Unknowns in Dual Use Dilemmas

Michael Smithson, The Australian National University

1. Introduction

Dual use dilemmas are defined as a consequence of the potential for the same piece of research to be used for harm and good. Miller and Selgelid[1] advise that ``Fine-grained ethical analyses of dual-use research in the biological sciences would seek to *quantify* actual and potential benefits and burdens, and actual and potential recipients/bearers of these benefits and burdens. These analyses would also identify a range of salient policy options.'' Desirable as such quantification may be, the path to it is obstructed by several yawning abysses in the form of unknowns. If unresolved or ignored, these unknowns can render fine-grained analysis and quantification impossible or arbitrary.

This paper investigates these unknowns and presents some approaches for dealing with them or, at least, taking them into account. These approaches are grounded in subjective expected utility (SEU) theory, whose primary tenet is that "rational" agents weigh up the potential consequences of acts by summing the products of the probability of every possible outcome and its utility. At least some of the probabilities and utilities might be based on subjective assessments, whence the "S" in SEU. SEU is employed here as a prescriptive or benchmark framework. My primary intent is to ask what an SEU-rational agent would conclude or choose, so that human decision makers can knowledgeably decide whether to take the agent's advice on board or reject it.

Our survey of unknowns comprises three sections:

1. Dilemma structures
2. State space indeterminacy
3. Imprecision and biases in judgements

The first section examines dual use dilemmas from the viewpoint of the standard social dilemmas framework. The primary purpose is to ascertain when a dual use dilemma is a mixed-motive game and therefore a genuine dilemma, and when it is a tradeoff. Dilemmas pose difficulties for rational self-interest that tradeoffs do not. The second section begins with the observation that dual-use dilemmas often are not limited to considering just two possible uses and may instead involve an indeterminate number of uses. Likewise, the number of response options also may be a matter of choice. In other words, the use and response state spaces are indeterminate. Both state spaces have consequences for decision making and these are elaborated in this section. The third section begins by pointing out the dangers in restricting judged probabilities and utilities to falsely precise representations and describing the decisional consequences of imprecision. It then brings in psychological considerations such as tendencies towards overconfidence in predictions and confirmation bias.

[1] Miller & Selgelid (2007: 42), emphasis in the original

2. Dilemmas or Tradeoffs?

When are the dilemmas actually social dilemmas, as opposed to tradeoffs? Genuine social dilemmas are harder to resolve than tradeoffs. They also present a fundamental difficulty for rational self-interested agents because the pursuit of self-interest in a social dilemma leads to the destruction of the common good. Moreover, the structure of a social dilemma partly determines the approaches needed to resolve it.

First of all, social dilemmas are social. They involve a game structure comprising at least two decision makers. Some ``dual-use dilemmas'' do not readily yield such a game structure because they are cast as single-agent decisions. An example is the concern that research conducted for beneficial purposes might be used by secondary researchers or other users to construct bioweapons. If these users would not be able to exploit the research if it were not conducted, then the situation reduces to a single-agent decision:

Research → potential benefits and risk of exploitation

versus

No research →potential costs and no risk of exploitation

While this decision may be difficult, it is not a social dilemma or even a dilemma in the sense of “damned if you do and damned if you don't”. Instead, this is arguably a tradeoff wherein each option combines potentially strong positive and strong negative consequences.

Dual-use dilemmas can become social dilemmas involving multiple agents if the decisions made by each agent alter the consequences for all of them. Biological research as an arms-race is perhaps the most obvious example. For instance, if researchers in country A revive an extinct pathogen and researchers in country B do not, country A temporarily enjoys a tactical advantage over country B while also risking theft or accidental release of the pathogen. If country B responds by duplicating this feat then B regains equal footing with A but has increased the overall risk of accidental release or theft. For country A the situation has worsened not just because it has lost its tactical advantage but also because the risk of release has increased. Conversely, if A restrains from reviving the pathogen then B may play A for a sucker by reviving it. It is in each country's self-interest to revive the pathogen in order to avoid being trumped, but the collective interest resides in minimizing the risk of accidental or malign release.

A similar example of a social dilemma is where countries A and B are considering whether to eliminate their respective stockpiles of smallpox. The payoff matrix is shown in Table 1. The entries are:
R = reward
T = temptation
S = sucker
P = punishment.

This matrix enables a definition of a social dilemma. A social dilemma exists iff these four conditions hold:

$R > P$

$R > S$

$2R > T + S$

$T > R \quad \text{or} \quad P > S$

Table 1: Payoff Matrix for a Two-Agent Game

		B: Eliminate	B: Retain
A	Eliminate	R_a, R_b	S_a, T_b
	Retain	T_a, S_b	P_a, P_b

There are 3 well-known dilemma structures, depending on how each country's decision maker rank-orders the consequences of the countries' joint decisions:

Chicken: $T > R > S > P$

Prisoner's: $T > R > P > S$

Trust: $R > T > P > S$

We do not require quantification of the matrix entries; they only need have a complete ordering for each player. We will denote the best outcome by 4 and the worst by 1. Of course, it is possible for the structure to differ between the two countries. In Table 2, the structure is Chicken for country A and Prisoner for country B.

Table 2: Chicken and Prisoner's Dilemma Combination

		B: Eliminate	B: Retain
A	Eliminate	3,3	2,4
	Retain	4,1	1,2

Table 2 makes it easy to see the roles played by greed and fear in a social dilemma. Each country can obtain its best outcome (rated 4) by retaining their supply if the other country eliminates theirs. Country B's worst outcome (rated 1) and country A's second-worst result if each eliminates supply while the other retains theirs. If both act on fear and/or greed and retain their supplies then the joint outcome is the worst of all four (rated 1 for A and 2 for B).

Different structures yield distinct pressures for and against eliminating smallpox stockpiles. A "cooperation index" is

$$K = \frac{R - P}{T - S}$$

which provides an overall indication of motivation for elimination. All else being equal, Prisoner will have a smaller value for K than Chicken or Trust. The cooperation index, in turn, may be decomposed into a Fear and Greed component:

$K = 1 - \left(K_f + K_g \right)$, where

$K_f = \frac{P-S}{T-S}$ and

$K_g = \frac{T-R}{T-S}$.

Thus, in Trust and Prisoner $K_f > 0$ whereas in Chicken $K_f < 0$, while in Chicken and Prisoner $K_g > 0$ whereas in Trust $K_g < 0$. In Chicken Greed is the component detracting from motivation to eliminate stockpiles and in Trust Fear is the detractor. Prisoner is the only dilemma in which both the Fear and Greed components exceed 0, so that both detract from motivation to eliminate. This is why K generally is lowest for Prisoner.

Another crucial characteristic of a dilemma is the "public" versus "private" nature of the consequences. This strongly influences whether institutional solutions such as privatization are potential solutions for social dilemmas. A good is subtractable if its use by one agent decreases the potential for its use by another. Attention is subtractable (devoting attention to one thing decreases the attention that can be given to others) whereas information is non-subtractable (simply acquiring information does not decrease its availability to others). A good is excludable if access to it can be restricted. Secrets and legally proprietary information are fairly excludable, whereas unsecured information on the internet is not. Goods are privatizable insofar as they are excludable and subtractable.[2]

Public goods (and bads) are strongly non-subtractable and non-excludable. The open-access and communalistic norms of scientists render research outputs a public good. A virulent, easily transmissible pathogen quickly can become a public bad. Common-pool resources, on the other hand, are goods that are subtractable but non-excludable. Air or water quality is an example of a common-pool resource. Toll goods are those, like proprietary information, that are excludable but non-subtractable. And finally, truly private goods are those, like well-guarded smallpox supplies, that are both excludable and subtractable.

The temporal dimension also can play an important role in dilemmas.[3] A large literature indicates that repeated dilemmas are more easily solved than one-shot dilemmas.[4] Repeated dilemmas permit agents to learn, build trust, or negotiate and verify compacts, whereas these are considerably more difficult in one-shot dilemmas. Consider ``Cat Out of the Bag'' (COB) consequences: It takes only one instance of the research to yield the potential for misuse or accident; subsequent research replication usually does not increase those risks. The COB risk associated with a particular research project can be the basis of a one-shot dilemma. However, if we consider the potential for multiple research efforts to throw up COBs then we have the makings of a repeated dilemma. Packaging one-shot dilemmas into a common category reframes them as repeated dilemmas, enhancing the chances of solving them.

Finally, it should be noted in passing that we have implicitly assumed that both agents know not only their own outcome preferences but each other's as well. Of course, it is also crucial to take into account each agent's perception of the other's payoffs, because those

[2] Ostrom, Gardner, and Walker (1994)
[3] e.g., Smithson (1999a)
[4] Danielson (1992)

determine what each agent believes the other's (rational) motives and best moves will be. Referring back to Table 2, if country A's intelligence is that country B's payoff matrix is identical to A's (i.e., Chicken instead of Prisoner) then A will underestimate B's motivation for retaining smallpox supplies. This is because B's actual outcome ranks for retention are {4,2} and for elimination they are {3,1} whereas A will believe they are {4,1} and {3,2} respectively.

Obviously there is much more to determining the nature of a dual use dilemma's structure than has been dealt with in this section. The intention here is merely to provide a starting-point by posing the question of whether a structure constitutes a social dilemma and, if so, what kind of social dilemma the structure corresponds to. The crucial difference between a social dilemma and a tradeoff is that a social dilemma entails a conflict between individual and collective interests that does not appear in tradeoffs. It is plausible, therefore, that the policies and procedures for dealing with dual use dilemmas also will need to distinguish between the two.

3. Partition Indeterminacy

Nearly all formal decision-making frameworks, including SEU, assume that all possible options and outcomes are known. In other words, the state space is predetermined. The nature of innovative research implies that in at least some dual use dilemmas that assumption is untenable on three counts. First, the potential outcomes of research often are not completely known. The accidental creation of a mousepox "superstrain"[5] is a case in point. Second, the uses of research outputs also sometimes are unanticipated. Witness the applications in cryptography of number theory, a sub-discipline that once was held up as the epitome of pure mathematics beyond reach of any applicability. Third, the variety of responses to the threat of research misuse is not predetermined. The first two sources of state space indeterminacy are matters to be taken into account by those who make judgements and decisions, and this is the topic of the next subsection. The third, however, can be a matter of choice, and this is discussed in the subsection thereafter.

3.1. Unknown Outcomes and Consequences

In most standard probability theories, on grounds of insufficient reason, a probability of $1/J$ is assigned to J mutually exclusive possible events when nothing is known about the likelihood of those events. For example, in a race involving three greyhounds, an agent who knows nothing about any of the dogs would assign a value of 1/3 to the probability of each greyhound winning. Moreover, even under alternative probability assignments the probabilities of the J events must sum to 1, meaning that the entire probability mass is concentrated on that set of events. Thus, a more knowledgeable rational agent who has assigned a probability of 1/2 to the first dog winning and 1/4 to the second dog is compelled to assign the remaining 1/4 to the third.

The number of possible elementary events or states in a space, is determined by the *partition* of that space. The greyhound race has been partitioned into 3 outcomes: Dog 1 wins, dog 2 wins, or dog 3 wins. Were we to allow ties, the partition would expand to $J = 7$. The ignorant agent now would assign a probability of 1/7 to each dog winning, and the

[5] Selgelid (2007)

more knowledgeable agent could distribute the remaining 1/4 probability across the remaining 5 events instead of having to allocate it all to the third dog winning. Thus, probability assignments are *partition-dependent*.

When partitions are indeterminate, partition dependence poses a problem for subjective probability assignments. This is not the same problem as unknown probabilities over a unique and complete partition (e.g., where we know that there are only red and black marbles in a bag but do not know how many of each).[6] It is more profound. In the absence of a uniquely privileged partition, there is no defensible prior probability distribution to be constructed.

Two separable problems for partitions may arise. One is an incomplete account of possible events. A unique and complete partition might be attainable in principle, but we lack the necessary information. The other problem is the absence of a privileged partition even when one has a complete account of those possibilities. Shafer[7] presented an example of this problem as a motivation for the belief functions framework. He asked whether the probability of life existing in a newly discovered solar system should be partitioned into {life, no life} or {life, planets without life, no planets}. This issue arises naturally when a decision must be made that involves a threshold or interval on a continuum. We shall revisit this particular problem in the next subsection.

Returning to the first problem, the most common situation confronting judges or decision makers is partial knowledge of the possible outcomes. We may know some of the potential uses and misuses of a new biotechnology but not all of them. We might even be willing to assign a subjective probability that party X will misuse this technology in ways we can anticipate. But what probability should we assign to X misusing the technology in ways we haven't anticipated? Likewise, what probability should we assign to party Y finding a new way to use the technology for good?

Smithson[8] presents strategies for dealing with partition dependence, distinguishing those that apply when a privileged or at least agreed-upon partition is attainable from those that apply when it is not.

1. Where a privileged or agreed-upon partition is attainable:
    a. Debiasing strategies
    b. Establishing criteria for choosing partitions
2. Where there is no privileged or agreed-upon partition:
    a. Using diverse partitions
    b. Modeling partition-dependence effects
    c. Using (nonstandard) probabilistic frameworks that avoid partition-dependence

Debiasing strategies are needed because human judges are strongly influenced by partitions in their subjective probability assignments. Two important manifestations of partition-dependence are distorted judgments of likelihoods of compound events and anchoring on an ignorance prior. A classic study[9] concerning people's assignments of

[6] Smithson (2009)
[7] Shafer (1976)
[8] Smithson (2009)
[9] Fischhoff, Slovic, & Lichtenstein (1978)

probabilities to possible causes of a given outcome (e.g., an automobile that will not start) revealed that possible causes that were explicitly listed received higher probabilities than when the same causes were implicitly incorporated into a "Catch-All" category of additional causes. The effect has since been referred to as the "Catch-All underestimation bias" and also sometimes the "pruning bias".[10]

Likewise, it has been empirically demonstrated[11] that subjective probability judgments are typically biased towards the ignorance prior determined by the partition salient to the judge. That is, people anchor on a uniform distribution of $1/J$ across all J possible events, even when taking into account prior evidence of how likely each event is. Because those adjustments typically are insufficient[12], judges' intuitive probability assignments are biased toward probabilities of $1/J$.

Criteria for choosing partitions and methods for exploring diverse partitions are not well established. One recently proposed set of criteria will be elaborated in the next subsection[13], but these have limited scope. Other criteria could be linked with strategies for manipulating and exploring judgement biases in informative ways. As a simple example, expert judges estimating probabilities of adverse consequences arising from the revival of an extinct pathogen could be randomly assigned to one of two conditions: A 2-fold partition (consequence vs no consequence) or a J-fold partition (a list of anticipated consequences plus a catch-all category for unanticipated ones). Partition dependence would predict that the average probability of an adverse consequence in the first condition should be less than the average sum of the probabilities across the J consequence categories in the second condition. The results would yield fairly defensible lower and upper expert estimates of the probability of adverse consequences. More sophisticated experimental designs would enable the construction and estimation of relevant partition dependence effects.

Finally, let us briefly consider non-standard probability frameworks that are not partition-dependent. These have appeared in the growing literature on generalized probability theories, and also in behavioral economics.[14] Walley[15] argues on normative grounds that imprecise probability frameworks can avoid partition dependence entirely. He proposes that when judges are permitted to provide a lower and upper probability judgment (i.e., imprecise probabilities) every ignorance prior should consist of vacuous probabilities {0,1}. In the greyhound race example, the ignorant agent could assign a lower probability of 0 and an upper probability of 1 to every event regardless of whether the partition is 3-fold or 7-fold. The lower and upper probabilities of the first dog winning would be 0 and 1 regardless of the partition, thereby avoiding partition dependence. Walley developed an updating method (the Imprecise Dirichlet Model) that is partition-independent and has generated interest within the community of imprecise probability theorists.

That said, recent studies[16] experimentally demonstrated that naïve judges are just as strongly influenced by partitions when making imprecise probability judgments as they are

[10] Russo & Kolzow (1994)
[11] Fox & Rottenstreich (2003)
[12] Tversky & Kahneman (1974)
[13] Smithson (2006, 2009)
[14] e.g., Grant & Quiggin (2004)
[15] Walley (1991, 1996)
[16] Smithson and Segale (2009)

when making precise probability judgments. Moreover, they demonstrated that many judges anchor on $1/J$ as the midpoint of their lower and upper probability judgments. No applicable debiasing strategies have yet been reported. Nevertheless, the possibility remains that allowing judges to express one kind of uncertainty (imprecision in their probability assignments) may militate against the impact of another kind (partition indeterminacy).

3.2. How Many Options?

Policies regulating responses to dual use dilemmas could be limited to two options, e.g., laissez-faire and bans. But what about a third option, such as oversight by a regulatory body? Or more than two additional options? Are there criteria that could indicate how many options a rational agent should prefer? How would we know whether each option was worth retaining? This appears to be a relatively unexplored topic, but reasonably important given that this is one aspect of dual use dilemmas where policy and decision makers actually have choices. It is directly related to partition indeterminacy because we are constructing a partition of a space of possible acts.

In the context of legal standards of proof, a typical threshold probability of guilt associated with the phrase "beyond reasonable doubt" is in the [.9, 1] range.[17] For a logically consistent juror a threshold probability of .9 implies the difference between the utility of acquitting vs convicting the innocent, is 9 times the difference in the utility of convicting vs acquitting the guilty.

Connolly demonstrated that the utility assignments to the four possible outcomes (convicting the guilty, acquitting the innocent, convicting the innocent, and acquitting the guilty) that are compatible with such a high threshold probability are counterintuitive. Specifically, "… if one does [want to have a threshold of .9], one must be prepared to hold the acquittal of the guilty as highly desirable, at least in comparison to the other available outcomes".[18] He also showed that more intuitively reasonable utilities lead to unacceptably low threshold probability values.

Smithson[19] showed that the incorporation of a third middle option (such as the Scottish Not Proven verdict) with a suitable threshold can resolve this quandary, permitting a rational (subjective expected utility) agent to retain a high conviction threshold and still regard false acquittals as negatively as false convictions. The price paid for this solution is a more stringent standard of proof for outright acquittal.[20] The main point here is that a consideration of preferences as expressed by the relative positions of utilities can aid in the choice of a partition of acts, due to the connection between these utilities and the threshold probabilities that determine when one act is chosen over another.

Applying Smithson's framework to dual use dilemmas, consider the simplest setup in which either some kind of misuse of a research output occurs or no misuse occurs. Suppose we must make a decision regarding the fate of a potential research project (e.g., whether to prohibit it or allow it to proceed), and we wish to do so on the basis of an estimated probability that the research output could be misused. Let us assume that choices will affect

[17] Connolly (1987)
[18] Connolly (1987: 111)
[19] Smithson (2006)
[20] for evidence that this also is what humans do, see Smithson, Deady, & Gracik (2007)

the utility of the no-misuse outcome because of inhibited scientific progress and/or resource expenditure in security arrangements. Let us also assume that the utility of the misuse outcome also will be affected by choice because the same considerations will be combined with the consequences of misuse, even if they are dwarfed by the latter.

Suppose we have a J-fold partition of acts R_j, for j = 0, 1, 2, …, J-1. There are two possible outcomes: No misuse and misuse. The act R_j has a utility H_j if there is no misuse and a utility G_j if there is misuse. We assume that the acts R_j are ordered so that $H_j > H_{j-1}$ and $G_{j-1} > G_j$ for any j. A straightforward argument shows that if the odds of no misuse exceeds an odds threshold defined by

$$w_{j-1j} = \frac{G_{j-1} - G_j}{H_j - H_{j-1}}$$

then the decision maker should prefer act R_j over R_{j-1}. The odds threshold w_{j-1j} therefore is determined by the ratio of utility differences.

Table 3: Two-Fold Partition of Acts

	R_1	R_0
	Laissez-F.	Prohibit
No misuse	H_1 = 1	H_0
Misuse	G_1 = 0	G_0

The simplest setup of this kind is shown in Table 3. There are two possible acts: Prohibition or Laissez-Faire. Without loss of generality we may assign H_1 = 1 (the best possible outcome) and G_1 = 0 (the worst). Therefore, the odds threshold is

$$w_{j-1j} = \frac{G_0}{1 - H_0}.$$

It immediately follows that if $G_0 < q_0$ for $0 < q_0 < 1$ then

$$H_0 > 1 - q_0 / w_{01}.$$

Suppose we also wish to restrict $w_{01} > y_0 > 1$. This should seem reasonable, because we are merely restricting the odds-of-no-misuse threshold to be above 1. Then

$$H_0 > (y_0 - q_0) / y_0.$$

For example, if q_0 = .1 and y_0 = 10 then H_0 > .99; and in fact if q_0 = 1 and y_0 = 100 then we also have H_0 > .99. Thus, no misuse under prohibition has nearly as high utility as no misuse under laissez-faire, implying that prohibition hardly decreases utility at all. Moreover, in the special case where prohibition obviates misuse so that $G_0 = H_0$, a high odds threshold yields a correspondingly high value for G_0 and H_0. For instance, y_0 = 10 implies G_0 and H_0 both must exceed 10/11.

The problem is the inability to simultaneously have a high value of y_0, a low q_0 and a relatively low H_0. The chief result is that a high (and therefore cautious) odds-of-no-misuse threshold for invoking the prohibition of research requires a belief that prohibition results in only a very small decrease in utility relative to the improvement in the (dis)utility of misuse.

As in the legal standard of proof case, this difficulty arises because we have only two possible acts. A way around this is to introduce a third act (middle option). Let us call it "Regulate". Table 4 shows the utility setup for this 3-fold partition.

Table 4: Three-Fold Partition of Acts

	R_2	R_1	R_0
	Laissez-F.	Regulate	Prohibit
No misuse	$H_2 = 1$	H_1	H_0
Misuse	$G_2 = 0$	G_1	G_0

The w_{01} threshold now determines when the Regulate option is chosen over Prohibit, and a new threshold, w_{12}, determines when Laissez-Faire is chosen over Regulate. Now, $H_1 > (y_1 - q_1)/y_1$ implies

$$w_{12} < \frac{G_1 - q_1}{(y_1 - q_1)/y_1 - H_1}$$

which in turn implies

$$H_0 > 1 - (G_1 - q_1)/w_{12} - q_1/y_1.$$

Setting $w_{12} = 5$ and $G_0 = .5$, for instance, and using the settings $q_1 = .1$ and $y_1 = 10$ gives

$$H_0 > 1 - (.5 - .1)/5 - .1/10 = .91.$$

If we are willing to lower the threshold to $w_{12} = 2$ and increase G_0 to .68 then

$$H_0 > 1 - (.68 - .1)/2 - .1/10 = .7.$$

The 3-fold partition therefore can express a belief that outright prohibition could substantially negatively affect research (in this last example, a decline in utility from 1 to .7). Nevertheless, there are limits if we take certain additional constraints into account. It seems reasonable to stipulate that misuse cannot yield a greater utility than no misuse, so we impose the constraint $G_0 < H_0$. As mentioned earlier, the case where $G_0 = H_0$ corresponds to the situation where prohibition of research eliminates the possibility of misuse of its outputs, so that there is no difference between the ``no misuse'' and ``misuse'' states. The restriction $G_0 < H_0$ and the constraint $w_{01} = 1$ imply that $H_0 > 1/2$. Higher odds thresholds increase the lower bound on H_0. It is easy to prove that the general relationship is $w_{01} = x$ implies $H_0 > x/(x + 1) = p_{01}$, the corresponding probability threshold. In the two examples above, $w_{01} = 5$ implies $H_0 > 5/6$ and $w_{01} = 2$ implies $H_0 > 2/3$.

Thus, extreme cases where prohibition of further research would hardly alter the (dis)utility of misuse of an existing technology impose severe restrictions on the utility if there is no misuse. Table 5 shows a setup like this, with similar low values of G_1 and G_2. We would be inclined to set the odds thresholds w_{01} and w_{12} to be very high, say $w_{01} = 100$ and $w_{12} = 1000$. The result would be that H_1 and H_0 both would be very close to 1: H_1 =.99999 and H_0 = .99989. Therefore, a substantial difference between H_1 and H_0 (say, due to the inhibition of scientific progress) can only arise if there is a substantial difference between G_1 and G_0 and a relatively low threshold odds of no misuse w_{01}.

Table 5: Extreme Disutility of Misuse

	R_2	R_1	R_0
	Laissez-F.	Regulate	Prohibit
No misuse	$H_2 = 1$	H_1	H_0
Misuse	$G_2 = 0$	$G_1 = .01$	$G_0 = .02$

Are there sets of utilities and threshold odds that could satisfy both the intuition that some security measures should be in place when there is only a very small chance of misuse, but that severe restrictions on research will have a substantial impact on scientific progress? What would these look like? Table 6 illustrates a setup similar to an earlier example that is compatible with these intuitions. The bottom row shows the odds thresholds. Resetting w_{12} to values greater than 10 has relatively little impact on w_{01} (or alternatively on utilities G_0 and H_0 if we wish w_{01} to remain at 2) because H_1 is already close to 1 and G_1 is close to 0. And of course it is possible to solve for H_1 and G_1 such that w_{12} takes a specific value greater than 10 while w_{01} is unaffected and remains at 2. Thus, in the 3-fold partition of acts we are free to set w_{12} to very conservative (high) values while still retaining flexibility regarding w_{01} or the utilities that comprise it.

Table 6: Extreme Disutility of Prohibition

	R_2	R_1	R_0
	Laissez-F.	Regulate	Prohibit
No misuse	$H_2 = 1$	$H_1 = .99$	$H_0 = .6933$
Misuse	$G_2 = 0$	$G_1 = .1$	$G_0 = .6933$
Odds thr.	$w_{12} = 10$	$w_{01} = 2$	

The setup is greatly affected, however, by changes in w_{01} because of the relationship described earlier between w_{12} and the lower bound on H_2. Increasing w_{01} from 2 to 5, as mentioned earlier, raises the lower bound on H_0 from 2/3 to 5/6. Preferences and intuitions regarding these effects will need to be guided by a sense of how harmful potential misuses are under prohibition versus regulation versus laissez-faire in comparison with the loss of potential knowledge and benefits when research is prohibited versus regulated. These comparisons are admittedly not easy to make, let alone quantify. Nevertheless, decisional thresholds do need to be set, and setting them in a considered manner requires comparisons of this sort.

Therefore some considerations about utility scales are appropriate to conclude this subsection. The utility scales used here are not absolute, or even ratio-level. They have neither an absolute zero nor a fixed upper bound. At best, they are interval-level scales, meaning that the difference between two utility assignments (e.g., $H_2 - H_1$) is a ratio-level scale. Recall that a ratio comparison of two such differences, $\left(G_{j-1} - G_j\right) / \left(H_j - H_{j-1}\right)$, determines the odds threshold w_{j-1j}. Smithson[21] defines two kinds of risk-orientation bias in the utility differences when utilities are restricted to the [0,1] interval. "A-bias" is measured by the sum of the log of the odds thresholds and refers to greater risk-aversion to one

[21] Smithson (2006)

outcome than the other. In our examples thus far, all $w_{j-1j} > 1$, indicating greater risk-sensitivity to misuse than to no misuse. "R-bias", on the other hand, is measured by

$$\sum_{j=1}^{J-1} \log\left(\frac{H_j - H_{j+1}}{H_{j-1} - H_j}\right) + \log\left(\frac{G_{j+1} - G_j}{G_j - G_{j-1}}\right)$$

and compares gains and losses in utility as the decision maker moves from one act to another. A positive sum indicates greater risk-sensitivity in choosing between acts for high *j*'s and a negative sum indicates greater risk-sensitivity in choosing between acts with low *j*'s. In Table 6 these log-ratios are 3.39 and 1.78, so there is greater risk-sensitivity in choosing between Regulate and Prohibit than between Laissez-Faire and Regulate. This is simply due to the greater changes in H and G utilities as we move from Regulate to Prohibit.

Finally, given that the utility scales have no absolute lower or upper bounds, a reasonable question to ask is whether some bounds are more useful or sensible than others. The [0,1] interval probably is not well-suited to human judgements because it lacks two features that have psychological significance: A reference-point representing the status quo and a distinction between being better off or worse off than the status quo. A well-established empirical and theoretical literature[22] informs us that people judge the utility of future outcomes relative to a reference point (usually the status quo) instead of in absolute terms, and that they are more sensitive to losses than to gains.

Table 7 presents one way of rescaling Table 6 according to these considerations. Suppose we assign 0 to represent the status-quo and represent the maximal loss by -100. Suppose also that we believe misuse of a research output under laissez-faire would yield a loss that is 10 times the magnitude of the gains that could be realized if no misuse occurred. Then G_2 = -100$ and H_2 = 10. The odds thresholds in Table 6 partially determine the remaining utility assignments. We require more one constraint, so let us repeat the loss due to misuse being 10 times the gain with no misuse under Regulate. The end result reveals that we believe we will be worse off than the status quo under the Prohibit option no matter whether there is misuse or not, but that will be our best option if the odds of misuse are shorter than 2 to 1.

Table 7: Rescaled Utilities from Table 6

	R_2	R_1	R_0
	Laissez-F.	Regulate	Prohibit
No misuse	H_2 = 10	H_1 = 9.9	H_0 = -26.4
Misuse	G_2 = -100	G_1 = -99	G_0 = -26.4

4. Imprecision and Bias in Judgements

Probability and utility judgements regarding dual use dilemmas ultimately must be made by human judges, and this last section discusses the most important issues regarding human judgements of this kind. We begin by considering issues of imprecision and conflict in judgements, and subsequently discuss relevant human tendencies toward over-confidence in predictions and confirmation bias.

[22] beginning with Kahneman & Tversky's (1979) Prospect Theory

Even when it is foreseeable, the probability of the misuse of a new technology or research output and the severity of its consequences almost never are known precisely, nor is there usually a consensus on their magnitudes. Imprecision and conflict are very likely to pervade judgements of probability and utility in dual use dilemmas. These uncertainties must not be denied or ignored; falsely precise estimates will be treated by decision makers as if they really are precise and decisions based on them will be far from robust. At the very least, decisions and their criteria should be subjected to sensitivity analyses to ascertain which components are the most affected by altering parameter values. In the preceding section, for instance, we saw that the three-option setup in Table 6 was robust against changes in w_{12} but sensitive to changes in w_{01}.

I shall leave conflict aside as even a brief treatment of it is beyond the scope of this paper, except to note in passing that some psychological investigations indicate that people prefer dealing with vague but consensual opinions to precise but disagreeing ones.[23] Thus, imprecision is viewed as a less severe kind of uncertainty than conflict.

Nevertheless, imprecision complicates decision making. A precise probability assigned to the misuse of a technology either exceeds or fails to exceed a decisional threshold of the kind discussed in the preceding section, so the choice among alternatives is clear. Precise probabilities bring decisiveness with them. However, a probability interval may lie entirely below or above the threshold, or may include it. Standard decision frameworks for imprecise probabilities treat the lower bound as the probability to use in betting on misuse and the upper bound as the probability to use in betting against it. Therefore, these frameworks claim there is no basis in the probabilities themselves for preferring the alternative on either side of a decisional threshold if the probability interval straddles it.

Suppose, for instance, that the setup in Table 6 is our decisional guide and we are confronted with a potential technological development for which experts estimate the odds of misuse to be somewhere between 5 and 50 to 1. This interval includes the threshold w_{01} = 10, so should we choose Regulate or Laissez-Faire? If we can defer this decision pending more information, should we do so? This issue is an active topic of research and attempting a resolution of it is beyond the scope of this paper, but the main purpose in raising it here is to point out that because imprecision really matters decision makers must work out how they will treat imprecise estimates differently from precise ones.

We now turn to probability judgements themselves. There is a large body of empirical and theoretical work on subjective probability judgements, but discussion will be restricted to just two judgement biases that are directly relevant. The first of these is probability weighting, which may be summarized by saying that people over-weight small and under-weight large probabilities. Note that this does not mean that people are necessarily under- or over-estimating the probabilities, but instead treating them in a distorted fashion when making decisions based on them. Rank-dependent expected utility theory[24] reconfigures the notion of a probability weighting function by applying it to a cumulative distribution whose ordering is determined by outcome preferences. Cumulative prospect theory[25] posits separate weighting functions for gains and losses.

[23] Smithson (1999b), Cabantous (2007)
[24] e.g., Quiggin (1993)
[25] Tversky & Kahneman (1992)

Two explanations have been offered for the properties of probability weighting functions. The first[26] is "diminishing sensitivity" to changes that occur further away from the reference-points of 0 and 1. The second is that the magnitude of consequences affects both the location of the inflection-point of the curve and its elevation. Large gains tend to move the inflection point downward and large losses move it upward.[27] Diminishing sensitivity has an implication for judgments and decisions based on imprecise probabilities as well as precise probabilities. A change from .01 to .05 is seen as more significant than a change from .51 to .55, but a change from .51 to .55 is viewed as less significant than a change from .95 to .99. An implication is that for decisional purposes people might view a probability interval [.01,.05] as less precise than [.51,.55], and so on. The prospect of large losses (as in the misuse of biotechnology) will exaggerate these effects for low probabilities. This issue is important for dual use dilemmas because at least some of the possible outcomes under consideration will have extreme probabilities attached to them.

The second relevant bias concerns confidence judgements and the elicitation of prediction or confidence intervals from human judges. Numerous studies demonstrate that both novices and experts tend to be overconfident in the sense that they construct prediction intervals that are much too narrow for their confidence criteria. A typical discrepancy is that when asked to construct an interval that has a 90% probability of including the correct prediction the actual hit-rate is below 50%.[28] However, recent findings[29] have suggested that when presented with prediction intervals people do not over-estimate their coverage-rates. The take-home lesson from this literature is that asking experts to estimate how likely the probability of, say, the theft of smallpox supplies from a particular source is between two values will yield more well-calibrated results than asking the experts to construct , say, a 95% confidence interval for that probability.

Finally, the catch-all underestimation bias described earlier is a special case of confirmation bias. This is a largely unconscious tendency in human information processing and judgement such that people seek out and pay more attention to information that confirms their beliefs than to disconfirming information. In the catch-all underestimation bias, confirmation bias manifests itself as a tendency to underestimate the likelihood of novel or unanticipated events. Unfortunately these are exactly the kind of events that policy planners and decision makers must be on the lookout for in dealing with dual use dilemmas. There are few recommendations on record for militating against confirmation bias. One is to construct inclusive teams containing members with diverse backgrounds and viewpoints and ensure that decision makers and planners listen attentively to those members with whom they disagree. However, this seemingly obvious strategy is complex and deceptively difficult to implement.[30] Another is the use of formal analyses, simulations and models to reveal consequences or possibilities that our preconceptions render invisible to us.[31]

---

[26] Camerer & Ho (1994)

[27] e.g., Etchart (2004)

[28] e.g., Russo & Schoemaker (1992)

[29] e.g., Winman, Hansson, & Juslin (2004)

[30] see, e.g., Brown (2010)

[31] see Lempert, Popper & Bankes (2002) for an example of this approach in the setting of policy making

5. Conclusions?

This paper largely neglects ethical considerations, which may seem odd given the predominantly ethical nature of dual use dilemmas. Ethical considerations have been set aside to enable a focus on some prerequisites for a ``fine-grained'' analysis of dual use dilemmas, namely a systematic investigation of specific unknowns that such an analysis would have to contend with. My hope is that ethicists will find useful guidance in this investigation, avoiding some of the pitfalls and traps awaiting the unwary.

Some pertinent unknowns have not been dealt with here, so this paper cannot be taken as anything like an exhaustive survey. Nevertheless, we have examined types of unknowns that are beyond the purview of standard decision theories, such as state space indeterminacy and imprecision. We have seen that there are genuinely different kinds of unknowns, not just different sources of the same kind, and that these play distinct roles. One of the key emergent points is that many of the unknowns in dual-use dilemmas (and in so-called "wicked problems") are interconnected. They can be traded against one another, and how one unknown is dealt with has ramifications for other unknowns. Allowing imprecision in probability assignments, for instance, offers a way of handling state space indeterminacy. Conversely, choosing the "right" number of options can rectify incompatibilities between preferences and decisional probability thresholds.

We may never be able to attain precise quantification of costs, benefits, and probabilities of outcomes arising from dual use dilemmas, so a fine-grained analysis in that sense also is unachievable. After all, accidental findings and consequences are legion in cutting-edge research and development, and so irreducible unknowns such as the catch-all under-estimation problem are likely to dog policy formation and decision making alike. Moreover, as I have argued elsewhere[32], if we value creativity, discovery, and/or entrepreneurship then we shall have to tolerate at least some irreducible unknowns.

Nevertheless, as Head[33] has pointed out, great uncertainty alone is not sufficient to render a problem "wicked" in the sense used in most of the literature on that topic. Wickedness also requires complexity and divergent or contradictory viewpoints about the nature of the problem and preferences regarding alternative outcomes. I have tried to show here that even rather simple formal analyses in the form of thought-experiments can frame and structure dual use dilemmas in useful ways that avoid some aspects of wickedness, so that at least some of our psychological foibles can be taken into account and even overcome.

References

Brown, V.A. 2010, Collective inquiry and its wicked problems, In V. A. Brown, J. Russell, and J. Harris (eds), *Tackling wicked problems through the transdisciplinary imagination*, Earthscan, London, pp. 61-84.

Cabantous. L. 2007, Ambiguity aversion in the field of insurance: Insurers' Attitude to imprecise and conflicting probability estimates, *Theory and Decision, vol. 62*, pp. 219–240.

[32] Smithson (2008)
[33] Head (2008)

Camerer, C. F. & Ho, T. H. 1994, Violations of the betweenness axiom and nonlinearity in probability, *Journal of Risk and Uncertainty, vol. 8*, pp. 167-196.

Connolly, T. 1987, Decision theory, reasonable doubt, and the utility of erroneous acquittals, *Law and Human Behavior, vol. 11*, pp. 101-112.

Danielson, P. 1992, *Artificial morality: Virtuous robots for virtual games*, Routledge. London.

Etchart-Vincent, N. 2004, Is probability weighting sensitive to the magnitude of consequences? An experimental investigation on losses, *Journal of Risk and Uncertainty, vol. 28*, pp. 217-235.

Fischhoff, B., Slovic, P., & Lichtenstein, S. 1978, Fault trees: Sensitivity of estimated failure probabilities to problem representation, *Journal of Experimental Psychology: Human Perception Performance, vol. 4*, pp. 330–344.

Fox, C. R. & Rottenstreich, Y. 2003, Partition priming in judgment under uncertainty, *Psychological Science, vol. 13*, pp. 195–200.

Grant, S. & Quiggin, J. 2004, *Conjectures, refutations and discoveries: incorporating new knowledge in models of belief and decision under uncertainty*, Paper presented at the 11th International Conference on the Foundations and Applications of Utility, Risk and Decision Theory (FUR XI--Paris), under the joint auspices of the Ecole Nationale Supérieure d'Arts et Métiers (ENSAM) and the Ecole Spéciale des Travaux Publics (ESTP), Paris, 2 July.

Head, B.W. 2008, Wicked problems in public policy, *Public Policy, vol. 3*, pp. 101-118.

Kahneman, D. & Tversky, A. 1979, Prospect theory: An analysis of decision under risk, *Econometrica, vol. 47*, pp. 263—291.

Lempert, R., Popper, S. & Bankes, S. 2002, Confronting surprise, *Social Science Computer Review*, vol. 20, pp. 420-440.

Miller, S. & Selgelid, M. 2007, Ethical and philosophical consideration of the dual-use dilemma in the biological sciences, *Science and Engineering Ethics, vol. 13*, pp. 523–580.

Ostrom, E. Gardner, R. & Walker, J. 1994, *Rules, games and common pool resources*, The University of Michigan Press, Ann Arbor.

Quiggin, J. 1993, *Generalized Expected Utility Theory: The Rank Dependent Model*, Kluwer, Boston.

Russo, J. E. & Kolzow, K. J. 1994, Where is the fault in fault trees? *Journal of Experimental Psychology: Human Perception Performance, vol. 20*, pp. 17–32.

Russo, J. E., & Schoemaker, P. J. 1992, Managing overconfidence, *Sloan Management Review, vol. 33*, pp. 7-17.

Shafer, G. 1976, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton.

Smithson, M. 1999a, Taking exogenous dynamics seriously in public goods and resource dilemmas, In M. Foddy, M. Smithson, S. Schneider, & M. Hogg (eds.) *Resolving Social Dilemmas: Dynamic, Structural, and Intergroup Aspects*, Psychology Press, Philadelphia, pp. 17-32.

Smithson, M. 1999b, Conflict aversion: Preference for ambiguity vs. conflict in sources and evidence, *Organizational Behavior and Human Decision Processes, vol. 79*, pp. 179-198.

Smithson, M. 2006, Scale construction from a decisional viewpoint, *Minds and Machines, vol. 16*, pp. 339-364.

Smithson, M. 2008, The many faces and masks of uncertainty, In Bammer, G. & Smithson, M. (eds.) *Uncertainty and Risk: Multidisciplinary Perspectives*, Earthscan, London, pp. 13-26.

Smithson, M. 2009, How many alternatives? Partitions pose problems for predictions and diagnoses, *Social Epistemology, vol. 23*, pp. 347-360.

Smithson, M., Gracik, L. & Deady, S. 2007, Guilty, not guilty, or … ? Multiple verdict options in jury verdict choices, *Journal of Behavioral Decision Making, vol. 20*, pp. 481-498.

Smithson, M. & Segale, C. 2009, Partition priming in judgments of imprecise probabilities, *Journal of Statistical Theory and Practice, vol. 3*, pp. 169-182.

Tversky, A. & Kahneman, D. 1974, Judgment under uncertainty: Heuristics and biases, *Science, vol. 185*, pp. 1124–1131.

Tversky, A. & Kahneman, D. 1992, Advances in prospect theory: Cumulative representation of uncertainty, *Journal of Risk and Uncertainty, vol. 5*, pp. 297-323.

Walley, P. 1991, *Statistical Reasoning with Imprecise Probabilities*, Chapman Hall, London.

Walley, P. 1996, Inferences from multinomial data: Learning about a bag of marbles, *Journal of the Royal Statistical Society, Series B, vol. 58*, pp. 3-34.

Winman, A., Hansson, P. & Juslin, P. 2004, Subjective probability intervals: How to cure overconfidence by interval evaluation, *Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 30*, pp. 1167-1175.