Risk-based Decision-making Fallacies: Why Present Functional Safety Standards Are Not Enough

Andreas Johnsen, Gordana Dodig Crnkovic, Kristina Lundqvist, Kaj Hänninen, and Paul Pettersson

School of Innovation, Design and Engineering Mälardalen University Västerås, Sweden

{andreas.johnsen, gordana.dodig-crnkovic,kristina.lundqvist,kaj.hanninen,paul.pettersson}@mdh.se

***Abstract*—**Functional safety of a system is the part of its overall safety, understood as freedom from unacceptable/unreasonable risks that depends on a system operating correctly in response to its inputs. Functional safety elements are examined at every stage of the software development life cycle, including requirement specification, design, implementation, verification, validation and deployment. Acceptability of risks is judged within a framework of analysis with contextual and cultural aspects by individuals who may introduce subjectivity and misconceptions in the assessment. While functional safety standards elaborate much on the avoidance of unreasonable risk, little is addressed on the issue of avoiding unreasonable judgments of risk. Through the studies of common fallacies in risk perception and ethics, we present a moral-psychological analysis of functional safety standards and propose plausible improvements of the involved risk-related decision making processes, with a focus on the notion of an acceptable residual risk. As a functional safety reference model, we use the functional safety standard ISO 26262, addressing potential hazards caused by malfunctions of hardware and software systems within road vehicles, and defines safety measures that are required to achieve an acceptable level of safety. Analysis points out the critical importance of a robust safety culture with developed countermeasures to the common fallacies in risk perception, which are not addressed by contemporary functional safety standards. We argue that functional safety standards should be complemented with the analysis of potential hazards caused by fallacies in risk perception, their countermeasures, and the requirement that residual risks must be explicated, motivated, and accompanied by a plan for their continuous reduction. This approach becomes especially important in contemporary developed autonomous vehicles with increasing computational applications.

I. INTRODUCTION

The foremost requirement in the development of safety- related E/E (electrical and/or electronic) systems is to not cause hazards that are more frequent and more severe than acceptable [1]. In other words, the risk associated with the system must be below a certain limit. To control that safety limits are not exceeded, governments enforce functional safety standards through which products and services must be certified. There exist a range of domain-specific standards, such as for automotive, aviation, railway, and nuclear applications, and a handful generic from which these typically are derived. For example, ISO 26262 [2] is an automotive-specific interpretation of the basic functional safety standard IEC 61508. ISO 26262 essentially provides a safety lifecycle reference model that complies with standardized safety requirements in the development of E/E systems within road vehicles. The reference model both addresses potential hazards caused by malfunctions and specifies safety measures through which safety is achieved. The standard defines safety as *the absence of unreasonable risk*:

> "*risk judged to be unacceptable in a certain context according to valid societal moral concepts*."

Other than a definition of risk: "*the combination of the probability of occurrence of harm and the severity of that harm*," there is no further elaboration on its meaning. As it turns out, functional safety standards are ultimately dependent on applied ethics, which may not be a surprise as harm is one of their central concerns. However, in front of the complex, theoretical nature of functional safety standards, the importance of their rather subjective foundation, i.e. judgment of right and wrong conduct, is easily forgotten.

The notion of harm originates from emotions; harm, physical as well as mental, instinctively causes unpleasant emotions. Together with the ability of reasoning, humans are able to develop models of right and wrong conduct – ethics. The "*do no harm*"-principle is fundamental to ethics [3], derived from the value of human dignity and the respect of the personal integrity. In applied ethics, the fundamental *principle of beneficence* refers to a moral obligation to "*act for the others' benefit, helping them to further their important and legitimate interests, often by preventing or removing possible harms.*" [4]. Principles like these may subsequently be applied to improve behavior and decision making such that harm to people and the environment is prevented or mitigated.

Models of right and wrong conduct change over time in response to gained experiences and knowledge and with an ever-changing environment. In recent centuries, changes have especially been made with respect to the development and expansion of computer software technology. In addition, there are as many interpretations and practices as there are individuals. It appears impossible to find a universally accepted set of rules of moral conduct that would be valid irrespective of context. However, by a continuous motivation for increased awareness, concepts and principles may be improved by continual refinements. Nevertheless, a critical problem still prevails, which is the fact that human reasoning often includes untrue impressions of the world and sometimes is driven by irrational motives. No matter on which level of expertise reasoning is conducted, scientific studies repeatedly show that even deliberated thoughts supported by statistics often include substantial errors [5]. If standards control functional safety by the absence of unreasonable risk, i.e. by the absence of risk judged to be unacceptable, we argue it is critical to also require the absence of unreasonable judgments.

The question is whether the reasoning behind an "acceptable residual risk" includes untrue axioms, inferences, or theorems. Such errors have the potential to result in unethical conduct, in spite of functional safety standards and their careful application. In addition, memories and emotion intensities are in many cases disproportional with respect to the stimuli that cause them. For example, unlikely harmful events often induce irrationally high intensities of unpleasant emotions or are ignored altogether. This is a valid problem of risk-based thinking as irrational feelings affect judgment even when evidence of their irrationality is presented [5]. Adjustments must therefore often be made for irrational feelings even when they seem to have been accounted for by reasoning. On the other hand, emotions to a large degree determine the well being of humans. Irrational emotions should consequently not be neglected in the judgment of unreasonable risk even though they correspond to untrue impressions, for the fear of an unlikely accident may be as harmful as the accident itself.

In this paper, we present a moral-psychological analysis [6] of risk-related decision making processes conducted within the area of functional safety. Moral psychology is a research field that brings together evolutionary, neuro-scientific, cognitive, psychological, cultural and societal perspectives to the questions of the nature of morality that are traditionally studied by ethics. The analysis is performed on generic concepts and principles of functional safety, where ISO 26262 is used as a functional safety reference model. We do not intend to answer whether contemporary functional safety standards correspond to right or wrong conduct, but rather to provide ideas of how risk-related decision making processes may be made more reasonable in the development safety-critical systems. Our approach is to first make concepts, principles, and ethical issues of functional safety explicit. We then identify relevant systematic errors of thinking and analyze the plausible impact these fallacies may have on functional safety in combination with ethical issues. Finally, the analysis is used to provide guidelines for how risk may be judged in a more reasonable manner. Kahneman's book "Thinking, Fast and Slow" [5] – partly based on papers [7][8][9][10] published by Tversky and Kahneman – is used as foundation in the study of fallacies.

The paper is organized as follows. In section II, an overview of functional safety as described by ISO 26262 is presented. In section III, the main ethical functional safety issues are presented. The systematic errors of thinking and examples of their negative effects in judgments of risk are then listed in section IV, which accordingly are followed by ideas and guidelines of plausible improvements in section V. Concluding remarks are finally presented in section VI.

II. OVERVIEW OF ISO 26262

ISO 26262 essentially addresses potential hazards caused by malfunction of safety-related E/E systems and provides the necessary safety measures to achieve an acceptable level of safety. Automotive safety integrity levels (ASILs) are provided for the classification of hazards – low to high risk. Classification of a hazardous event is essentially based on its frequency of occurrence, the human controllability to avoid an accident in case of its occurrence, and the potential severity of the resulting harm or damage. In turn, each ASIL specifies safety requirements, such as mechanism for error detection and error handling that must be achieved to reach an acceptable residual risk. Defined confirmation measures, such as examination and assessment, must finally be performed to ensure achievement.

As a reference process model, the standard uses a V-model to represent the different phases of the system development. The model mainly consists of three phases: Concept phase (part 3 of the standard), Product development (part 4, 5, and 6), and Production and operation (part 7).

In the concept phase, the item to be developed in compliance with the standard is firstly defined. This entails in defining the functional, non-functional, legal, and already known safety- requirements of the item. Potential hazards of the item are then identified and ASIL-classified through hazard analysis and risk assessment. Safety goals (SGs), which inherit the ASILs of the corresponding hazards, shall concurrently be formulated for the identified hazards. These SGs describe characteristics needed to avoid hazards or to reduce risk associated with the hazards to an acceptable level. Functional safety requirements (FSRs) shall then be specified for each SG. FSRs describe basic safety mechanisms, implementation-independent safety- related behavior, and safety measures that have to be provided by elements in the primarily assumed system architecture for complying with the SGs and their ASILs. FSRs do only consider functional aspects of the system and not how these are technically implemented in software or hardware. FSRs inherit the same ASILs as the corresponding SGs and shall be allocated to elements of the primarily assumed system architecture.

FSRs are decomposed in the product development phase into technical safety requirements (TSRs), which describe how to implement the safety mechanisms and safety measures described by the FSRs in software or hardware. TSRs are succeeded by the development of a system design. Verification must be conducted to ensure compliance of the design with respect to the TSRs.

The Production and operation phase finally impose necessary directives on the production and maintenance process to ensure functional safety.

In addition, the standard provides a vocabulary (part 1), requirements of the institution responsible for the complete safety lifecycle and its individual activities (part 2), supporting processes (part 8), and ASIL-oriented and safety-oriented analyses directives (part 9). Part 2 includes some requirements on the safety culture, defined as:

> "*policy and strategy used within an organization to support the development, production and operation of safety-related systems*."

The main requirement of a safety culture, according to the standards, is:

> "*The organization shall create, foster, and sustain a safety culture that supports and encourages the effective achievement of functional safety*,"

where evidence of competence, organizational-specific rules and processes for functional safety, and evidence of quality management, must be produced. ISO 26262 examples of a good safety culture include: traceable accountability; safety as the highest priority; a system that rewards effective achievement of functional safety and penalizes those who take short- cuts that jeopardize safety or quality; appropriate degree of independence in the integral processes; proactive attitude towards safety; the required resources and competences are allocated; intellectual diversity is sought and used to advantage; and existence of supporting communication and decision-making channels, where self-disclosure and disclosure of discovery by anyone else are encouraged.

III. FUNCTIONAL SAFETY ETHICS ISSUES

The foundation of applied ethics is the principle of beneficence [4]. The central role of technological ethics is therefore to derive policies that at least protect humans and the environment from harm induced by technology, and at best also support their flourishing. However, a possibility of harm exists in any given situation, whether it be maliciously intentional or accidentally unintentional. The limit from which justified (acceptable) harm turns into unjustified (unacceptable) harm must consequently be analyzed in order to provide guidance in situations where harm seems unavoidable. General challenges of ethics, such as privacy, trustworthiness, respect, responsibility, fairness, caring, values, virtues, and balancing freedom with authority, do also apply in the development of technology.

The innovation, evolution, and dramatically increased use of computer software technology have led to tremendous benefits in addition to a platform through which societies may evolve faster than ever before. However, the pace at which it has been integrated into our everyday lives raises concerns about whether we have ignored the possible unethical implications. A fast pace of development causes a degree of speed blindness, where benefits might be focused upon, while potential losses are ignored. For example, the responsible source of harm is seldom evident when harm is caused by malfunctioning computer systems, partly due to its complexity, the human-machine interaction, and the vast number of actors behind its design, production, operation, maintenance, and certification. In addition, control by non-human systems causes a decreased sense of responsibility as the distance and time between the human act (e.g. software design) and its possibly negative effects (e.g. accident) are increased. Such conditions make it difficult to identify the source of wrong conduct and unreasonable design decisions.

The evolution of computer technology may intuitively be viewed as causing more benefits than losses. However, an intuitive impression does not imply that it is ethically justified in every respect. No benefit justifies unjust means and humans are easily blinded when benefits are great. Perhaps the largest issue in this regard is environmental sustainability. The environment through which life is possible has been shown to be endangered by our own creations. The industrial revolution, for example, has led to tremendous benefits but also vast amounts of pollution and fatalities. The use of fossil fuels has, through the greenhouse effect, the potential to change the climate to the point where human life is no longer possible. An increased awareness of the long-term effects of the use of fossil fuels has altered the judgment of its acceptance. On the other hand, the industrial revolution has also paved the way for advanced technology, which now in many cases is used as a means of reducing the amount of pollution.

For example, mechanics and hydraulics within vehicles and aircraft are being replaced by E/E systems partly to make them lighter and more fuel-efficient. In addition, fuel engines are being replaced with more efficient electric motors. The question is whether the long- term effects of an increase of electricity generation, electrical wiring, electronic components, and batteries is environmentally sustainable, and if not, whether it is less negative than the effects of fossil fuel consumption and weather the benefits are worth the risk of collapsing the ecological system. In order to completely replace fossil fuels with renewable energy, the increase and side effects, such as pollution from batteries, ground disturbance from land and vegetation clearing for wind turbine and solar panel installations, water flow disturbance due to hydroelectric power stations, etc., will be substantial and must not be neglected to avoid unreasonable judgments.

The world is constantly changing, and so are also our attitudes toward right and wrong. The future of technology cannot be predicted with certainty. With such premises, judging whether a design will have major negative consequences or not appear nearly impossible. The only means of achieving morally justifiable practices seems to be continuous regulation by a broadly framed perspective of knowledge and awareness. In the following subsections, we describe six areas of ethical issues, which we argue are critical to risk-related decision making processes within the area of functional safety.

A. Diversity of Judgments

The amount of risk an individual judges unacceptable differs greatly between individuals. Some individuals find the high risk of injury in professional boxing as unacceptable whereas some find it acceptable. The same principle applies to the society depending on the context; professional boxing and public transportation are typically accepted activities even though the risk of injury in professional boxing would be unacceptable in public transportation. Attitudes toward risk vary among both individuals and contexts. In general, there is a significantly larger acceptance towards voluntary risk com- pared to involuntary risk (risk which is out of an individual's control or knowledge). However, attitudes toward risk tend to be normally distributed and the majority rule tends to be accepted as the ethical solution when unanimous agreements cannot be achieved. A public judgment of unacceptable risk based on averages and the majority rule is by these premises the moral method to publicly judge risk. Such judgments are sound as long as the majority possesses a truthful impression of the actual risk. This is often not the case as described in Section IV, especially in unfamiliar contexts. For example, an irrational amount of fear is typically induced in contexts which contradict natural human conditions, such as flying, where the typically more accurate standpoint of domain experts may vastly differ from the majority of the public.

B. Vision Zero and Zero Tolerance

The principles of *vision zero* and *zero tolerance* are often applied by governments to behaviors that cause harm. Vision zero is applied in engineering of road systems with the aim of achieving traffic with no fatalities or serious injuries. Zero tolerance is applied to eliminate harassment, violence, illegal narcotics, driving under the influence of alcohol, and illegal weapons. The strength of these principles is the evident goal they indicate. Goals, if embraced by individuals, tend to make a large difference on how they make decisions in regard to actions that influence them [5]. The argument behind vision zero is that lives can never be exchanged with societal benefits. Neither fatality nor severe injury is by definition acceptable. In the domain of functional safety standards, on the other hand, the principle is rather that a degree of fatalities and injuries larger than some small number is unacceptable. In other words, a small degree of fatalities and injury is acceptable, but not more.

The question is whether an application of the vision zero principle to functional safety would result in safer systems even though total safety cannot be achieved. The immediate problem is how the certification process would be conducted since a product or service never could comply with the principle. A solution could be to force enterprises to provide legitimate arguments to their decisions of why an increase of safety cannot be achieved, whether it be due to cost or lack of better methods. This requirement for providing the justification for residual risks, which would complement the existing requirements for demonstration of risk prevention, would raise awareness of their existence and increase the probability of future mitigations.

C. Wants vs. Needs

Neither flying, driving, artificial intelligence, or energy generation through nuclear power plants, are human needs of survival and reproduction, but rather human wants. The question is whether it is ethical to promote such wants even though they inevitably result in environmental degradation, fatalities, and severe injuries. One could argue that the industrial and technological evolution has resulted in an increased population, standard of living, and life expectancy. Such arguments do not take into consideration the resultant ecological footprint, the right and standard of animal and plant life, and the potentially catastrophic long-term effects. Only after taking into account broad perspectives of both benefits and losses in the decision making process, we are able to determine if a decision is justified. Since human reasoning commonly is misled by focusing illusions and narrow framing of problems, as will be discussed in Section IV and Section V, there is a need for policies that remind decision makers of their existence.

D. Business

Unacceptable risk is often determined in relation to benefit and cost. The main driver of enterprises developing safety-critical systems, however, is not safety but rather profit. From a business perspective, benefit is profit and risk translates to cost, such as liability and deficit (degraded marketing) in response to accidents. Mykytyn et al. define product liability as: "*the legal liability of manufacturers and sellers to compensate buyers, users, and even bystanders for damages or injuries suffered because of defects in goods purchased*" [11]. Liability is further divided to intentional liability and strict liability. The former requires an intentional act that is reasonably foreseeable to cause harm and the latter requires no intent or negligent act. Both may lead to punishment. According to Dowlatshahi: "*courts have shown little mercy for manufacturers who neglect safety and who produce products that later prove to be unsafe*" [12]. Kienle et al. emphasize, based on a literature study on liability risks, that "*it is important that the company can show that it follows general guidelines (e.g., professional codes of conduct/ethics/practice) as well as applicable (safety) standards*" [13]. However, from the results of an industrial questionnaire, it is evident that the respondents rank safety culture as more important than external standards to deal with liability concerns. In fact, safety culture is also rated as more important than risk analysis, internal standards, and legal council. These results suggest moral concepts applied in everyday practices are more important for safety than meeting standards alone. Based on the findings, functional safety standards should enforce a more extensive set of requirements on safety culture. We argue enterprises should conduct design decisions following *the principle of beneficence*, with the highest imperative of increasing safety and not profit. Profit should be a result of a value for the customer, where safety is a core value that cannot be offered for profit. In such a value-oriented business model, where safety is paramount in the decision making process, the task of governments would rather be to control that increased safety is rewarded than to control that systems do not introduce unacceptable amounts of risk. It is also reasonable to believe such a model would reduce long-term negative effects as safety requires quality which is fundamental to sustainability [14].

E. Law, Regulations, and Policies

Leveson presents a socio-technical control model of the interactions between organizations and external governmental stakeholders [15]. The model essentially describes that the legislature (typically composed of politicians as legislators) enacts laws, which may be further refined by regulatory agencies with more concrete guidelines. These laws are further shaped by court decisions, which interpret their meaning to real cases, denoted as case law. Enterprises then respond to the legal system by creating internal guidelines, standards, and safety policies. These are further decomposed within individual projects to concrete practices. From this point, it is critical to engage feedback in the opposite direction, where experiences may cause changes to internal policies and standards, and where incidents and accidents may cause changes within the legal system [3]. We believe that transparency between risk- based decision making processes and legislatures is crucial for the optimization of the legal system, such that the root causes of unreasonable design decisions can be avoided through regulation.

F. Evolution, Innovation, and Sustainability

The history of civilization yields an evolution through which the biophysical environment continuously has been used by humanity for our own purposes, on the expense of other species. Perhaps even in disfavor of our own in the long term. Nevertheless, each species evolve more or less under the same principle. The difference between humanity and other species is that our creativity has reached the ability to control and manipulate the entire ecosystem, possibly to a point of mass extinction [16]. Some of the biggest threats are pollution, nuclear technology, biotechnologies, and artificial intelligence, where nuclear weaponry and power plants have the potential to collapse the global ecosystem within a short period of time. There is evidently a need to protect humanity from its own creations. According to Kemp [17], many view material products rather than social progress as improved life quality. Kemp argues that the environmental problems are caused by the results of science and technology, which individuals alone cannot solve due to their complexities.

Evolution is a gradual process where changes require time. Slow development of an ecosystem is a requirement for its stability, so that the environment has enough time to compensate for changes. The critical question is whether the immediate environmental changes now caused by humanity have induced instability within the ecosystem, possibly where thresholds have been exceeded such that deflections are magnified instead of compensated. With respect to functional safety in particular, the question is whether the benefits of computer software technology are used in a sustainable manner and whether revolutionary steps are conducted too quickly or with too much uncertainty. For example, applications within the domains of automotive and aviation generally reduce the cost of transportation due to increased efficiency. It is reasonable to assume that the technology develops to an increase of traffic, which is not unproblematic from a safety perspective. Safer systems, because of the rebound effect, might lead to an increased amount of pollution, in addition to a total increase of fatalities and injuries, by using a broader frame of perspective. We argue that possible misuses of benefits should be addressed in the decision processes and the system design, to prevent development of unsustainable systems. This measure may also counteract cognitive biases, which in many cases may induce naivety to such indirect effects, as we will discuss in the following section.

IV. FALLACIES OF RISK PERCEPTION

The perception of risk is related to the concept of trust [1]. Trust (or mistrust) is not instantaneous or permanent, nor all or nothing. Trust has a spectrum and is developed (or deteriorated) over time and experiences and varies among individuals.

A. Issues of Using The Majority Rule

Determining the public's opinion on an unacceptable level of risk is consequently difficult. Based on such premises, an application of the majority rule appears to be the ethical choice of conduct. However, there are scenarios in which judgment by the majority rule may be highly irrational compared to judgment by domain experts. If the majority base their attitudes toward risk by compromising two opposed positions, i.e., assuming the truth lies between the two extremes, the judgment will likely be untruthful if any of the two positions propagate false information. This is an informal fallacy referred to as *argument to moderation* [18]. Furthermore, the public is highly shaped by media, which has the potential to induce an unrealistic impression to the majority through biased information. The shaping effect is also substantial between individuals in *groupthink* as it causes a loss of individual creativity, unique- ness, and independent thinking due to group pressure. Finally, stereotypes and publicly accepted theories induce a blindness to their flaws, known as *theory-induced blindness* [5]. Theories correspond to the currently most accurate description of phenomena and are constantly improved as science evolves. On some occasions, flaws may later condemn theories as false and be rejected altogether. However, the transition from a proven false theory to a corresponding practical change of the society is long and effortful. The problem is that we are reluctant to change believes once they have been founded. Even the scientific community possesses this property [5]. On a more individual level, similar problems emerge from *stereotypes* [5]; humans tend to be willing to infer the general from the particular but unwilling to deduce the particular from the general.

B. Fallacies of Individual Judgment

Individual reasoning frequently suffers from systematic errors. One essential problem is that reasoning is based on emotions, which intensities are not linearly distributed over the scale of stimuli. The critical problem with judgment of risks is when they are small or large. In such judgments, we tend to either give them far too much weight or ignore them altogether, depending on the context. These properties essentially explain the rather lucrative businesses of lottery and insurance. Humans also tend to become irrationally risk seeking when all options result in a loss and irrationally risk aversive when all result in a gain [5]. These properties explain why favorable settlements often are rejected and unfavorable settlements often are accepted between disputing parties. In addition, a significant number of humans overestimate their abilities and in many cases have a poor sense of probability. In [19], a survey was conducted to conclude how drivers estimate their driving skills. The conclusion is that 90 percent of drivers believe they are better than average, which cannot be true given that skills are normally distributed and that estimations are based on the same definition of what driving skills are. This is known as *the above-average effect*. Related to this phenomenon is *the Dunning-Kruger effect* [20], where unskilled people do not only have a tendency to reach erroneous conclusions and poor decisions, but also have a tendency of being unable to realize their incompetence. With respect to probability, Seymour and Veronika [21] discovered through an experiment that a majority of the subjects preferred a winning chance of 9/100 compared to a chance of 1/10. The choice of probability representation turns out to be highly important to avoid unreasonable decisions. This fallacy is known as the *denominator neglect* [5], where the vividness a number brings makes humans ignore the context on which it is based.

There are several causes for these fallacies. Perhaps the main problem is an excessive confidence and a reliance on heuristics: opinions based on memory availability, guesses, and feelings. We typically know less than we believe we know and are reluctant to acknowledge our ignorance and the uncertainty [5]. The problem is amplified by the ~~notion of~~ *confirmation bias*: humans tend to search for, interpret, favor, and recall information that confirms their believes. Judgments thus often stand in direct contrast to sound scientific methods, where hypotheses are tested by trying to find evidence that refutes them, rather than evidence that confirms them. Confirmation bias also causes a tendency to *belief perseverance*: the inability to change beliefs, even when exposed to evidence of the contrary. Humans also tend to generalize bits of in- formation and assume they are true for all properties of the studied phenomenon, even when there is little to no correlation between the properties. This type of cognitive bias is known as *the halo effect* [5].

C. Biases due to Memory Mechanisms

The mechanisms of memory also influence decision making. For example, the ease with which issues can be retrieved from memory determines its relative importance, known as *the availability effect* [5]. By this premise, media to a large degree determine the public's attitude to risk within functional safety. In addition, familiarity causes cognitive ease, which in turn causes an impression of truthfulness. Humans also tend to arrange events into a coherent description even though they are random and independent. Coherency of events, similarly to familiarity, also causes cognitive ease and thereby an impression of truth. Judgments of probability are highly vulnerable to coherency. A problem is that we tend to confuse coherency with probability, known as *the conjunction fallacy*. Tversky and Kahneman [22] conducted a study where the subjects were asked: "*Which alternative is more probable? (1) Linda is a bank teller, or (2) Linda is a bank teller and is active in the feminist movement*." The vast majority of respondents chose the second alternative, which cannot be true since the probability of the second option is a subset of the first. The second option, in contrast to the first, causes a feeling of coherency and thereby, in this case, a false impression of truthfulness.

Extreme events, such as extraordinary accidents, are thus likely to be assigned a cause-and-effect-based explanation instead of luck or misfortune. Extreme events do happen by chance, especially in systems with many uncertainties. Severe accidents within the domain of commercial aviation, for example, are often caused by multiple unrelated extraordinary events that unfortunately took place in a sequence that lead to the accident. A chain of unrelated improbable events that may lead to an accident is close to impossible to predict, not least to predict them all. A related problem is that the outcome of an event (or chain of events) will change our memories of what we believed prior to the event in line with the outcome, known as *hindsight bias* [5]. Consequently, the unlikely chain of events that led to the accident will appear as certain in hindsight. However, the certainty is illusory. This, in turn, makes us prone to unjustifiably criticize good judgments with unlucky outcomes and unjustifiably award bad judgments with lucky outcomes. Another related problem is the phenomenon of *regression toward the mean* [5]. Extreme events will by the principles of probability be followed by less extreme events. Since we are reluctant to accept the random nature of the world, we are likely to assign an illusory cause-and-effect- based explanation to a change toward the mean, instead of this principle.

D. Biases by Responsibility

Humans have a tendency to react stronger to mistakes of commission compared to mistakes of omission even though the resultant harm is the same. Decisions to act, in contrast to the default state of inaction, cause a higher feeling of responsibility. Humans are therefore prone to not respond to risky situations with actions even though they are beneficial.

E. What You See Is All There Is

Another critical problem of decision making is *the question substitution* humans tend to make when faced with difficult questions [5]. Instead of answering the original question, we tend to answer an easier one instead, often without being aware of the substitution. The substitution is typically made toward a more intuitive, subjective, question. For example, the question "Is it safe to drive?" could be substituted with "Do I feel safe when driving?". Such questions do not only not answer the initial question, but are also distorted by a *focusing illusion* [5]. We tend to treat problems in isolation, i.e. we think that what we see is all there is (denoted WYSIATI – What You See Is All There Is – by Kahneman [5]), and also overestimate their importance when we think about them. The general driver generally feel safe when driving, but when asked explicitly, experiences may be retrieved such that the driver may think otherwise. Consequently, even the question "Do I feel safe when driving?" will often not be answered, but rather "Do I feel safe when driving when I think about it?". Acquired information from memory causes emotions that tend to make us jump into conclusions, known as *affect heuristics* [5]. There is thereby a tendency to let decisions dominate over the arguments on which they are made. Once a decision has been made, it is treated as correct and its flaws tend to become invisible to the human thought.

F. Status Quo Bias

Although quick jumps and shortcuts are preferred by confidence and ignorance in reasoning, humans have a reluctance to change current state of affairs, known as *status quo* bias [5]. It is caused by a tendency to apply more weight to losses caused by a change compared to its benefits. Many situations are consequently not changed even though there are better alter- natives, which is troublesome in the field of functional safety. For example, many pilots are used to control aircraft through mechanics and hydraulics and may oppose a change toward fly-by-wire systems even though such replacements could make their job safer, easier, and more environment-friendly. Furthermore, stakeholders, enterprises, engineers, authorities, etc., might be reluctant to change practices, processes, tools, cultures, codes of conduct, etc., even though it could result in safer systems at lower costs. On the other hand, if investments are made to make a change, humans are reluctant to abandon the idea in the process of achieving it in case the situation develops to the worse, known as the *sunk-cost fallacy* [5].

G. Biases by Subconscious Processes

Subjective decision making processes are also highly affected by subconscious processes, where acquired information has an effect even though the information is not consciously recognized, known as *the priming effect* [5]. Related to priming effect is *the anchor effect*, where a quantitative or qualitative value of some property subconsciously leads judgments in its direction, even when the value obviously is false.

V. COUNTERACTING FALLACIES

Common fallacies of risk perception support a judgment by domain experts as the most reliable and ethically justified method. Nevertheless, completely rational experts view the world in terms of numbers and logics, where important properties of harm in relation to emotions are easily excluded. Unpleasant emotions are harmful too, even if they correspond to an irrational response. A long-term, strong fear of an accident that never takes place, may be equally harmful as the accident itself. In addition, acceptance of fatalities, injuries, and environmental damages is judged differently depending on the context even if the physical damage is the same; fatalities to adults are viewed differently from fatalities to infants; fatalities by unintentional human errors are viewed differently from fatalities caused by malicious intentions or careless ignorance; accidents in extreme weather conditions are viewed differently from accidents in optimal conditions; damages to endangered species are viewed differently from damages to non-endangered species; and so forth. From this perspective, the public may be better than experts at morally weighing types of harm in terms of unacceptability according to Kahneman [5]. By these premises, the optimal moral method seems to be judgment by domain experts with an adjustment in the direction of the public opinion.

Both judgments by domain experts and by the public will include errors as discussed in section IV. Since the existence of these is known, it would be morally invalid to not take them into consideration in risk-related decision making processes. There is a number of methods to mitigate them. First of all, since errors exist in every individual judgment, it is important to de-correlate them in collective judgments such that they are suppressed rather than magnified (e.g. by groupthink and untrue anchors). Making sources of information in judgments independent from each other can achieve this.

The wisdom of crowds stems from this property [5]. Although each individual is poor in guessing, the average of a crowd tend to be accurate if the individuals (and guesses) are independent from each other. Some will guess too high and some too low, but the errors tend to cancel each other out if judgments are made independently.

Furthermore, extreme results are by nature more likely to be found in small samples compared to larger, known as *the law of small numbers* [5]. Essentially, whenever a phenomenon is studied, samples must be sufficiently large to be reliable. Violations of this law are common even in the scientific community, commonly due to a convincing intuition, often causing untruthful impressions and judgments [5]. Such principles are made central in clinical science and justice systems, e.g. witnesses are not allowed to interact before a testimony and a single witness is given little weight compared to several consistent, and we argue these also are important in decision making processes within functional safety. In addition, the halo effect is suppressed by making sources of evidence independent such that the quality of a property does not affect the impression of other properties.

To further rationalize decision making processes, principles that raise doubts and reduce overconfidence of stakeholders should be implemented within the processes. Two basic principles are to criticize the strongest objective arguments and to put focus on the weakest parts of subjective information [10]. The availability effect can be suppressed by forcing parties to provide additional arguments in favor of their judgments [10]. The more arguments that one must come up with the less intuitive they become, thus causing cognitive struggle and thereby a reduced confidence in their initial judgments. The same principle can be used to suppress hindsight bias in case of unexpected negative effects, incidents, and accidents. By forcing parties to list more scenarios through which an event could be avoided, the less confident they become in that it was avoidable. Moreover, when case specific information is available, such as from an accident, we tend to neglect the statistical base rates, i.e., humans are unwilling to deduce the particular from the general. In order to suppress this phenomenon, parties should be forced to recall the statistics when case specific information is presented.

An opposite problem of the availability effect is that humans tend to have difficulties with imagining something worse than what has been experienced [5]. The most catastrophic experience is most likely not the worst there can be, and there are probably many more scenarios through which these can take place. The importance of thinking outside of the box, i.e. *broad framing* of the problem, is therefore critical in process of making risk-related decision. The principle is to make decisions based on an analysis of a wide set of possible options rather than individual decisions on each option in isolation.

Since humans are reluctant to accept the random nature of the world, we tend to assign an illusory cause-and-effect-based explanation to a change toward the mean. Nevertheless, regression toward the mean is not based on such a phenomena, but rather on random errors in a natural distribution around the mean. In terms of liability of accidents, enterprises within safety-critical domains are thereby to some extent punished for being unlucky and rewarded for being lucky. Authorities, and the society as whole, on the other hand, are statistically being rewarded for punishing unlucky enterprises and punished for rewarding lucky ones. This behavior cannot be right conduct as it incorporates unjustified harm. This type of wrong doing can be mitigated by adjusting hindsight judgments according to regression to the mean principle.

VI. CONCLUSION

Functional safety of a system is defined as the part of its freedom from unacceptable/unreasonable risks that depends on a system operating correctly in response to its inputs. Its elements are analyzed at every stage of the software development life cycle, from requirement specification, design, implementation, verification, validation to deployment. Functional safety standards, such as ISO 262626, define safety as the absence of unreasonable risk.

In this paper, issues important to functional safety are analyzed together with common fallacies in risk perception in order to derive precautions for the involved risk-related decision making processes. In particular, the notion of an acceptable residual risk stipulated by functional safety standards is explored, including with respect to long-term negative effects of the technological evolution. We also address the governmental principles of *vision zero* and *zero tolerance* often applied to other societal safety issues. Kahneman's book "Thinking, Fast and Slow" [5] is used as foundation for the analysis of unreasonable risk judgments. We question if it is ethical to apply a view of definitive acceptance towards an amount (even if very small) of fatalities, injuries, and damages, and in that case, what degree of risk is morally acceptable and whether such decisions include systematic errors of thinking. We propose that, besides the existing requirement for the demonstration of risk prevention in the certification of safety critical systems, the requirement for justification of residual risks should be added in order to raise awareness of their existence and increase the probability of future mitigations.

Based on Kahneman's studies, the goals people set are highly important to what they do and feel about risk. By these premises, and under the assumption that the safety culture has the largest impact on the safety of critical systems as suggested in [13], an application of the *vision zero principle* and a larger focus on safety culture requirements within functional safety standards might be a more morally valid approach to the regulation of risk and possibly lead to safer system designs. Krause writes "*A culture that truly values ethical (and safe) behavior must be led by men and women committed to principle for its own sake, not solely for the purpose of compliance. Compliance alone does not require a deeper understanding, and without a deeper understanding, the ability to make functional safety safer is reduced.*" [23]. The view that safety culture based on openness, learning, adaptability and sharing of experiences is central for safety is supported by the study of the ethical aspects of cyber-physical technology [24-26], which also comprise hardware controlled by software. The authors emphasize the evolutionary character of technology, which is being improved iteratively and consecutively, because many of the phenomena in the real world applications are emergent and impossible to predict from the beginning. The constant improvements and sensitivity to safety issues are central and can only be upheld if the whole safety culture is built around them.

Since reasoning evidently is distorted by a focusing illusion and a reluctance to frame problems broadly, safety also relies on the ability of available tools and usable artifacts. Consequently, principles to avoid or suppress systematic errors of thinking described in this paper should be incorporated in safety standards. ISO 26262 specifies groupthink and exclusion of dissenters as examples of a poor safety culture; however, no additional fallacies or specific guidelines to prevent them are presented. Additional fallacies that may be threats to functional safety and possible countermeasures are presented in this paper. Regarding an application of the vision zero principle instead of an acceptance toward a predefined degree of residual risk, we argue it deserves to be studied in the domain of functional safety as it has the potential to significantly improve safety cultures and reduce negative long-term effects of the technological evolution, which otherwise may be ignored by a biased focus on short-term benefits. The current development of autonomous systems controlled by artificial intelligence makes an analysis of this change in attitude even more urgent.

ACKNOWLEDGMENTS

This research is supported by the Swedish Foundation for Strategic research (SSF) project SYNOPSIS – Safety Analysis for Predictable Software Intensive Systems – and the knowledge foundation (KK-stiftelsen) project DPAC – Dependable Platforms for Autonomous systems and Control.

REFERENCES

[1] A. Avižienis, J.-C. Laprie, and B. Randell, *Dependability and Its Threats: A Taxonomy*. Boston, MA: Springer US, 2004, pp. 91–120.

[2] International Organization for Standardization, “ISO 26262-1:2011 Road vehicles - Functional safety,” Geneva, Switzerland.

[3] R. Hugman, E. Pittaway, and L. Bartolomei, “When ’Do No Harm’ Is Not Enough: The Ethics of Research with Refugees and Other Vulnerable Groups,” British Journal of Social Work, 2011.

[4] T. Beauchamp, “The Principle of Beneficence in Applied Ethics,” in Stanford Encyclopedia of Philosophy, 2008.

[5] D. Kahneman, Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011.

[6] J. M. Doris, The Moral Psychology Handbook. Oxford University Press, 2010.

[7] A. Tversky and D. Kahneman, “Availability: A heuristic for judging frequency and probability,” Cognitive Psychology, vol. 5, no. 2, pp. 207 – 232, 1973.

[8] D. Kahneman and A. Tversky, “On the psychology of prediction,” Psychological Review, vol. 80, no. 4, pp. 237–251, 1973.

[9] A. Tversky and D. Kahneman, “Subjective probability: A judgment of representativeness,” Cognitive Psychology, vol. 3, no. 3, pp. 430 – 454, 1972.

[10] D. Kahneman and A. Tversky, “Belief in the law of small numbers,” Psychological Bulletin, pp. 105–110, 1971.

[11] K. Mykytyn, P. P. Mykytyn, and C. W. Slinkman, “Expert Systems: A Question of Liability?” MIS Q., vol. 14, no. 1, pp. 27–42, Mar. 1990.

[12] S. Dowlatshahi, “The role of product safety and liability in concurrent engineering,” Computers & Industrial Engineering, vol. 41, no. 2, pp. 187 – 209, 2001.

[13] H. Kienle, D. Sundmark, K. Lundqvist, and A. Johnsen, “Liability for Software in Safety-Critical Mechatronic Systems: An Industrial Questionnaire,” in Proceedings of the 2nd International Workshop on Software Engineering for Embedded Systems, June 2012.

[14] R. Sapru and R. Schuchard, “CSR and Quality: A Powerful and Untapped Connection,” The Global Voice of Quality, ASQ and BSR Quality Press, 2011.

[15] N. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety (Engineering Systems)*. The MIT Press, 2012.

[16] G. Ceballos, P. R. Ehrlich, A. D. Barnosky, A. Garcia, R. M. Pringle, and T. M. Palmer, “Accelerated modern human–induced species losses: Entering the sixth mass extinction,” Science Advances, vol. 1, no. 5, 2015.

[17] P. Kemp, *En teknolgietik*. Stockholm, Sweden: Brutus Östlings Bokförlag Symposion, 1991.

[18] S. Gardner, *Thinking Your Way to Freedom: A Guide to Owning Your Own Practical Reasoning*. Temple University Press, 2008.

[19] O. Svenson, “Are we all less risky and more skillful than our fellow drivers?” Acta Psychologica, vol. 47, no. 2, pp. 143 – 148, 1981.

[20] J. Kruger and D. Dunning, “Unskilled and Unaware of It: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self- Assessments,” Journal of Personality and Social Psychology, vol. 77, pp. 1121–1134, 1999.

[21] S. Epstein and V. Denes-Raj, “Conflict Between Intuitive and Rational Processing: When People Behave Against Their Better Judgment,” Journal of Personality and Social Psychology, 1994.

[22] A. Tversky and D. Kahneman, “Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment,” Psychological Review, pp. 293–315, 1983.

[23] T. Krause, “The Ethics of Safety,” http://ehstoday.com/safety/best-practices/ehs_imp_67392, October 2016.

[24] G. Dodig Crnkovic and B. Çürüklü “Robots: ethical by design,” Ethics and Information Technology, vol. 14, no. 1, pp. 61–71, 2012.

[25] A. Thekkilakattil and G. Dodig-Crnkovic, “Ethics Aspects of Embedded and Cyber-Physical Systems“ In: IEEE Proceedings of COMPSAC 2015: The 39th Annual International Computers, Software & Applications Conference, Symposium on Embedded & Cyber-Physical Environments (ECPE). Taichung, Taiwan - July 1-5, pp. 39-44, 2015. DOI: 10.1109/COMPSAC.2015.41

[26] G. Sapienza, G. Dodig-Crnkovic, and I. Crnkovic, “Inclusion of Ethical Aspects in Multi-Criteria Decision Analysis.” Proc. WICSA and CompArch conference. Decision Making in Software ARCHitecture (MARCH), 2016 1st International Workshop. Venice April 5-8 2016. DOI: 10.1109/MARCH.2016.5, ISBN: 978-1-5090-2573-2. IEEE