Quaestio facti. Revista Internacional sobre Razonamiento Probatorio
Quaestio facti. International Journal on Evidential Legal Reasoning
Sección: Ensayos en homenaje a Larry Laudan
2025 I 10 pp. 1-30
Madrid, 2025
DOI: 10.33115/udg_bib/qf.i10.23144
Marcial Pons Ediciones Jurídicas y Sociales

ISSN: 2604-6202
Recibido: 25/04/2025 | Aceptado: 0/11/2025 | Publicado *online*: 19/12/2025

STANDARDS OF PROOF FOR FUTURE CRIMES AND DECISION THEORY

Hylke Jellema
Willem Pompe Institute for Criminal Law and Criminology,
Utrecht University

ABSTRACT: What is the proof standard for applying preventive criminal sanctions? This is an open question in various legal systems. Some authors suggest that we can answer it by using decision theory. On this approach, the proof standard is conceptualized as a probabilistic threshold: a preventive sanction can only be imposed if it is sufficiently probable that a person will commit a crime in the future. According to decision-theorists, how high this probability of a future crime should be, can be determined by means of a utilitarian calculus. However, such a decision-theoretic analysis requires wrestling with a number of difficult questions. This article surveys these questions and explores some avenues for answering them. It does so by considering a standard of proof for a fictional preventive sanction and offering a decision-theoretic justification for that standard.

KEYWORDS: Preventive sanctions; decision theory; legal proof; evidence theory; dangerousness.

SUMMARY:

1. INTRODUCTION

In many countries, the (criminal) legal system is increasingly being used as a tool for preventing future crime (Carvalho, 2017). Part of this trend is the introduction of various criminal sanctions intended to prevent recidivism[1]. These are sanctions that can only be applied to a person if (among other things) they are deemed "dangerous" —i.e., likely to engage in criminal behavior in the future[2]. But when is someone sufficiently dangerous to be the subject of these sanctions? Another way of phrasing this question is "what is the proof standard for future crimes?" This is an open question in various legal systems (see e.g., Bijlsma & Meynen 2023; Tadros, 2013; Scurich, 2016; Schopp, 1996; Schopp & Quattrocchi, 1995; Slobogin, 1989; 2006).

One way of thinking about proof standards for dangerousness is as probabilistic thresholds that specify the minimum level of certainty that is required to prove a given factual statement. In the case of past crimes, such a factual statement may be "that the defendant committed the alleged criminal acts". For preventive sanctions, it could be "the defendant will (again) commit criminal acts in the near future". Or it may be a more specific proposition, such as "if not committed, the defendant will commit a violent act within the next year." But how high should we set the probabilistic threshold? One method for answering this question is *decision theory*. This is an approach to decision-making that is used in various fields, including economics, psychology and the law (Scurich & John, 2011, p. 90). It has also been extensively applied to legal standards of proof, especially the beyond a reasonable doubt standard (see e.g., Vorms & Hahn, 2021).

The underlying idea of the decision-theoretic approach to proof standards is that any proof standard should be set at the level where applying a (preventive) sanction maximizes expected utility. Where this point lies depends on the relative desirability (i.e., utilities) of the possible outcomes of a decision (not) to apply a (preventive) sanction. For instance, for past crimes the relative desirability of the possible outcomes is usually linked to Blackstone's (1962) remark that it is better that ten guilty persons escape than that one innocent suffer. In other words, a false conviction is at least ten times worse than a false acquittal[3]. This fact is commonly thought to jus-

[1] This is also known as risk-based sentencing (Eaglin, 2017).

[2] Whether someone is "dangerous" depends on the probability that they will cause future harm as well as on the magnitude of this harm (Scurich & John 2010, p. 446-7). Slobogin opts for a more fine-grained account, according to which "risk" comprises four elements: (i) a probability that (ii) a particular type of offence outcome will occur within (iii) a specific period of time (iv) in the absence of an intervention (Slobogin, 2021, p. 38).

[3] This ratio should not be taken too literally; various other ratios have been proposed. For examples, such as Voltaire's claim that "Tis much more Prudence to acquit two Persons, tho' actually guilty, than to pass Sentence of Condemnation on one that is virtuous and innocent" or Benjamin Franklin who wrote that "it is better a hundred guilty persons should escape than one innocent person should suffer" (Laudan, 2006, p. 63).

tify a high criminal standard of proof, somewhere between 90% and 99% (see e.g., Jellema, 2023, p. 102)[4]. After all, the higher we set the standard, the more difficult it will be to convict someone, and therefore fewer false convictions will occur. Yet a high standard also means that more people will be acquitted, including some who are guilty. Blackstone's remark implies that we are willing to tolerate a fair amount of such false acquittals to prevent a smaller number of false convictions and that the standard for proof beyond a reasonable doubt should therefore be high. For preventive sanctions, the assumptions that ought to guide a decision-theoretic analysis of the proof standard are less clear[5].

The literature on decision theory and proof standards has mostly been concerned with the beyond a reasonable doubt standard for past crimes. Nonetheless, there have been several authors who have applied the approach to preventive sanctions (Nagel, Neef & Schramm, 1977; Mossman, 1995; Scurich & John, 2010; 2011; 2012; Vars, 2012; Scurich, 2015; 2016; 2018). This body of work on decision theory and proof standards includes a number of scholars, such as Larry Laudan, who are skeptical of this approach (Laudan & Saunders, 2009). The result of this sprawling discussion has arguably been greater confusion rather than clarity. What we *can* conclude is that applying decision theory to proof standards means wrestling with a number of difficult philosophical, legal and practical questions. The goal of this article is to explore some of these questions as well as possible ways of answering them. In order to make this exploration as clear as possible, I use a fictional but realistic example of a preventive sanction and propose a proof standard for that sanction[6]. I then ask how one might justify this standard on decision-theoretical grounds, what questions this raises and how one might answer these questions.

The structure of this article is as follows. First, I explain the basics of decision theory (section 2). After that, I describe the preventive sanction and the associated standard of proof that will be used as a running example are introduced (section 3). Next, I offer a decision-theoretic justification of this proof standard (section 4). One of the key ideas developed in this section is that this justification can be reached by means of a "reflective equilibrium"—approach. Additionally, this part of the article also discusses the underexamined question whether we should include the utilities of correct outcomes when determining the proof standard, or whether we can limit ourselves to the utilities of errors. Section 5 deals with the question whether proof

[4] Though see section 3 for criticisms of the relationship between the Blackstonian ratio and the standard of proof.

[5] Though some suggest that a low proof standard is warranted because they consider a dangerous person who goes on to commit a crime (a false negative) worse than a preventive sanction being applied to a non-dangerous person (Monahan, 1977).

[6] The example used is fictional for two reasons. First, I do not mean to give the image that the aim of this article is to make a policy proposal about some existing preventive sanction. Rather, the goal is to explain decision theory. Second, the discussion in this article is philosophical in nature. I want to sidestep legal questions regarding criminal procedural law.

standards for future crimes can best be expressed in vague but flexible, qualitative terms or precise but inflexible probabilities. Finally, section 6 explores two questions regarding the relationship between predictions of dangerousness by experts and the fact-finder's (judge or jury) determination whether the proof standard is met. The first question is to what extent normative, utilitarian assumptions can be (or should be) made by the expert rather than the fact-finder. Second, I discuss the idea proposed by some authors and adopted in a number of legal systems that, aside from the probability of future crime, there should also be a proof standard for how confident we are in that probability.

2. DECISION THEORY AND PROOF STANDARDS

Decision theory is a method to balance costs and benefits in a decision dilemma. Before we turn to explaining this approach, it is helpful to distinguish it from a closely related idea with which it is sometimes confused. Many authors assume that a key aim of the beyond a reasonable doubt standard is to distribute errors fairly (see e.g., Jellema, 2023, p. 102). The underlying idea is that the higher we set the proof standard, the less likely it is that someone will be falsely convicted. However, this reduction in false convictions comes at a cost—the higher the proof standard, the more often defendants will be falsely acquitted. After all, the higher the proof standard is, the larger the group of defendants will be with strong evidence against them, but where the evidence is not sufficient to meet the standard of proof. The idea defended by various authors is that the proof standard should be set such that the criminal legal system produces roughly ten false acquittals for every false conviction (to the extent that such errors cannot be avoided). This ratio is based on Blackstone's (1962), that it is better that ten guilty persons escape than that one innocent suffer. This suggestion—that the proof standard ought to act as a method for distributing errors, has also been suggested in the context of preventive sanctions (see e.g., Monahan, 1977). The question for preventive sanctions would then be how many false positives (individuals subject to preventive sanctions, who would not have committed crimes in the future) are acceptable compared to every false negative (individuals to whom no preventive sanction was applied who went on to commit a crime).

While popular, this notion of error distribution has been the subject of critique. For example, as a number of authors note the distribution of errors depends on other factors aside from the proof standard (DeKay, 1996; Lillquist, 2002, Laudan & Saunders, 2009; Scurich & John, 2010). This includes unknowable factors such as how many defendants are actually innocent and guilty, the quality of the evidence against them and the degree to which courts or juries accurately evaluate this evidence. Furthermore, Laudan (2015) argues that (at least in the American criminal legal system) other criminal procedural rules, aside from the standard of proof influence the error distribution. In particular, there are numerous rules aimed at benefiting the defendant, for instance regarding the admissibility of evidence, thereby further shifting the error distribution towards false acquittals. So, no matter how one sets the proof standard, we cannot guarantee that this will result in the desired error distribution.

In response to the aforementioned arguments, various scholars writing on the beyond a reasonable doubt standard suggest that Blackstone's ratio should not be taken as stipulating an optimal distribution of errors (see e.g., Lillquist, 2002; Laudan & Saunders, 2009). Instead, this ratio can also be interpreted as specifying the relative *utilities* of the two types of errors. This idea lies at the core of the decision-theoretic approach, according to which we can use decision theory to clarify standards of proof. The first to explore this approach were Kaplan (1967) and Cullison (1969), working independently. Their decision-theoretic analysis of the beyond a reasonable doubt standard begins with the assumption that the disutility of a false conviction is ten times greater than that of a false acquittal. These relative utilities can then be used to calculate how high this proof standard ought to be set so that the desirability of convicting would either be equal to or above the desirability of acquitting.

What would such a decision-theoretic analysis look like? The decision-theoretic framework relies on three types of variables: the courses of action one can decide to take, the probability that certain outcomes will materialize if a given course of action is chosen and the utilities associated with those outcomes. These utilities should not be thought of as the costs or benefits that an outcome *has* in some absolute sense. Rather, the notion of utility is simply a means of comparing how "good" and "bad" the potential outcomes are relative to one another (Scurich, 2016, p. 172; Scurich & John, 2010, p. 432).

To give a commonly used example: suppose that you have to decide whether to bring an umbrella when going outside or not. If you bring the umbrella, you would have to carry it, which is slightly cumbersome. In other words, bringing the umbrella involves a small amount of disutility. However, if you do not bring an umbrella and it rains, you will probably get wet, which you want to avoid. The disutility of getting wet is much greater than that of bringing the umbrella. So, should you bring it? Aside from the relative utilities of these outcomes, this also depends on the probability that these outcomes will come about. In this example, the most important probability is that of rain. If it is extremely unlikely that it will rain, it may not be worthwhile to bring an umbrella, because it is very likely that carrying it will not yield any benefits. If the chances of rain are very high, it will probably be in your best interest to bring it. It is then very likely that it will help you avoid getting wet. What you are trying to estimate is what choice—to bring or not bring an umbrella—maximizes your expected utility. The higher the probability of rain, the greater the expected utility gained by bringing the umbrella is. There is a certain point where the probability of rain is high enough that it is better to bring the umbrella than to not bring it. It is at that point that you should take your umbrella with you.

"If the probability of event X is at least probability P, take decision Y". The X and Y in this statement can be substituted for different kinds of events and decisions. This includes legal decisions. For instance, "if the probability that the defendant committed the alleged criminal acts is at least 95%, convict them". This last sentence is one way to understand the proof of guilt beyond a reasonable doubt standard. As said, for criminal cases, most scholars agree that a false conviction is much worse than a false acquittal. And, according to many, this implies that the level of probability for proof of guilt beyond a reasonable doubt ought to be very high. For instance, according to Kaplan (1967), if we assume that a false conviction is ten times worse than a false acquittal, we end up with a proof standard of 91%. The formula Kaplan uses to arrive at this figure is as follows:

$$\text{Proof standard} = \frac{1}{1 + (U_{\text{false negative}} / U_{\text{false positive}})}$$

Where "U" standards for "utility". In the context of past crimes, a false negative is the acquittal of a guilty person. For future crime, it refers to a person to whom no preventive sanction was applied, who went on to commit a crime. In the context of past crimes, a false positive refers to an innocent person who is convicted. For future crime, a false positive occurs when we apply a preventive sanction to a person who would not have committed a future crime. When we take into account the relative utilities of these two types of errors, we arrive at a number between 0 and 1, expressing a proof standard between 0% and 100%.

While elegant, Kaplan's (1967) idea that proof standards can be determined by considering the utilities of errors has been criticized by various authors. The problem with his proposal is that there are more than two possible outcomes of a trial. More precisely, there are four:

Table 1 The four possible outcomes of the decision to convict or acquit

	Convicted	**Acquitted**
Guilty	True conviction (true positive)	False acquittal (false negative)
Innocent	False conviction (false positive)	True acquittal (false positive)

We can see that, aside from errors (false positives and negatives), a trial can also result in a true positive (the criminal conviction of a guilty person) or a true negative (the criminal acquittal of an innocent person). The same is true of preventive sanctions. Here the possible decisions are (i) applying a preventive sanction, or (ii) not applying such a sanction. Both decisions can be correct (true) or an error (false):

Table 2 The four possible outcomes of the decision (not) to apply a preventive sanction

	Sanction applied	**Sanction not applied**
To dangerous person	True positive	False negative
To non-dangerous person	False positive	True negative

For instance, a true positive occurs when we apply a preventive sanction to a person who would otherwise have reoffended (i.e., a "dangerous person"). A false negative means that no preventive sanction is applied, yet the person turned out to be dangerous and goes on to commit a crime. Yet while there are four possible outcomes, Kaplan's formula only takes into account two of them. As various authors have argued, decision theory is incomplete if it does not include the utilities associated with correct outcomes (e.g., Tribe, 1971; Lillquist, 2002; Laudan & Saunders, 2009; Nance, 2016, p. 24-5). To solve this problem, Tribe (1971) proposed a formula to calculate the standard of proof that includes all four outcomes:[7]

$$\text{Proof standard} = \cfrac{1}{1+\cfrac{(U_{\text{false negative}} \,/\, U_{\text{false positive}})}{(U_{\text{true negative}} - U_{\text{false positive}})}}$$

Including correct outcomes can have a significant effect on the resulting proof standard. For example, recall Blackstone's widely shared intuition that a false conviction is ten times worse than a false acquittal. Laudan & Saunders, (2009, p. 11-12) list various existing proposals that include this assumption but make different (reasonable) assumptions about the utilities of correct outcomes. Depending on which of these assumptions we adopt, the proof standard can be anywhere between 55% and 95%[8]. Let us consider one of these proposals to illustrate how the formula above works. Lillquist (2002, p. 109) offers the following example of how one could set the required utilities in his discussion of the beyond a reasonable doubt standard:

True acquittal: 0, True conviction: 1, False acquittal: -1, False conviction: -10.

He justifies these figures as follows: he assumes that there is no utility (positive or negative) for a true acquittal. He assigns a positive utility of 1 for true convictions, because it is the best possible outcome. Then he proposes that the downside of a false acquittal is that a guilty person is not convicted and that, hence, the negative social utility associated with this is the converse of the positive utility of a true conviction, i.e., -1. Finally, Lillquist assumes that a false conviction is ten times worse than a false acquittal, namely -10. Working from these assumptions, the resulting standard of proof would be 83%[9].

Because the utilities of correct outcomes can have a large impact on the resulting standard of proof there seems to be a good reason to include them when calculating this standard. However, there is also a downside to this. The equation that includes

[7] This formula was simplified by Laudan & Saunders (2009, p. 5) as follows:
Proof standard = $(U_{\text{true negative}} - U_{\text{false positive}}) / \{(U_{\text{true positive}} - U_{\text{false negative}}) + (U_{\text{true negative}} - U_{\text{false positive}})\}$

[8] Furthermore, they argue that certain plausible assumptions regarding these utilities could even lead to a standard of proof below 50% (Laudan & Saunders, 2009, p. 12).

[9] 1 / {1 + [(1 - -1) / (0 - -10)] } = 0.833. However, we can easily justify different numbers, that would result in a different standard of proof (Lillquist, 2002, p. 110).

all four outcomes makes it much harder to reach conclusions about the desired proof standard. For instance, if we want to conclude that the proof standard is high, it is no longer enough that the disutility of an erroneous conviction be much greater than the disutility of an erroneous acquittal. Instead the value of $U_{\text{true positive}} - U_{\text{false negative}}$ has to be greater than the value of $U_{\text{true negative}} - U_{\text{false positive}}$. Furthermore, as Laudan & Saunders (2009, p. 13-14) point out, the ratio I just mentioned is not a ratio of utilities, but a ratio of differences between utilities. One consequence of this is that we could, for example, subtract 10 from each of the utilities mentioned above by Lillquist (2002) and the result of the formula would be the same. However, in that case the ratios of the utilities would be quite different. For example, the utilities assigned to the two errors would be -11 for a false acquittal and -21 for a false conviction. These no longer stand in a 1 to 10 ratio to one another. In other words, utility ratios, such as Blackstone's ratio now tells us next to nothing at all about how high the proof standard ought to be. We can no longer just ask, "how much worse is a false positive compared to a false negative?" Instead, we have to answer the much more complicated question, what is greater, the utility difference between a true positive and a false negative, or the utility difference between a true negative and a false positive?

We now face a dilemma. On the one hand, it seems that we should not ignore correct outcomes, lest we end up with a proof standard that does not accurately reflect our utilitarian assumptions. On the other hand, including them means that it becomes much more difficult to determine how high the proof standard ought to be. Some authors attempt to circumvent this dilemma by making the simplifying assumption that the utilities of correct outcomes can be ignored. For instance, in the example of whether to take an umbrella, we did not consider the utilities of correct outcomes. This assumption is warranted if the utilities of both correct outcomes (taking the umbrella when it rains and not taking the umbrella when it does not rain) are roughly equal. In that case, Tribe's formula can be reduced to Kaplan's, where we only have to take into account the ratio of utilities of errors, a much simpler endeavor. As I argue in section 4, this solution is feasible for a decision-theoretic account of the proof standard for future crimes (while it may not be feasible for such an account of proof standards for past crimes).

There are further reasons why proof standards for preventive sanctions lend themselves more readily to decision-theoretic analysis than the standards for past crimes. First, decision theory involves a utilitarian calculus. Some may find this worrisome when it comes to decisions about past crimes, as these involve such unquantifiable considerations as justice, guilt and desert. Weighing these values against monetary costs, or against practical benefits such as deterrence may seem inappropriate. Yet preventive sanctions are part of the shift to preventive justice, where (criminal) law is used as a means to prevent future crime. This is an inherently utilitarian project (and is therefore also subject to the usual objections to the utilitarianism, such as that it ignores such values as justice). Second, decision theory involves analyzing proof standards in terms of quantitative probabilities. Tribe (1971) argued that such quantification leads to a number of practical difficulties, including whether jurors or courts will be able to understand probabilistic instructions. However, dangerousness criteria involve predictions of the future, which is inherently a probabilistic assessment (whereas proof of past crimes is only probabilistic according to certain theories of evidential reasoning). This is all the more so because courts often base their judgments of danger based on structured risk assessment estimates experts that involve quantitative language (Vars, 2010, p. 873-4; Eaglin, 2017). Hence, this worry also seems less great when it comes to proof standards for future crime (Scurich & John, 2010).

Having explained the basics of decision theory, let us turn to the hypothetical preventive sanction and associated proof standard that will be used as a running example in this article.

3. A HYPOTHETICAL STANDARD OF PROOF

In this section I describe the proof standard that we will use as an illustrative example throughout this article. As said, this standard is intended to be fictional but realistic. Suppose that a criminal legal system allows for the following preventive sanction:

> *Treatment facility for repeat violent offenders:* Those who have been convicted of a violent crime on multiple occasions can be committed to a specialized institution to prevent recidivism. This commitment is for a period of up to two years. Within this detention center treatment options are offered, aimed at resocialisation.

Suppose that several legal criteria have to be met before this measure can be imposed. This includes a ´dangerousness´ criterion: there must be a danger that the felon will commit another violent offense if this preventive measure is not imposed. This dangerousness requirement is encapsulated in the following legal provision:

> *Proof standard:* The preventive sanction can only be applied if it is likely that the person will commit serious harm to others.

This proof standard specifies a minimum probability of violent recidivism. In this case, the word "likely" is intended to express a probability of around 50% that the person will commit certain acts. This required level of probability is justified using decision-theory in the next section. However, the legal provision does not contain a precise, quantitative probability, as the standard allows for some degree of flexibility in the minimum level of dangerousness required for applying a sanction. This flexibility is desirable on decision-theoretic grounds (see section 5).

4. JUSTIFYING THE THRESHOLD

The proof standard proposed in the previous section states that it must be "likely" that a person will commit serious harm to others. The word ´likely´ is a verbal expression of a probability. Such verbal expressions can, and often are, interpreted in different ways within different contexts and by different people (see e.g., Willems, Albers & Smeets, 2020). In the current context the word “likely” is intended to convey a probability of recidivism around 50%[10], though the exact value may vary from case to case (see the next section). In this section, I explain how one could justify a standard of 50% for the fictional preventive sanction described in section 3. The overarching aim of this section is to show the kinds of questions one runs into during such a decision-theoretic justification and to explore some ways of answering them. In particular, one of the main ideas developed in this section is to propose a “reflective equilibrium”—approach to this process of justification (see 4.2).

4.1. Rejecting the empirical approach

The first step in our decision-theoretic analysis is to come up with utilities to put into Tribe’s equation mentioned above. How do we do this? One common answer is that the proof standard for future crimes should be set according to the preferences of society in general or of a more specific subgroup such as lawmakers, psychiatrists or judges. What those preferences are is a question that can be, and has been, empirically investigated. Here are a few examples. Mossman & Hart (1993) investigated societal attitudes towards civil commitment. They asked participants which was worse, “being attacked by a man with a knife [a false negative, failing to commit a dangerous individual], or spending a certain time period as a patient in a state psychiatric hospital [a false positive, committing a non-dangerous individual]’. Similarly, Mossman (2006) asked mental health professionals to compare having to spend various lengths of time as a patient in state hospital to being attacked by a man wielding a knife. Scurich & John (2011) had two stakeholder groups, former mental patients and psychiatrists fill in a questionnaire about the use of restraints in psychiatric hospitals. They asked them which was worse and by how much, the unnecessary use of such restraints on patients (a false positive) or failing to use restraints when required (a false negative).

One problem with these studies is that they focus only on the costs associated with errors. As we saw in the previous section, this perspective may be too limited (however, see below for a defense of this approach). Laudan & Saunders (2009, p. 18) propose a more comprehensive approach to eliciting utilities comes in the context of the beyond a reasonable doubt standard for past crimes. According to Laudan and Saunders (2009), we can ask fact-finders (judges or jurors) how much they would pay to convert a false acquittal to a true conviction or a false conviction into a true acquittal. This approach would take all four outcomes into account.

[10] See Janus & Meehl who show that several US courts use the “likely” standard for civil commitments, but are typically unclear about what probability value they intend to convey with this. The authors tentatively propose that this standard can be translated into a minimum probability of recidivism of 50% (Janus & Meehl, 1997, p. 41).

Yet all these studies face a debilitating problem. If existing research shows anything, it is that people's intuitions about utilities differ wildly. For example, the studies of Mossman & Hart (1993) and of Mossman (2006) received such a variability in responses that summary statistics (e.g., the average or mean) would likely to have been misleading (Scurich, 2016, p. 175). The same problem appears in the study by Scurich & John (2011). They found that former mental health patients thought false positives were much worse than false negatives. The median patient tradeoff was that 141 false negatives were equivalent to one false positive. However, doctors thought that false negatives were worse than false positives. Their median tradeoff was seven false positives to one false negative. This suggests that personal experience may have an enormous impact on how one values the desirability of the outcomes.

These findings are in line with research on error ratios for past crimes, where great variability in responses was found (see Lillquist, 2002, p. 143-146). Furthermore, as Lillquist (2002, p. 144-5) argues, it may also be the case that people's responses would differ strongly if they had to make a decision in a real case, rather than being presented with an abstract, fictional case. However, if more detailed examples are used, it would be unclear how much the results would generalize. To summarize, there appears to be little hope for using empirical research to determine the societal consensus on the required utilities. Such a consensus appears simply not to exist.

4.2. Reflective equilibrium

The upshot of the above is that a decision-theoretic analysis of proof standards for future crimes cannot rely on a pre-existing social consensus regarding the relevant utilities. However, there is an alternative approach. In particular, we can propose a proof standard and show that the assumptions that underly this standard are reasonable.

One idea that can help us arrive at such a reasonable proof standard, resting on reasonable assumptions, is the philosophical *method of reflective equilibrium* (*cf.* Knight, 2023). On this approach, we begin by considering our opinions about some subject matter, then formulate systematic principles that account for these judgments. We then look for discrepancies between these principles and our judgments. If there are any, we adjust both principles and judgments until we reach a state in which they agree with one another[11]. I propose that we can use this approach to

[11] The version of the reflective equilibrium that I have in mind was already hinted at in the results found by Nagel, Lamm and Neef (1981, p. 368) who asked respondents about the desirability of false convictions compared to false acquittals and found that the reported values would sometimes lead to a surprisingly low standard for proof beyond a reasonable doubt. The authors go on to note that some respondents reevaluated their stated utilities after it was pointed out to them that their assigned utilities would result in such lenient proof standards. This reevaluation is in line with the notion of reflective

justify a proof standard for a preventive sanction. Specifically, my version of this approach for determining the standard of proof asks us to answer four questions:

1. What are the costs and benefits associated with the outcomes of the decision to (not) apply a preventive sanction?

2. What are the relative utilities of these outcomes given the results of step 1?

3. What is the resulting proof standard given the results of step 2?

4. Is this proof standard reasonable?

If the answer to the last question is "no"—i.e., if we have arrived at a standard that is unreasonably high or low, we return to steps 1 and 2, and ask whether we have overlooked any costs and benefits and whether we should change our assessment of the relevant utilities. If the answer to the final question is "yes", we end up with a clear justification for the assumptions that underly this standard. Those who wish to criticize this standard can then do so by questioning these assumptions.

4.3. Justifying the standard of proof

Let us apply the reflective equilibrium approach to the fictional preventive sanction and the associated proof standard described in section 3. The first step is to consider the associated costs and benefits associated with each outcome.

As Morris & Miller (1985, p. 23) argue, the two most important considerations when it comes to the utility of preventive sanctions are (i) how serious the interference with liberty involved is when the sanction is applied, and (ii) how serious the injury from the subsequent crime would be, were the sanction not applied to a dangerous person. To these two main considerations, several others can be added. For instance, being preventively detained because you are "dangerous" can have a stigmatizing effect, damage one's career, relationships and general health (Mossman, 1995, p. 111). Stevenson & Mayson (2022, p. 730) also mention the cost of detention as a cost of preventively detaining a person and as a potential cost of not applying such detention the harm to the victim's family and friends if a dangerous person does commit a crime. We can also draw inspiration from Laudan and Saunders (2009, p. 14-21) who list a number of utility-determining factors for true and false convictions and acquittals. Not all of these are applicable to preventive justice. For instance, preventive sanctions are not aimed at giving criminals their just deserts. Yet other factors mentioned by them are more relevant, such as "crime reduction by deterrence" and "crime reduction by incapacitation". Using these remarks as inspiration, I arrive at the following list of costs and benefits of each of the four outcomes (with arrows representing whether something is a cost or benefit and the number of arrows indicating of how great a cost or benefit it is):

equilibrium, as long as these respondents also updated their utility judgments to be coherent with the desired proof standard and if the resulting utility judgments remained reasonable to themselves.

Table 3: The costs and benefits associated with a decision (not) to apply the preventive sanction

	Correctly applied	**Incorrectly applied**
Sanction	*True positive*	*False positive*
	Deprivation of liberty ↓↓	Deprivation of liberty ↓↓
	Stigmatization (damage to career, personal relationships, health etc.) ↓↓	Stigmatization (damage to career, personal relationships, health etc.) ↓↓
	Cost of incarceration ↓	Person unjustly treated as dangerous ↓
	Crime reduction by deterrence ↑	Cost of incarceration ↓
	Crime reduction by incapacitation ↑	Crime reduction by deterrence ↑
	Sense of security community ↑	Sense of security community ↑
	Resocialisation ↑	
Non-sanction	*True negative*	*False negative*
	No costs or benefits, maintains status quo	Harm to victim(s) ↓↓↓
		Harm to family/friends victim(s) ↓
		Damage to sense of security community ↓↓

Many of these costs and benefits can be debated. For instance, is there indeed a distinct injustice in treating someone as dangerous when they are not? If so, should this be part of a utilitarian calculus, or is this strictly a deontological claim? My aim here is not to resolve such debates, but to explicate assumptions that will underly my decision-theoretic analysis in the hope that these will be reasonable to others (and to allow them to be criticized).

The second step is to consider the relative utilities of the four outcomes in the light of the above. A few assumptions seem reasonable. First, I assume that a true negative has a utility of 0, as it has no associated costs or benefits[12]. Second, for both true positives and false positives the associated costs are greater than the associated benefits. Hence, their utility is net negative[13]. Of these two, a false positive is worse (but not much worse) than a true positive. Finally, a false negative, where a dangerous person is not detained and hence commits a crime, is the worst outcome. More precisely, I follow Monahan's (1977, p. 370) suggestion that it is ten times worse than detaining a non-dangerous individual[14].

We can now arrive at numbers expressing the relative utilities of these numbers. Mossman (1995, p. 108-109) suggests setting the best outcome to 1 and the worst outcome to 0 in order to normalize the scale and to permit the easy evaluation of outcomes that fall between these extremes[15]. However, because three of the four outcomes have a net negative utility and because the best outcome has a net utility of 0, I will use a scale of 0 (best) to -1 (worst). Because a false negative is the worst outcome, it gets a value of -1. As just said, I assume that a false positive is ten times worse than a false negative, so the latter gets a value of -0.1. A true positive is better than a false positive. Hence, I assign it a value of -0.05. Finally, a true negative receives a value of 0.

If we enter these values into Tribe's formula mentioned in section 2. The result is as follows:

$$\text{Proof standard} = \frac{1}{1 + \{(-0.05 + 1) \,/\, (0 + 0.1)\}} \approx 0.1$$

In other words, the assumptions just made yield a proof standard of roughly 10%. The preventive sanction that we are using as an example can therefore only be applied to a person if there is a probability of at least 10% that they will commit a violent crime in the future. Yet this seems far too low. It would allow us to detain individuals even if there is only the slightest suggestion that they may soon commit

[12] Vars comes to the same conclusion. However, some might argue that not being labeled "dangerous" removes a stigma for a person and hence has some small amount of positive utility (Vars, 2012, p. 887).

[13] This may strike some as unreasonable for a true positive. Is it not a good thing to prevent a dangerous person from committing a crime? Yes, it certainly is. However, its goodness derives from the fact that it leads to avoiding a false negative and the associated costs mentioned in the table above. We should avoid double-counting such costs and benefits.

[14] One important consideration for preventive detention that was not included in this analysis is that, in practice, preventive detention is typically evaluated regularly. This dampens the potential harmfulness of false positives. After all, it creates the possibility of correcting such errors will be corrected within a relatively short time-period (Slobogin, 2018, p. 402).

[15] This also emphasizes that utilities are tools for comparing outcomes.

another crime. The resulting proof standard therefore fails the fourth step of the reflective equilibrium method.

We now have to return to steps one and two: should the costs and benefits and/or the utilities be changed? A candidate for revision immediately springs to mind. The assumption made above was that a false negative is ten times worse than a false positive (a reverse Blackstone ratio). This assumption was supported by citing Monahan (1977, p. 370). But let us look at what Monahan actually wrote:

> Paraphrasing Blackstone (1962), it might be better that ten "false positives" suffer commitment for three days than one "false negative" go free to kill someone during that period .

However, as Mossman (1995, p. 111-2) argues, "not all acts of violence, are murders. Some assaults frighten victims but cause little physical harm, and their evil approximates the evil of needless hospitalizations more closely than does an act of murder." Additionally, Monahan (1977) speaks of a hospitalization of three days, whereas the preventive sanction described in section 3 allows for detainment for up to two years. Hence, it seems reasonable to alter the ratio between false positives and negatives. Let us assume that a false negative is still worse in this case, but only slightly so. Because a false negative remains the worst option, it still has a value of -1. We can assign a value of -0.7 to a false positive to express this assumption that it is slightly less bad than the false negative. Finally, we shall retain the assumption that a true positive is half as bad as a false positive and that a true negative is a neutral outcome[16]. If we put these values into the equation, we get a proof standard of around 50%, which seems more reasonable[17]. In other words, we have arrived at a reflective equilibrium. While not all will find this standard reasonable, those who disagree with it can challenge the assumptions that underly it, namely that (i) a false negative is the worst possible outcome; (ii) a false positive is slightly less bad than a false negative, (iii) a true positive is less bad than a false positive, but still a negative outcome, (iv) a true negative is a neutral outcome.

4.4. Can we ignore correct outcomes?

As I explained in section 2, there are good reasons to include the utilities of correct outcomes in our calculation of the proof standard. Indeed, as we saw just now, true positives have various costs and benefits associated with them. Yet as I mentioned, the downside of including all four outcomes is that it makes calculating the proof standard significantly more complex. For this last reason, many authors make the simplifying assumption that the utilities of the correct outcomes are equal, and hence, cancel one another out (Lempert, 1976). This would mean that Tribe's formula can be reduced to Kaplan's, on which we simply compare the (dis)utilities of errors. Scurich (2016, p. 173-174) suggests that we can also make this assumption in

[16] Therefore, I assign these values of -0.35 and 0 respectively.

[17] $1 / \{1 + [(-0.35 + 1) / (0 + 0.7)] \} \approx 0.52$.

the context of preventive sanctions. However, he does not explore whether and why such an assumption might be reasonable. In fact, given the discussion above, one could argue that the two correct outcomes should not be regarded as equal. After all, we saw that true negatives have no associated benefits or costs, whereas true positives do have all kinds of costs and benefits. Nonetheless, I believe that an argument can be made for this simplification in the context of preventive sanctions.

To the extent that authors have tried to justify this assumption, they use one of two methods, which I call the "regret-based approach" and the "reductive approach". The regret-based approach is used by Lempert (1976, footnote 41). It is tied to the idea that the utilities one assigns in a decision-theoretic analysis can be viewed as the regret one may feel (or should feel) after making a decision. For example, a fact-finder might feel a certain amount of regret if they found out that they wrongfully applied a sanction to a non-dangerous person. This regret may be less than the regret they would feel if they wrongfully did not apply a sanction to a dangerous person. According to Lempert (1976), an ideal fact-finder should not feel regret after making a correct decision. Hence we can set the utility of both types of correct decisions to 0, thereby canceling them out.

Stevenson and Mayson (2022, p. 764) offer an argument against the regret-based approach. They note that ignoring the costs of correct decisions makes sense for adjudications of guilt, "where it is permissible to discount the harm inflicted on a person who is accurately convicted and punished because, at least in theory, that harm is deserved." However, this finding of guilt is absent in the preventive context. Therefore, Stevenson and Mayson (2022, p. 765) propose, even correct preventive decisions are costly because they "subordinate the welfare of the detained person to the public good."

The reductive approach is described (and rejected) by Laudan & Saunders (2009, p. 14-15). On this approach we do take the utilities of correct outcomes into account, but we do so by assigning negative versions of them to errors. For example, one benefit of a true conviction is that it gives a guilty person their just desert. We could say that a negative consequence of a false acquittal therefore is that it fails to give the party their just desert. If we strip away all costs and benefits of correct outcomes, we can set their utility to 0. However, as Laudan & Saunders (2009, p. 14-15) point out, this strategy does not work for proof of past crimes. They write:

> Consider the assignment of "a failure to deliver just deserts" to a false acquittal. But it is not only false acquittals that fail to give just deserts to the guilty. So do true acquittals and false convictions. We could readily add to the liabilities of a false acquittal that it too fails to give just deserts. Still, how do we capture [the fact that] a true acquittal fails to deliver an important benefit associated with a true conviction? The only reasonable way to do that is to reduce the utility of a true acquittal relative to a true conviction (or, alternatively, raise the utility of the latter). But either modification would undermine the whole enterprise, since the object of this maneuver was to render true convictions and true acquittals as neutered and thus dispensable, each possessing a utility of zero.

So, whereas the regret-based approach allows us to simply ignore all costs and benefits associated with correct outcomes, on the reduction account we must assign these costs and benefits to incorrect outcomes. In the case of convictions and acquittals, this account fails according to Laudan and Saunders (2009). But let us consider the aforementioned costs and benefits associated with preventive detainment. We saw that true negatives have no costs and benefits. Hence we can focus on the costs and benefits of true positives. Can we assign the converse of these to false negatives? I believe that we can. See the table below (converted costs and benefits in italics):

Table 4: the costs and benefits associated errors

	Incorrectly applied
Sanction	*False positive*
	Deprivation of liberty ↓ ↓
	Stigmatization ↓ ↓
	Person unjustly treated as dangerous ↓
	Cost of incarceration ↓
	Crime reduction by deterrence ↑
	Sense of security community ↑
Non-sanction	*False negative*
	Harm to victim(s) ↓ ↓ ↓
	Harm to family/friends victim(s) ↓
	Damage to sense of security community ↓
	No deprivation of liberty ↑ ↑
	No stigmatization ↑ ↑
	No cost of incarceration ↑↑
	No crime reduction by deterrence ↓
	No crime reduction by incapacitation ↓
	No resocialisation ↓

None of these conversions of costs or benefits appear to be problematic. So, it seems that we can reduce Tribe's formula to that of Kaplan:

$$\text{Proof standard} = \frac{1}{1 + (U_{\text{false negative}} / U_{\text{false positive}})}$$

To justify the 50% rule, we would have to conclude that a false negative is roughly equally bad as a false positive. This may seem surprising. Had we not said that a false negative is worse? However, note that we also assigned the converse of all attributes of a true positive to a false negative. Because I assumed that a true positive had a negative utility, we now add positive utility to the false negative. And because we are simply moving the utilities from one part of the equation to the other, the result is the same.

One downside of this approach is that it makes determining the costs and benefits of the errors complex, because we have to convert each value to its negative. A benefit of the regret-based approach is that it is simpler in this regard. Rather than retaining the costs and benefits of correct outcomes in a different form, it ignores them[18]. As Stevenson and Mayson (2022) argue, however, this may count against this approach.

5. THE VALUE OF AN IMPRECISE STANDARD

The previous section showed how one can justify the proposed "likely" standard, which is intended to be a percentage around 50%. This justification was based on considering several costs and benefits associated with the different outcomes. However, these costs and benefits can differ between individual cases. For example, the type of violent crime that we expect to occur in the case of a false negative can differ in severity. Should the standard for the preventive sanctions we are considering therefore differ between cases too? This idea, that the proof standard should be flexible to account for the utilities in individual cases, has been argued for by several authors (Stoffelmayr & Diamond 2000; Lillquist, 2002; Vars, 2010). These authors claim that a quantified standard (i.e., one that states an explicit probability such as "50%") would not be flexible enough. Instead, they propose vaguer terms, of which "likely" is an example[19]. This section examines this suggestion in relation to the fictional preventive sanction central to this article. It will be argued that formulating standards of proof requires striking a balance between flexibility and normative guidance.

5.1. The decision-theoretic argument for flexibility

Vars (2010, p. 21) summarizes the point well, when he writes that the strongest argument against a quantified proof standard is that it is insufficiently flexible, as it "prohibits the trier of fact from balancing the costs of false positives and false negatives in a particular case and from adjusting the standard of proof accordingly." He therefore suggests using verbal rather than numerical standards. Similarly, Stoffelmayr & Diamond (2000, p. 783) write that a "single uniform standard across

[18] However, note that this means that the two approaches lead to different proof standards. For example, earlier, I wrote that I assigned a false positive a value of -0.7 and a false negative a value of -1. If we enter these values into Kaplan's formula, we get a proof standard of around 41%, rather than of 50%.

[19] Janus & Meehl (1997, p. 40) show that several US courts use the "likely" standard for civil commitments, but are typically unclear about what probability value they intend to convey with this.

cases is not an optimal resolution when the decisions to which the standard is being applied carry different costs." The most extensive treatment of the argument is given by Lillquist (2002) who proposes that the notorious vagueness of the beyond a reasonable doubt standard is desirable, as it allows fact-finders to use a flexible standard – tailored to the utilities of the individual case. Both Lillquist (2002, p. 88) and Stoffelmayr & Diamond (2000, p. 769) cite empirical research suggesting that, in practice, fact-finders do engage in such tailoring of the proof standard to the crime and to the punishment.

Lillquist (2002, p. 162-76) discusses the benefits and risks of having a flexible rather than a fixed standard for past crimes. The biggest problem with a flexible standard is that fact-finders may not accurately weigh the utilities in a given case and therefore apply a different standard than what is desirable. As Lillquist (2002, p.166 writes:

> It may be simpler for the juror to apply a rule that says "find the defendant guilty if you are ninety percent certain" than to apply a standard that, at best, implicitly invites the juror to decide what the appropriate standard should be[20].

He proposes two questions that we should ask when deciding whether variable or fixed standard of proof would be desirable: (i) is the preferred standard of proof likely to vary widely in different cases? (ii) are fact-finders likely to apply a standard of proof in a way that varies widely from the one that we would prefer in a given case? (Lillquist, 2002, p. 168). When we answer "yes" to the first question and "no" to the second, then a variable standard is preferable. If the answer to the first question is "no" and to the second "yes", we should prefer a fixed standard. He comes to the conclusion that a variable standard is preferable, in part on the ground that there can be very much variation in the utilities involved in individual criminal cases. Lillquist (2002, p. 170), points out that many legal systems have criminalized very minor crimes, such as manufacturing burglary tools, which are much less serious than, for example, murder and rape (and that there are many crimes in between these extremes, such as tax fraud). Similarly, the punishments can vary between a fine and life in prison (or even the death penalty in some legal systems) (Lillquist, 2002, p. 149-50). Therefore we would want a proof standard that can vary widely too.

What about proof standards for future crimes? As we saw earlier, people tend to have very different opinions on the utilities involved in preventive detention. On the other hand, while the beyond a reasonable doubt standard spans a wide range of cases and potential punishments, dangerousness criteria are often tied to a specific type of preventive sanction. The fictional sanction from section 3 is an example of

[20] Though applying a numerical threshold may not necessarily be easier for fact-finders. After all, judges and jurors have notorious difficulty with interpreting numerical information and find it difficult to critically question these numbers. This is also another reason why quantified standards may conflict with the fact-finder's task to consider the specifics of the case. As Tribe (1971, p. 1376) puts it, fact-finders may be "induced by the persuasive force of formulas and the precision of decimal points to perceive themselves as performing a largely mechanical and automatic role."

this, as it is connected to the treatment facility for repeat violent offenders. The range of cases covered by this standard should differ much less than those that fall under the beyond a reasonable doubt standard for past crimes. Nonetheless, as Mossman (1995, p. 110-1) writes:

> Not all acts of violence, however, are murders. Some assaults frighten victims but cause little physical harm, and their evil approximates the evil of needless hospitalizations more closely than does an act of murder. Moreover, involuntary hospitalization of persons who, if left alone, would have been harmless is more than a mere annoyance. It can be stigmatizing and damaging to an individual's career, marriage, and nervous system.

How much can the proof standard for our example differ if we adopt a variable standard of proof? I will consider two examples, that I take to be the most extreme cases. For the sake of simplicity, I will adopt the regret-based account discussed in the previous section and only consider the relative utilities of errors. The first example is based on Monahan's (1977, p. 370) suggestion that 'it might be better that ten "false positives" suffer commitment for three days than one "false negative" go free to kill someone during that period.' In other words, a relatively short period of commitment that will help prevent a murder. Let us therefore use Monahan's (1977) suggestion of a ratio of one false positive to ten false negatives as one of the extremes. An example from the opposite end of the spectrum would be inspired by Mossman's (1995) remark above, namely a long commitment of two years to prevent recidivism in the form of a violent assault that leaves the victim mostly unscathed. In the case of a false positive, a non-dangerous person is committed for two years. A false negative would mean that the aforementioned assault occurs. It does not seem unreasonable to reverse the ratio in this case, to one false positive for every ten false negatives.

When we put these values into Kaplan's formula, we get proof standards of 9% and 91% respectively. Of course one could argue that cases where a person who is likely to commit a murder only receives treatment for a brief moment are rare, as are cases in which someone who is not particularly dangerous will be committed for the maximum amount of time allowed. Nonetheless, even if we relax these assumptions, the point holds. For example, if we assume a 2:1 and 1:2 ratio the proof standard would vary between 33% and 66% respectively. So, it seems that we should answer Lillquist's first question with a resounding "yes".

As for his second question, Lillquist (2002, p. 174) admits that it is difficult to determine how often fact-finders would wrongly weigh the relative utilities. This is made all the more difficult because these utilities are dependent on the particulars of the case. Nonetheless, Lillquist (2002, p. 175) suggests that there is cause for optimism with respect to that question, in part because the reasonable doubt standard (with which he is concerned) has been vague for so long, and the legal system appears to be content with this vagueness. Yet, what about preventive sanctions? It is for this reason that I now turn to what I consider to be the strongest argument against a flexible standard.

5.2. The need for a precise standard

The argument against a vague, flexible standard is that it may leave fact-finders *too* free in their judgments. For example, as Stoffelmayr & Diamond (2000) argue, one criterium for an adequate standard of proof is that it should be precise enough to be distinguishable from other standards and that it should create at least some consistency (e.g., between jurors or judges in the same case). Yet as Kagehiro (1990, p. 194-197) found, jury instructions such as "very likely" and "extremely likely" all elicited essentially the same verdicts. In a similar vein, Morse (1982, p. 72) argues that standards for civil commitment are too vague and have no generally agreed upon meaning. He writes that each person therefore "injects his or her own private meaning into the criteria, rendering the system essentially lawless."[21] This point of lawlessness is also relevant in the criminal legal context. A well-known problem with preventive criminal justice compared to traditional criminal law, namely the lack of clear procedural rules and case-law that governs the process of proof (see e.g., Ashworth, Zedner & Tomlin, 2013; Bijlsma, 2024).

While variance in the standard of proof can sometimes be desirable from a decision-theoretic perspective, too much variance can lead to an unequal treatment of equal cases and a lack of predictability, which clashes with the requirement of legal certainty. Furthermore, as Tillers & Godfried (2006, p. 155) put it, triers of fact should not be permitted "to strike a balance that is wildly at variance with the values of society at large."[22] As we saw earlier, it is doubtful whether there truly exists a (measurable) societal consensus on the relevant utilities. Nonetheless, it is easy to imagine utility assignments that are deemed unreasonable by society (e.g., standards that are far too low or far too high). Finally, a precise standard would also make it easier to determine whether the fact-finder correctly applied the standard in individual cases.

According to Tillers & Godfried (2006, p. 155) it is up to lawmakers and appellate judges to select a standard of proof that best accommodates the competing interests[23]. They suggest that this should be a precise, quantitative standard. Yet there are also other options for the lawmaker or appellate court, which may strike a better balance between precision and flexibility. For instance, Stoffelmayr & Diamond (2000, p. 782) propose that "an alternative to [a] single probability standard (…)

[21] This claim is supported by the study of Monahan & Silver (2003) who asked judges how they would set proof standard for short-term involuntary civil commitment. They found values between 1% and 56%. See also (Slobogin, 2021, p. 48).

[22] Similarly, Scurich & John (2010, p. 446) argue that "any departure from the normative social cost policy should be justified by more than an appeal to judicial discretion" because otherwise it may be too easy to apply preventive sanctions, leading to an increase in the number of false positives.

[23] Similarly, Redmayne states that the weighing of utilities is a policy decision "better made by the legislature than by judges" (Redmayne, 1999, p. 183)

might be to provide a range (e.g., . 87 to .92) that [fact-finders] would be invited to apply according to their assessments of the costs of error associated with a particular offense." One objection to this proposal would be that, as we have seen, the proof standard for future crimes can vary wildly. An instruction that asks fact-finders to put the standard somewhere between "9% and 91%" (or even between 33% and 67%) is hardly helpful. Another approach might be to create more fine-grained standards, that limit the variance in cases (Stoffelmayr & Diamond, 2000, p. 782). For example, one could create multiple proof standards for the treatment facility for repeat violent offenders discussed in section 3, based on the utilities involved.

A further possibility suggested by Stoffelmayr & Diamond (2000, p. 783) is a non-quantitative standard that instructs the fact-finder to balance different costs and benefits. Lillquist (2002) argues that finding the right wording for this is extremely difficult. However, as Vars (2012, p. 893) points out, there are some courts that have formulated these types of jury instructions for dangerousness criteria. For example, one such instruction reads:

> In assessing the risk of reoffending, it is for the fact finder to determine what is "likely." Such a determination must be made on a case-by-case basis, by analyzing a number of factors, including the seriousness of the threatened harm, the relative certainty of the anticipated harm, and the possibility of successful intervention to prevent that harm.

This kind of instruction is aimed at jurors in common law systems. Yet in civil systems, where it is usually up to the judge to make such determinations, similar instructions could, for instance, be (and sometimes are) given by the lawmaker. Such an approach could also help avoid another potential problem with vague standards. We want to avoid situations in which the fact-finder includes considerations in their utility calculus that are socially undesirable (e.g., where they adopt a lower proof standard for certain social groups than for others). If the instructions were to include the factors that the fact-finder should weigh, this gives them guidance on how to go about determining the relative utilities in an acceptable way, without limiting their freedom to tailor the proof standard to the case at hand.

To summarize the above, the proof standard should be formulated in a way that balances the need for flexibility against the need for clear normative guidance to the fact-finder. The "likely" standard proposed here may allow too much flexibility. One could solve this by formulating several fine-grained quantitative standards, or by offering qualitative instructions on how the fact-finder should weigh the relevant utilities.

Does a proof standard that contains a precise probability requirement provide sufficient normative guidance? According to Slobogin (2021, ch. 2), it does not. He argues that policymakers should not only offer clarity about the probability of a future crime sufficient to justify preventive intervention, but also regarding the nature and time-frame of the predicted crime (Slobogin, 2021, p. 49-56). In addition, he opts for a subsidiarity requirement according to which a far-reaching intervention such as preventive detention is only permissible if there are no less intrusive measures available to reduce the risk of a future crime to acceptable levels (Slobogin, 2021, p. 52-6). The proof standard used as an example in this article ("it must be likely that the person will commit serious harm to others") does not contain such details. Adding them would provide additional clarity for the fact-finder. However, note that the discussion above, regarding the balance between flexibility and normative guidance also holds for such additional criteria, where more normative guidance leads to less flexibility.

6. PROOF STANDARDS AND RISK ASSESSMENTS

On the decision-theoretic approach, the proof standard is conceptualized as a probabilistic threshold. In other words, to apply a preventive sanction, the probability of recidivism must be sufficiently high. The most important evidence for determining the probability of a future crime is often the testimony of an expert (typically a forensic psychiatrist) who uses a structured risk assessment instrument for determining the likelihood that the person under consideration will recidivate (Eaglin, 2017). Such instruments are increasingly "actuarial" (Vars, 2010, p. 873-4; Eaglin, 2017). This means that the expert uses statistical algorithms to combine and weigh various risk variables—such as age, gender and past violent behavior—and to combine these into a single probability judgment. The result of such an actuarial risk assessment may, for instance, be that a defendant is categorized as a "high risk individual", which might indicate a risk of violent recidivism of 60% within the next two years.

This section discusses two issues regarding the link between the expert's risk assessment and the fact-finder's decision whether the proof standard is met. The first is to what extent it can ever be up to the expert to weigh the utilities in a given case. The second is a suggestion made by some authors and adopted by some legal systems, that aside from a proof standard for the probability of a future crime, there should also be a proof standard for our confidence in this probability.

6.1. Utilitarian assumptions in risk assessment technologies

As we saw so far, on the decision-theoretic approach, we determine the level of the standard of proof by considering the utilities at stake. Furthermore, these utilities may differ between cases and hence so may the standard of proof. So far, I have assumed that the weighing of utilities is strictly up to the lawmaker or the appellate court and the fact-finder (i.e., the judge or jury, depending on the legal system). This fits with the literature on proof standards for past crimes. An expert is not supposed to offer verdicts on how to interpret the standard of proof or on whether a defendant's guilt is proven beyond a reasonable doubt. However, this is arguably different for dangerousness assessments. For those types of decisions, it is not unheard of for the expert (or for the developers of the risk assessment technologies that the expert uses) to make normative, utilitarian assumptions. To give an example, Min & Ferris (2020, p. 14) discuss the Harm Assessment Risk Tool (HART). They write that this risk assessment technology is "calibrated to err on the side of caution, because it regards under-estimations of risk levels as a more serious error than over-estimations." In other words, the way in which the risk levels are reported is adjusted based on a normative belief about the required utilitarian tradeoff. As Eaglin (2017) argues, developers of risk-assessment tools make numerous of such normative judgments about when persons should be marked as "dangerous" or "high risk", including how much risk we should tolerate as a society and what tradeoff between false negatives and positives is desirable.

One reason why the fact-finder may not be aware of these normative choices made by experts and by the designers of risk assessment technologies is that risk assessments are often communicated to the fact-finder in categorical terms, such "low/medium/high risk", rather than in terms of probabilities. Scurich (2018) argues against this practice, suggesting that it obscures what is fundamentally a value judgment about the relative costs and benefits of correct and incorrect outcomes[24]. After all, whether a probability of recidivism of, say, 50% counts as high risk (which would, presumable justify applying a preventive sanction) depends on how one balances the utilities of the various outcomes. However, when risk assessments are communicated in categorical terms, the court may assume that the expert is using a different proof standard (based on different assumptions about the utilities involved) than they actually are (Bijlsma & Meynen, 2023, p. 273). This is especially problematic because courts typically follow the expert's judgments when it comes to whether an individual is "dangerous". When an individual is marked as "high risk", courts almost always opt for preventive interventions (Slobogin, 2021, p. 48).

A potential argument against leaving the balancing of utilities and the determination of the proof standard strictly up to the fact-finder is that a forensic expert may have what philosophers call "epistemic authority" (Zagzebski, 2012). In other words, the expert may be in a better position than the judge or jurors to assess what the consequences of applying the sanction to the defendant would be, as well as determining the consequences of recidivism. For example, they may have had in-depth conversations with the defendant when assessing how dangerous this person is and they may be able to draw on their own extensive experience with similar cases. Furthermore, it is likely that experts are better at interpreting the statistical evidence generated by actuarial risk assessment technologies. After all, many people find it difficult to interpret and draw correct inferences when reasoning about probabilities

[24] An even more neutral method of reporting risk assessments is offered by Slobogin. He proposes that experts should not report that a person has a specific probability of reoffending. Instead they should report only "that the offender received a risk score that is consistent with the scores of a group studied in previous research, X percent of which offended" (Slobogin, 2021, p. 45).

(Kahneman, 2011). Judges and jurors can also succumb to various kinds of errors in drawing conclusions for probabilistic evidence (Dahlman, 2024). Whether and to what extent the expert should make decision-theoretic determinations will depend on how we weigh their expertise on these matters against the lawmaker's democratic and the fact-finder's legal authority.

6.2. Proving a probability beyond a reasonable doubt

Some authors propose that, aside from a standard for the minimum probability of recidivism, there should also be a proof standard for the quality of our evidence for that probability. This idea was first discussed by Monahan & Wexler (1978). They make a distinction between two levels of proof standards. First there are those standards that measure the probability of future crimes, and that are expressed in terms such as "likely" or "highly likely" (I will call these "first-order standards of proof"). Second, there are standards that express how confident the fact-finder is about the evidence for this level of probability (which I will call "higher-order standards"). According to Monahan and Wexler (1978), such higher-order standards are expressed using terms like "clear and convincing evidence" and "beyond a reasonable doubt." They offer several real-life examples of proof standards for civil commitment requiring that a court find "beyond a reasonable doubt" that an individual is "likely" (or, in another example "more likely than not") to "do substantial harm to another" (Monahan & Wexler, 1978, p. 40)[25].

What would it mean to prove "beyond a reasonable doubt" that it is "likely" that the defendant will commit a violent crime in the future? There are at least two answers to this question. The first is offered by Monahan & Wexler (1978, p. 39) themselves. They note that risk assessment methods place an individual in a statistical group that has an associated risk of dangerous behavior. Suppose that membership of this group is based on having a certain age, sex, a psychiatric diagnosis, past criminal behavior and an addiction. We can verify all these facts about the person with a great degree of certainty, e.g., by checking their police records, bringing in witnesses and so on. In this example, proving a risk beyond a reasonable doubt means adducing solid evidence that the defendant falls into a given risk-category.

The second suggestion on what it means to prove a risk of recidivism beyond a reasonable doubt is explored by Vars (2012). He notes that estimates of risk themselves come with error – expressed in terms of confidence-intervals (Vars, 2012, p. 873-4; Van Koppen, 2008). Vars proposes that we can interpret higher-order proof standards in terms of being confident to a certain degree that the true probability of recidivism is *at least* a certain level. He gives the example of proving beyond a reasonable doubt – i.e., with a confidence of at least 90% – that there is a probabil-

[25] See Scurich & John (2010, p. 447-8) for further examples.

ity of at least 50% that an individual will recidivate (Vars, 2012, p. 876). According to Vars (2012), this 90% confidence can be achieved by requiring a higher predicted level of recidivism risk. The more the predicted level of recidivism risk exceeds 50%, the greater our confidence that the "true" recidivism risk lies above this point. How high the predicted level of risk ought to be to achieve this level of confidence depends on how accurate our risk-assessment methods are[26]. For instance, Vars (2012) discusses data concerning a real-life risk-assessment method. He concludes that for this method, 90% confidence that there is over 50% chance of recidivism for a given individual is achieved when the risk-assessment instrument predicts a risk of recidivism of 65% or higher (Vars, 2012, p. 876)[27].

Why would we want to have a two-tiered proof standard for preventive sanctions? One possibility is that such a standard reflects the idea that a preventive sanction such as the one used as an example here, can have far-reaching effects and that the decision to apply this sanction should not be taken lightly[28]. Even if our decision-theoretic analysis leads to a relatively low first-order standard, we may nonetheless want to be as confident as possible in our probabilistic assessment. This idea connects to a similar debate surrounding proof standards for past crimes: even if convicting a defendant maximizes expected utility, such a conviction may not always be warranted. In particular, many people have the intuition that we cannot convict someone based only on statistical evidence, even if this evidence makes it highly probable that they committed the alleged crime (*cf.* Dahlman, 2020; Günther, 2024a; 2024b). Some authors suggest that a conviction for a past crime is therefore only justified if there is a high probability of guilt and this probability is "stable"; (or "robust", "resilient" or "safe") (*cf.* Ho, 2008, 278; Stein, 2005, 88; Di Bello, 2013; 2015; Urbaniak, 2018; Günther 2024a; 2024b; Jellema, 2024). These authors have provided accounts of what it means for a probability to be "stable", why convictions based purely on statistical evidence are not stable and why convicting based on an unstable but high probability is undesirable. For future crimes, the statistical evidence (the output of a risk-assessment instrument) is typically the most important (and sometimes even the only) evidence based on which a preventive measure is applied. One unexplored question is whether the arguments from the debate on naked statistical evidence for past crimes also give reasons to desire "stable" probabilities for future crimes[29].

[26] I.e., it depends on how tight the associated confidence intervals are.

[27] In contrast Slobogin proposes that proving a recidivism risk of 51% to 90% certainty means that "the state would need to show that roughly 45 percent (51 percent x 90 percent) of people in the offender's risk category will commit 'a violent act constituting a threat to society'" (lobogin, 2021, p. 47).

[28] This idea might, for instance, underly the following remark by a court: "society has a substantial interest in the protection of its members from dangerous deviant sexual behavior. But when the stakes are so great for the individual facing commitment, proof of sexual dangerousness must be sufficient to produce the highest recognized degree of certitude" (Vars 2012, p. 868).

[29] A related question is whether (and when) proof of future crimes is based purely on statistical evidence. For this we'd need an account of the difference between statistical and individualized evidence. One such account is offered by Günther (2024a).

A potential, argument *against* a two-tiered standard of proof is offered by Slobogin (2006, p.144). As he puts it, the two-tiered standard of proof may be a "sleight of hand" because it "purports to require a high level of proof that the person will offend,' when in fact the conception really only considers the confidence in the prediction, rather than the risk". In other words, the suggestion of a two-tiered standard is nothing more than a method of masking a low standard of proof as high one. Another counterargument to such standards is offered by Laudan (2006, p. 119) regarding proof standards for past crimes. He proposes that the criminal legal system has a "pro-defendant bias". What he means by this is that various legal rules, such as those regarding the admissibility of evidence, are designed to benefit the defendant. According to Laudan (2006, p. 119-44) the intuition that underlies such pro-defendant rules may be mistaken. He argues that such rules increase the number of false acquittals above the desired level enshrined in the proof standard, thereby ignoring the negative utilities associated with that kind of error. A suggestion such as that by Vars (2012) may be subject to a similar critique. After all, on this account we assume that uncertainty about the probability that someone will commit a crime should benefit the defendant[30]. This could similarly be argued to be an example of the kind of "pro-defendant bias".

7. CONCLUSION

This article discussed how we may use decision theory to clarify standards of proof for future crimes. On this approach we weigh the utilities of the possible outcomes of the decision (not) to apply a preventive sanction. This utilitarian calculus yields a probabilistic threshold that has to be met before the preventive sanction can be applied. The principal aim of this article was to show that such a decision-theoretic analysis requires wrestling with a number of difficult questions and to explore some possible answers to them.

One of these questions was how we ought to determine the relevant utilities. This article rejected the suggestion that we can determine them by means of empirical studies. Instead, a "reflexive equilibrium approach" was proposed on which we use an iterative process to arrive at a reasonable standard of proof, based on clearly explicated assumptions. Another question that this article dealt with was whether our utilitarian calculus should include the utilities of correct outcomes. I tentatively suggested that this need not necessarily be the case and outlined two philosophical accounts that can support this conclusion (the "regret-based approach" and the "reductive approach").

[30] Another way of making this point is by noting that such so-called "higher-order uncertainty" about our "first-order probabilities" gives us just as much reason to revise our probabilities upwards as to revise them downwards (*cf.* Henderson, 2022).

A third question was how specific the proof standard for future crimes ought to be. Can it best be expressed in vague but flexible, qualitative terms or in terms of precise but inflexible probabilities? This article argued in favor of the former, but did note that the desired flexibility must be weighed against the normative guidance that this standard offers.

The final part of this article dealt with two questions regarding the relationship between predictions of dangerousness by experts and the fact-finder's determination whether the proof standard is met. First, to what extent can (or should) normative, utilitarian judgments be made by the expert (as often happens in practice)? Second, aside from a standard for the required probability of future crimes, should there also be a standard for the quality of the evidence on which that probability is based?

The aforementioned questions are part of the overarching issue where the limits of preventive justice lie. The aim of preventing future crime must always be weighed against the importance of individual liberty. Decision-theory is a method of formalizing and thereby clarifying this balancing act.

REFERENCES

Ashworth, A., Zedner, L., & Tomlin, P. (Eds.). (2013). *Prevention and the Limits of the Criminal Law*. Oxford University Press.

Bijlsma, J. (2024). Risicostrafrecht en rechtsorde: Kent noodzaak geen recht?. *University of Groningen.*

Bijlsma, J., & Meynen, G. (2023). Aannemelijkheid van strafuitsluitingsgronden en sanctievoorwaarden. Welke mate van onzekerheid is aanvaardbaar?. *Rechtsgeleerd Magazijn Themis*, *2023*(5), 268-276.

Blackstone, W. (1962). *Commentaries on the Laws of England* (Vol. 2). Beacon

Carvalho, H. (2017). *The preventive turn in criminal law*. Oxford University Press.

Cullison, A. D. (1969). Probability analysis of judicial fact-finding: A preliminary outline of the subjective approach. *Toledo Law Review,* 1: 538–598.

Dahlman, C. (2020). Naked statistical evidence and incentives for lawful conduct. *The International Journal of Evidence & Proof*, *24*(2), 162-179.

— (2024). A systematic account of probabilistic fallacies in legal fact-finding. *The International Journal of Evidence & Proof*, *28*(1), 45-64.

DeKay, M. L. (1996). The difference between Blackstone-like error ratios and probabilistic standards of proof. *Law & Social Inquiry*, 21(1), 95-132.

Di Bello, M. (2013). *Statistics and probability in criminal trials*. Stanford University.

Eaglin, J. M. (2017). Constructing recidivism risk. *Emory LJ*, *67*, 60-122.

Günther, M. (2024a). Legal proof should be justified belief of guilt. *Legal Theory*, *30*(3), 129-141.

Günther, M. (2024b). Probability of Guilt. *Canadian Journal of Philosophy*, *54*(3), 189-206.

Henderson, L. (2022). Higher-order evidence and losing one's conviction. *Noûs*, 56(3), 513-529.

Ho, H. L. (2008). *A philosophy of evidence law: Justice in the search for truth*. Oxford University Press on Demand.

Janus, E. S., & Meehl, P. E. (1997). Assessing the legal standard for predictions of dangerousness in sex offender commitment proceedings. *Psychology, Public Policy, and Law*, *3*(1), 33-64.

Jellema, H. (2023). (Im) probable stories: combining Bayesian and explanation-based accounts of rational criminal proof. *University of Groningen* [PhD thesis]

— (2024). Reasonable doubt, robust evidential probability and the unknown. *Criminal Law and Philosophy*, *18*(2), 451-470.

Kagehiro, D. K. (1990). Defining the standard of proof in jury instructions. *Psychological Science*, *1*(3), 194-200.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kaplan, J. (1967). Decision theory and the factfinding process. *Stan L. Rev.*, *20*, 1065-1092.

Knight, C. (2023). Reflective Equilibrium, In: Zalta, E.N. & Nodelman, U. (eds.). The Stanford Encyclopedia of Philosophy. plato.stanford.edu/archives/spr2025/entries/reflective-equilibrium

Laudan & Saunders (2009). Re-thinking the criminal standard of proof: Seeking consensus about the utilities of trial outcomes. *International Commentary on Evidence*, 7(2), article 1 (online journal).

Laudan, L. (2006). *Truth, error, and criminal law: an essay in legal epistemology*. Cambridge University Press.

— (2015). Why Asymmetric Rules of Procedure Make It Impossible to Calculate a Rationally Warranted Standard of Proof. *Available at SSRN 2584658*.

Lempert, R. O. (1976). Modeling relevance. *Mich. L. Rev.*, *75*, 1021-1057.

Lillquist, E. (2002). Recasting reasonable doubt: Decision theory and the virtues of variability. *UC Davis L. Rev.*, *36*, 85-197.

Min, B., & Ferris, G. (2020). Regulating Artificial Intelligence for Use in Criminal Justice Systems in the EU. *Fair Trials,* 2-35.

Monahan, J. (1977). Strategies for an empirical analysis of the prediction of violence in emergency civil commitment. *Law and Human Behavior*, *1*(4), 363-371.

Monahan, J., & Silver, E. (2003). Judicial decision thresholds for violence risk management. *International Journal of Forensic Mental Health*, *2*(1), 1-6.

Monahan, J., & Wexler, D. B. (1978). A definite maybe: Proof and probability in civil commitment. *Law and Human Behavior*, *2*(1), 37-42.

Morris, N., & Miller, M. (1985). Predictions of dangerousness. *Crime and Justice*, *6*, 1-50.

Morse, S. J. (1982). A preference for liberty: The case against involuntary commitment of the mentally disordered. *Calif. L. Rev.*, *70*, 54-106.

Mossman, D. (1995). Dangerousness decisions: An essay on the mathematics of clinical violence prediction and involuntary hospitalization. *U. Chi. L. Sch. Roundtable*, *2*, 95-138.

— (2006). Critique of pure risk assessment or, Kant meets Tarasoff. *U. Cin. L. Rev.*, *75*, 523-610.

Mossman, D., & Hart, K. J. (1993). How bad is civil commitment? A study of attitudes toward violence and involuntary hospitalization. *The Bulletin of the American Academy of Psychiatry and the Law*, *21*(2), 181-194.

Nagel, S., Lamm, D., & Neef, M. (1981). Decision theory and juror decision-making. *The trial process*, 353-386.

Nagel, S., Neef, M., & Schramm, S. S. (1977). Decision Theory and the Pre-Trial Release Decision in Criminal Cases. *University of Miami Law Review*, *31*(5), 1433-1491.

Nance, D. A. (2016). *The burdens of proof.* Cambridge University Press.

Redmayne, M. (1999). Standards of proof in civil litigation. *Mod. L. Rev.*, *62*, 167-195.

Schopp, R. F., & Quattrocchi, M. R. (1995). Predicting the present: Expert testimony and civil commitment. *Behavioral Sciences & the Law*, *13*(2), 159-181.

Schopp, R. F. (1996). Communicating risk assessments: Accuracy, efficacy, and responsibility. *American Psychologist,* 51, 939-944.

Scurich, N. (2015). Criminal justice policy preferences: Blackstone ratios and the veil of ignorance. *Stanford Law & Policy Review*, *26*, 23-35.

— (2016). Structured risk assessment and legal decision-making. *Advances in Psychology and Law: Volume 1*, 159-183.

Scurich, N. (2018). The case against categorical risk estimates. *Behavioral Sciences & the Law, 36(5), 554-564.*

Scurich, N., & John, R. (2010). The normative threshold for psychiatric civil commitment. *Jurimetrics*, 425-452.

— (2011). Constraints on restraints: A signal detection analysis of the use of mechanical restraints on adult psychiatric inpatients. *S. Cal. Rev. L. & Soc. Just.*, *21*, 75-107.

Scurich, N., & John, R. S. (2012). Prescriptive approaches to communicating the risk of violence in actuarial risk assessment. *Psychology, Public Policy, and Law*, *18*(1), 50-78.

Slobogin, C. (1989). The ultimate issue issue. *Behav. Sci. & L.*, *7*, 259.

— (2006). *Minding justice: Laws that deprive people with mental disability of life and liberty*. Harvard University Press.

— (2018). Preventive justice: A paradigm in need of testing. *Behavioral sciences & the law*, *36*(4), 391-410.

— (2021). *Just algorithms: Using science to reduce incarceration and inform a jurisprudence of risk*. Cambridge University Press.

Stein, A. (2005). *Foundations of Evidence Law*. Oxford University Press.

Stevenson, M. T., & Mayson, S. G. (2022). Pretrial detention and the value of liberty. *Virginia Law Review*, *108*(3), 709-782.

Stoffelmayr, E., & Diamond, S. S. (2000). The conflict between precision and flexibility in explaining" beyond a reasonable doubt". *Psychology, Public Policy, and Law*, *6*(3), 769-787.

Tadros, V. (2013). Controlling risk. *Prevention and the Limits of the Criminal Law*, 133-55.

Tillers, P., & Gottfried, J. (2006). A collateral attack on the legal maxim that proof beyond a reasonable doubt is unquantifiable. *Law, Probability and Risk*, 5, 135–157.

Urbaniak, R. (2018). Narration in judiciary fact-finding: A probabilistic explication. *Artificial Intelligence and Law*, 1–32.

Van Koppen, P. J. (2008). De beschaving van risicostrafrecht: Tussen goede opsporing en prima gevaarspredictie. *Proces*, *2008*(2), 36-46.

Vars, F. E. (2010). Toward a general theory of standards of proof. *Cath. UL Rev.*, *60*, 1-45.

— (2012). Delineating sexual dangerousness. *Hous. L. Rev.*, *50*, 855-898.

Vorms, M., & Hahn, U. (2021). In the space of reasonable doubt. *Synthese*, *198* (Suppl 15), 3609-3633.

Willems, S., Albers, C., & Smeets, I. (2020). Variability in the interpretation of probability phrases used in Dutch news articles—a risk for miscommunication. *Journal of Science Communication*, *19*(2), A03.

Zagzebski, L. T. (2012). *Epistemic authority: A theory of trust, authority, and autonomy in belief.* Oxford University Press.