms are well founded, the EVC should be rejected since the right strategy to adopt when faced with a dual-use problem will *not* necessarily be the one that offers the greatest expected value. It may, for example, be the one that minimises *risk*, or that minimises the likelihood of *intentionally causing* harm.

I am not in a position to assess the arguments for and against agent-relativity or the rationality of risk aversion here. But I will briefly consider how a proponent of the EVC might seek to accommodate these criticisms. She might argue that, even if we accept agent-relativity and that it is rational to be risk averse, it may be that the EVC still serves as a useful starting point—as a default position from which alternative criteria and procedures for assessing responses to dual-use problems might be derived. Thus, we would assume that all equally likely and equally valuable outcomes should be given the same weight *unless* some sound nonconsequentialist argument could be supplied for prioritising or discounting outcomes to which we bear a certain relationship. Similarly, we would take a risk-neutral approach unless a sound argument for risk aversion could be found.

One reason to take an expected-value-based approach as a starting point is that it captures certain core considerations while taking all other possible factors to be irrelevant until proven otherwise: everyone should agree that in confronting dual-use problems, it matters what good and bad outcomes might result, how likely they are, and how good or bad they are. An expected-value approach captures the relevance of these considerations. Some might argue that other factors, such as risk and the agent's relationship to the good and bad outcomes, are also important; but this would be controversial, so it might seem appropriate to adopt, as a default position, a decision procedure that does not take them into account.

A second reason to adopt an expected-value approach as a point of departure is that the formal framework of expected-value theory may provide a helpful tool for thinking about other approaches as well. As we saw above, the expected value of a gamble over a number of different states of the world can be thought of as a weighted average of the values of those states, with the weights given by their probabilities. But we could also weight the values using further factors to accommodate agent-relativity or rational risk aversion. Consider a nonconsequentialist approach that takes into account the *intentions* of an agent posed with a dual-use problem as well as the consequences of her decision. It may be helpful to think of such an approach as a variant of the expected-value approach that weights the values of different possible states of the world according to both their probability *and* whether they were intended.[22] Adopting such a formalised approach may help to ensure that we are clear and explicit about precisely *how much* the intentions of the agent matter.

Conclusions

I have outlined an approach to dual-use problems that involves: 1) identifying possible outcomes of each alternative course of action; 2) identifying possible states of the world that might result, each consisting of a combination of these outcomes; 3) explicitly assessing the values and probabilities of these states of the world; 4) using these probabilities and values to calculate the expected value associated with each alternative course of action; and 5) selecting the alternative with the highest expected value.

This expected-value procedure is attractive because it captures some commonsense intuitions about how to respond to risks and benefits, is consistent with a major school of ethical thought (consequentialism), and has a track record in government and corporate decision-making. It is, however, also susceptible to several criticisms, which may justify deviation from it. Criticisms adverting to the lack of an adequate metric for ranking states of the world and to empirical and moral uncertainties suggest that we may need to regard expected values as providing only an imperfect indication of which course of action to pursue. Criticisms appealing to bias and demandingness suggest that, even if the right course of action to take when faced with dual-use problems is always the one that maximises expected value, it may be best to adopt a decision procedure that does not involve calculating expected values at all. Finally, criticisms based on agent-relativity and the rationality of risk aversion go even further: they suggest that in some cases the right course of action will *not* be the one that maximises expected value.

I have not assessed whether these criticisms are persuasive. But I have tried to show that, even if they are, there may be some value in retaining the expected-

[22] The expected value assigned to an alternative would then be relative to the agent making the decision between alternatives; it would no longer be a value that could be ascribed from any standpoint.

value criterion as a default position from which deviations must be justified. Thus, there may remain an important role for expected values in thinking about dual-use problems.

Further questions

Suppose that we do keep the EVC in mind, at least as a starting point for further discussion. Two important further questions arise.

First, what outcomes should be included among the positive and negative outcomes of a course of action? For example, should the production of knowledge be included as a positive outcome independent of any positive applications of that knowledge? This will depend on the controversial question of whether knowledge has any intrinsic value.

A second question is, to the extent that *anyone* should ever explicitly assess the expected value associated with a particular strategy for preventing misuse, *who* should do it? For example, should the decision be made by some form of expert committee, or should it be made by a political or representative body?

I leave these questions as potential topics for future discussion.