Definitions for Safety Engineering

Peter Bernard Ladkin
Causalis Limited and University of Bielefeld

This document contains terms to be defined (singular *definiendum*), definitions of those terms (*definiens*), and commentary on some of those definitions. Definienda are rendered in boldface, definiens in normal, and commentaries in italic font, with commentary titles in boldface italic.

Some attention has been paid to defining all the terms used, even some commonplace terms which either have or need to be assigned a technical meaning, to eliminate ambiguity. The definition of a term may usually be found in proximity to its first use, after that use (that is, the presentation style is "lazy" rather than "greedy"). Commonplace words used with commonplace meanings are left undefined.

Harm. Human injuries and fatalities, classified by severity and accumulated, property loss and damage, environmental loss and damage including pollution; any other loss of worth considered by regulation or stakeholders to be deleterious.

Severity (of harm). Amount of harm stated in terms of agreed units (for example, numbers of fatalities, severe injuries, minor injuries; damage to property or objects in terms of amortised replacement cost where applicable; expected cost of cleanup of environmental damage).

Risk. Expected value of harm. More properly, the expected value of the severity of harm. This may be restricted to specific groups of people or to a portion of the environment, as specified, e.g. "risk to inhabitants living within 10 miles of the plant". The term "expected value" is used here as in probability theory.

Commentary on the term Risk.

1. This is the definition of De Moivre from his 1711 Royal Society paper De Mesura Sortis, in which risk was first defined and which notion is still used in finance and commerce. Most engineering notions of risk derive from this notion, but often proceed via the notion of hazard, and associated likelihoods and severities. Under substantial and often unrealistic probabilistic-independence assumptions, these engineering notions of risk are usually seen to be near to or the same as that of De Moivre. One may surmise that the original purpose of the engineering notions was to enable methods to be devised for calculating good approximations to De-Moivre-type risk. De Moivre's definition has the advantages of precision and brevity, as well as long-standing success amongst its users.

2. This definition of risk, using the terms of probability theory, may give a mistaken impression that all that matters are numbers. A number may indeed come out of it if all types of harm are taken to be comparable to each other, as for example in attempts to assess the "value of a statistical life", or VSL, which is an amount of money. Traditional Cost-Benefit Analysis takes all types of harm to be comparable, usually reducing these to monetary units using Willingness to Pay or other techniques. However, there is no presumption attaching to the definition given here that all types of harm are comparable. One may well have a vector of types of harm, for example

<property damage, human deaths, animal deaths, human disabilities, other major human injury,

species extinguished, area of ecosystems destroyed, greenhouse-gas contribution>

in which elements of the vector are not taken to be comparable to each other. Risk would then also be a similar vector, containing one value for each type of harm, with varying units, for example

< €30 million, 10, 500, 100, 300, 3, 1,000 sq. km., 5 million tonnes>

Although the values in this example are numerical, there is also no presumption that values must be numerical. Values may indeed be taken from a ratio scale, for example numbers, but may be equally be taken from an ordinal scale, as in aviation certification procedures[1], or even from scales in which some elements are incomparable[2].

3. This definition may also give the impression that all that matters are averages. But many of the significant events that worry safety engineers are "Black Swans", in the words of N. N. Taleb[3]. Averages would be a misleading indicator of risk if one believed, as Taleb, that extreme events are more likely to exhibit Mandelbrotian "scalability" than Gaussian distribution. But there is nothing in this definition to tie "expected value" to a Gaussian distribution. The expected value of harm could well include a list of Talebian extreme events along with some kind of assessment, different from those based on Gaussian distributions, of how likely one is to occur and what kind of harm may result.

Societal Risk. Expected value of harm, unrestricted.

Individual Risk. Expected value of harm, to a single person, including damage to hisher property. A specific individual may be mentioned, or an individual risk to an unknown person may be stated.

Safety. The contrary of risk.

Commentary on the term Safety.

*The term "contrary" may need some explanation. There is a certain amount of technical (formal) logic involved. Two propositions **p** and **q** are said to be contraries if they cannot both be true. Two contrary propositions can, however, both be false[4]. Risk is not a proposition, but is a property of some enterprise, that is, some general behavior, in the case of safety engineering involving (but not restricted to) the behavior of a engineered system S. Properties may be said to be contraries if, when asserted of the same object, contrary propositions result. So, for example, "red" and "not red" are contrary*

---

1 *"minor", "major", "hazardous", "catastrophic"*. See E. Lloyd and W. Tye, Systematic Safety, Civil Aviation Authority, London, 1982.

2 For example, "*ignorable*", "*intolerable*", "*acceptable as is*", "*not acceptable as is*", used in evaluations according to ALARP. Here, "*acceptable as is*" is a value indicating that costs of further risk reduction are "greatly disproportional" to tbe benefits (to adapt the wording of Lord Asquith of Bishopstone in Edwards vs. National Coal Board, 1949); and "*not acceptable as is*" a value denoting that further risk reduction is "reasonably practicable". One could also argue that such a scale best represents the political values attaching to such catastrophic harms as a 9/11-type terrorist attack, in which the societal value changed from "*acceptable as is*" to "*not acceptable as is*" in the course of a few days in September 2001. Such a scale has two ordinal chains: ("*ignorable*" < "*acceptable as is*" < "*intolerable*"), and ("*ignorable*" < "*acceptable as is*" < "*intolerable*"), as well as two values which are strictly incomparable: ("*acceptable as is*" || "*not acceptable as is*").

3 Nassim Nicholas Taleb, The Black Swan, Penguin Books, 2007. See also what I regard as a more readable work, Fooled By Randomness, Penguin Books, 2004. Taleb is concerned with rare, extreme events with significant consequences, and rejects use of the concepts of the Gaussian distribution in domains in which such events occur.

4 See the entry for *contraries* in The Oxford Companion to Philosophy, New Edition, ed. T. Honderich, Oxford University Press 2005. The entry refers to the article Contradictories and Contraries, in Logic Matters, P. T. Geach, Blackwell Publishers, Oxford 1973, but I would not recommend this article to those without training in traditional scholastic logic.

properties: "X is red" and "X is not red" are contrary propositions. There is, then, a question as to what the associated proposition would be with respect to a property like risk, which is not Boolean (that is, something has the property or does not) but admits of variation amongst many different values. Suppose those values can be taken in a Boolean lattice (a lattice with complements), then the notion of contrary may be extended as follows: properties P and Q are contraries if, for any X, the value of property P of X is the complement (in the lattice) of the value of the property Q for X. The further question now arises as to what X might be here if P is "risk" and Q is "safety". The behavior of a system in a given environment is what is to be assessed with respect to risk and safety, so the full propositions are something like "the behavior of system Z in given circumstances C results in risk R", of which the associated proposition for safety would be "the behavior of S in given circumstances C results in a level of safety S". This account says that if risk is valued on a Boolean lattice, then the associated value of safety is the complement in the lattice of the value of risk. This does require that we value risk on a Boolean lattice. But I believe this is a reasonable formal criterion, since most real-world scales on which engineering risk is evaluated turn out to be Boolean lattices. If one evaluates risk on an ordinal scale, say as a simple example "improbable" <"unlikely" < "likely" < "high", then one defines the lattice supremum to be the maximum of two values, the infimum to be the minimum, and the complement of "improbable" to be "high" and vice versa; similarly with "unlikely" and "likely". Mutatis mutandis for any ordinal scale, including the discrete scales associated with risk matrices, and any of the other decision methods for acceptance and non-acceptance of risk.

If safety is the contrary of risk, the question arises as to what are the units of a safety measure[5]*. Even if we deal with complements as above, it may not be the case that units can be calculated for safety, even though they exist for the associated risk measure. Consider the contraries "selfish" and "selfless", for example. One might choose to measure the degree to which a person P exhibited selflessness in 2008 by summing the money heshe gives to charities aiding disadvantaged people in 2008. So this measure would have units of "euros per year", say. But what units could we reasonably assign to hisher selfish tendencies in 2008? There is no obvious answer. However, suppose we measure P's selflessness in 2008 by summing the money heshe gives to charities aiding disadvantaged people in 2008 as a proportion of disposable income. Then P's selflessness has no units, but is a pure number between 0 and 1. Our Boolean-lattice calculation applies to show P's selfishness as (1 – (P's selflessness)). So whether there may be units to "safety" depends very much on how we choose to measure risk.*

For an example of choosing units related to risk and safety, consider the case of road travel in private vehicles. The individual chance of dying on German roads is roughly 1 in 15,000 per year (for serious injury, one can roughly divide this figure by ten). This seems to be a good way of stating risk. The units here are "per year" (taking "1 in 15,000" as usual to be one-fifteenthousandth). Concomitantly, safety can be expressed as 14,999 in 15,000 per year. The same units are used for both. Now say I express my risk of road travel as the number of injurious accidents I expect to have per million kilometers. Then I can express my safety as the number of kilometers I expect to travel without an injurious accident. Those units are complementary: "number of accidents per unit distance" and "distance per unit accident". Suppose now I measure my individual risk of commercial air travel as my chance of dying per flight taken (calculated for a certain group of airlines in a well-known study 20 years ago as being about 1 in 4.5 million). Then how do I measure my safety? Maybe as expected number of flights I can take without dying. Now, these units make be regarded as intuitively complementary, but I am not sure how far this intuition leads to a rule. It seems to be that one may choose to measure risk in such a way that safety has a complementary value (as explained above) with complementary units, but also that one may choose to measure risk in such a way that there may be no reasonable units of its contrary,

5 I thank Peter Bishop for raising this point.

safety. Although "unit calculations" work for physical quantities in physics and there respect complements (the units of the complement are the complement of the units), I do not know that they must work for arbitrary units of varying-value properties in general.

Accident. An event whose causal consequences include harm. The harm must usually be significant for the term to be appropriate.

Event. A change of state. An event is specified by giving two states. It may be identified logically with the ordered pair of the two states, which are in turn termed the **pre-state** of the event and its **post-state**.

State. Of a collection of objects. A complete collection of properties of the objects under consideration and their relations to each other. "Complete" means that nothing may be added. In specification, one is normally concerned about a **relative state** in which only a subcollection of properties and relations are considered.

Object. Anything reasonably denoted by a noun. A **collection of objects** may be identified logically with the set of those objects.

System. A collection of objects, with or without specified properties and relations of concern.

Environment. Of a system. The collection of objects which do not belong to the system, but which have relations with some of the objects constituting the system. Usually specific relations are of concern, and others not. In this case, we speak of the **world** as containing all objects which have some relation or other, and the environment as the subcollection of the world containing those objects which have relations of concern.

Behavior. Of a system, a system with its environment (**joint behavior**), or a collection of systems (**joint behavior**). A temporal sequence of events whose constituent states involve objects in the system, or system and its environment, or collection of systems. The sequence may be of unspecified length, and may also be non-terminating.

Process. A term for a behavior which, for the purposes of a specific causal analysis, is taken to be a unit, as if it were an event with more states than two, drawn out in a measurable but bounded interval of time.

Phenomenon. Concerning a system S, or concerning a system together with its environment S+E. A state involving objects in the system or an event or process involving objects of the system in its constituent states.

Consequence. Of an event. The harm that causally results from a given failure or accident.

Failure. A phenomenon of a system or subsystem in which the system or subsystem does not perform its required or expected function. A **dangerous failure** is a failure of which the risk (of continued system operation) in the presence of the failure is substantially raised over the risk of continued system operation had the failure not occurred.

Fault. Associated with a failure. A specific phenomenon, in the causal history of a failure, that is regarded by engineers or other specified persons as anomalous or out-of-specification, and whose occurrence in combination with other phenomena which are within expectation or specification, as well

as possibly with other faults, caused the failure. Or such a specific phenomenon which would so cause a failure but has not yet done so.

Commentary on the term Fault*. Since there are many phenomena in a causal history (see definition below) of a failure F, the notion of a fault which caused F selects any of them. Specific phenomena from amongst these are selected, and thereby termed "faults", usually through some method of selecting particular types of anomaly for attention, or by comparing with a specification.*

Cause. Of a phenomenon P. A phenomenon occurring in the causal-factor graph of a specified phenomenon P.

Causal-Factor Graph. Of a phenomenon P. One constructs the graph as follows. One enumerates a collection C of **necessary causal factors** (NCFs) of P, such that the phenomena in C constitute a causally sufficient collection for the phenomenon P to occur. Then one performs the same operation for each phenomenon in C, in turn. And repeats the operation as long as desired. When finished, accumulating all the phenomena enumerated at any stage, including P itself results in a collection CE. The **causal-factor graph** is the collection CE together with the binary relation *phenomenon A is a necessary causal factor of phenomenon B* for phenomena A and B in CE. Such a set with a binary relation constitutes logically a mathematical graph.

Commentary on the term "causal-factor graph".

1. Often in engineering the phrase "chain of events leading to", or "causal chain of events leading to" a phenomenon P is used. This phrase is misleading in that, in most cases, the phenomena causing P will be arranged not in a causal chain, which is a linear sequence of phenomena, but in a causal network, in mathematical terms a graph. Neither will those phenomena be restricted to events, but they will also include states, as well as maybe specific behaviors which are not further analysed. The term "causal-factor graph" fits this situation more closely than the term "causal chain of events", which is misleading.

Necessary causal factor. Phenomenon A is a necessary causal factor (NCF) of phenomenon B just in case A and B both occur but B would not have occurred had A not occurred. This is the counterfactual or contrary-to-fact notion, well established amongst researchers into causality as well as in many causal-analysis methods in engineering.

Commentary on the term "necessary causal factor".

The counterfactual notion was considered in detail in the early 1970's by the experts in causality J. L. Mackie[6] and David Lewis[7], and has withstood critique for over three decades[8] The counterfactual notion of causation for explanation of individual phenomena (as opposed to repeatable regularities) is well-established amongst expert commentators on causality, and originates over two hundred years ago with David Hume (his "second definition"). The counterfactual notion is used in accident explanation for example in Andrew Hopkins's Accimaps, used by the Australian Transport Safety

6 In Chapter 2 of J. L Mackie, The Cement of the Universe: A Study of Causation, Oxford University Press, 1974

7 In David Lewis, Causation, J. Philosophy 70:556-67, reprinted in David Lewis, Philosophical Papers Volume II, Oxford University Press, 1986, along with an extensive Postscript considering (and mostly answering) points raised in critical discussion of the notion in the decade after publication.

8 A useful recent collection of articles by prominent experts and researchers in causality is John Collins, Ned Hall and L. A. Paul, eds., Causation and Counterfactuals, MIT Press, 2004.

Bureau, which are counterfactual causal-factor graphs containing a selection of generalised factors. It is also used in Why-Because Analysis (see below), and by air accident investigators for the U.S. Air Force. For Accimaps, see Andrew Hopkins, Safety, Culture and Risk, CCH Australia Limited, Sydney 2005, and Lessons from Longford, CCH Australia Limited, Sydney 2000. For the U.S.A.F.'s methods, see Air Force Instruction 91-204, July 2004, and the book by the author of the instruction, Richard H. Wood and Robert W. Sweginnis, Aircraft Accident Investigation, 2nd Edition, Endeavor Books, 2006. For Why-Because Analysis, see Causal System Analysis, by Peter B. Ladkin, or the Why-Because Analysis Home Page available at www.rvs.uni-bielefeld.de . The sociologist Scott Snook uses the counterfactual notion in his Causal Map of the friendly fire shootdown of two U.S. Army helicopters in Iraq in 1994 by two U.S.A.F. Fighters, in Figure 1.3, p 21 of Scott A. Snook, Friendly Fire, Princeton University Press, 2002. However, Snook mistakenly rejects the counterfactual notion as inadequate to present an explanation of the incident, although it is clear that his Causal Map is inadequate. It tuns out that Snook's painstaking and insightful sociological analysis of the causes may be represented, in detail, and including the sociological and psychological theories he adduces, in a Why-Because Graph, as Ladkin et al. showed. See Peter Ladkin and Jörn Stuphorn, Analysis of a Friendly Fire Accident with WBA, in the Third Bieleschweig Workshop, on-line proceedings available through www.rvs.uni-bielefeld.de .

However, many engineers and experienced accident investigators use an intuitive meaning for causal factor (or necessary causal factor) that roughly coheres with the counterfactual notion but is not as rigorous. This notion is roughly that phenomenon A is a NCF of phenomenon B if and only if an occurrence of B requires (in some sense of "requires") an occurrence of A. The main question here is the question of necessitation, what "requires" means: in particular "necessary" with regard to which background of events and states? A "normal" background in some sense? Or one which is abnormal but not necessarily rare? One in which the laws of physics hold? Exactly how such intuitive notions differ from the counterfactual notion is not clear to me.

If the meaning of the counterfactual "B would not have occurred had A not occurred" follows the rigorous semantics for counterfactuals given by David Lewis[9], then the causal-factor graph is called a ***Why-Because Graph*** *or WBG.*

Causal History. Of a phenomenon P. The collection of phenomena in the causal-factor graph of P.

Factor. In a causal-factor graph of P. A phenomenon which belongs to the causal history of P.

Potential Factor. A specific phenomenon which is considered for inclusion in the causal history of P.

Root-Causal Factor. Of a phenomenon P. A phenomenon in the causal history of P for which there are no causal factors in the causal history of P. Equivalently, a leaf node in the causal graph of P.

Root Cause. Another name for *root-causal factor*. However, there is a related but somewhat variant use as a collective term for a group of related root-causal factors.

Commentary on the term "root cause".

For an example of the use of the term as a collective, consider the following. Equipment operators working together perform various inappropriate actions, which become root-causal factors of an

9 David Lewis, Counterfactuals, Basil Blackwell Ltd, Oxford 1973, reissued Blackwell Publishers, Oxford 2001.

incident. It may be said,collecting all these inappropriate actions together, that a root cause of the incident was inappropriate operator behavior or (more sophisticated but semantically identical) poor personnel-resource management.

Stopping Rule. The construction of a causal-factor graph of a phenomenon P must terminate, if the causal-factor graph is ever to be used. The process of inquiring after potential factors of factors must therefore terminate. A **stopping rule** is an explicit criterion for terminating the inquiry into potential factors of a factor Q. One **stops with Q**.

Factor-Identification Rule. In constructing a causal-factor graph of a phenomenon P, a finite list of potential factors must be chosen in a way that is manageable within the resources devoted to the construction. The entire list of properties and objects in the world does not constitute a list manageable within the resources of any such project. Thus objects, their properties and relations with other objects, and phenomena associated with these, must be selected. A factor-identification rule is an explicit method for this selection.

***Commentary on the rules for causal-factor graphs**. Since many factor-identification rules may be defined, as well as many stopping rules, there will not generally be one unique causal-factor graph for a phenomenon P, but rather one for each selection of factor-identification rule and stopping rule, which together uniquely determine a causal history of P.*

Error. An anomalous or out-of-specification phenomenon associated with requirements specification or analysis, design, implementation, operation or maintenance of a system. The general term "error" includes all these categories, but the definition of an error within each category is very different. A **requirements-specification error** is a mismatch between phenomena in the environment of a system in use and the use of a system as specified in its requirements specification. A **design error** is the failure of a system design to fulfil its requirements specification. An **implementation error** is the failure of a system to fulfil its design specification. A **maintenance error** is an event whereby either specified known faults are left in the system, or faults are introduced into the system, caused by system maintenance activity. An **operational error** is an event whereby a system is brought into mismatch with its behavioral specification without a design fault being present. A **non-specific error** is an anomalous phenomenon which does not fall into any of the above categories.

Hazard. A phenomenon of a system, or its environment, or both, which substantially raises risk, although the likelihood of an accident still remains less than certain. A **system hazard** is a system phenomenon which is a hazard. An **environmental hazard** is an environment phenomenon which is a hazard.

***Commentary on Hazard**. A variation on the term "hazard" defines a system hazard to be a system phenomenon, which along with some reachable state of the environment, is a sufficient cause of an accident; mutatis mutandis with environmental hazard. Here, "reachable state" means a state which can plausibly occur, although its occurrence is not necessarily likely. Such a definition must be extensively refined, however, to exclude states which inevitably arise as a necessary causal consequence of operating a system according to intent: as it stands it would recognise even the routine flight of a commercial aircraft to be a "hazard", because flying into a convective storm of sufficient strength (a "reachable state of the environment") would cause a loss of control. How to refine such a definition adequately is a matter for further research.*