Toward Risk Frameworks for Autonomous Systems that Take Societal Safety-related Benefits into Account

Ellen J. Bass, Steven Weber
Drexel University
ellen.j.bass@drexel.edu, steven.weber@drexel.edu

Abstract

Current risk frameworks such as probabilistic risk analysis methodologies do not take societal safety-related benefits into account. To inform human-AI collaborative system development, this manuscript highlights the need for updated risk frameworks and suggestions for relevant considerations.

Introduction

Systems with embedded machine learning and artificial intelligence (AI) technologies may enhance performance and increase safety. For operations with small robots and drones where failures do not produce major damage, the public is likely to accept higher levels of risk due to the reduction of safety risks to human workers. However, current risk frameworks such as probabilistic risk analysis (PRA) methodologies do not take such societal safety-related benefits into account. This manuscript highlights the need for updated risk frameworks to inform human-AI collaborative systems.

A Safety Management System is a top-down approach to managing safety risk and assuring the effectiveness of safety risk controls (Federal Aviation Administration (FAA) 2020). The main components are safety policy, safety risk management (SRM), safety assurance, and safety promotion. SRM is composed of system analysis including developing a safety risk acceptance plan; identifying hazards; and analyzing, assessing, and controlling safety risk (FAA 2023). Understanding safety risk components requires examining factors that increase or decrease the likelihood of events that can result in unwanted accidents or incidents. Analysts assess each hazard's associated safety risk against risk acceptance criteria from the safety risk acceptance plan to determine the acceptable safety risk level (FAA 2023).

Typically, risk is determined by severity and likelihood. Severity is the potential consequence or impact of a hazard (degree of loss or harm). Typical questions include:

- What are the credible outcomes (i.e., catastrophic, hazardous, major, minor, minimal)?
- Why (e.g., data, expertise, rationale for how the safety analyst or team arrived at their determination)?
- How do existing controls and additional mitigations change the technology, the operators, and/or operating environment, such that the severity is reduced?

Likelihood is the estimated probability or frequency, in quantitative or qualitative terms, of the outcome(s) associated with a hazard. Questions to consider include:

- What is the likelihood of credible outcomes? (e.g., frequent, probable, remote, extremely remote)
- Why? (e.g., data, expertise, rationale for how the safety analyst or team arrived at their determination)
- How do mitigations change technology, operators, and/or operating environment, so that the likelihood is reduced?

Unfortunately, these types of questions ignore alternative solutions and their potential consequences. Questions related to what risks can be avoided if a technology is used should be as important as the risks introduced by a technology. This situation requires considering alternative methods.

Alternative Risk Assessment Methods

When operations can increase safety with respect to societal need, societal benefit should be considered as part of risk analyses. Alternative methods for addressing risk should be investigated. Such strategies are consistent with a National Academies of Sciences, Engineering, and Medicine (NASEM) report focusing on the integration of remotely supervised aircraft and related questions (NASEM 2018):

- What are the benefits and limitations of alternative risk assessment methods? How do these alternative methods compare to PRA methods as well as severity and the probability metrics traditionally used?
- What state-of-the-art assessment methods are currently in use by industry, academia, and other organizations?
- What are key challenges or barriers to overcome to implement the recommended risk assessment methods?

The 2018 NASEM report concluded that the public is likely to accept for domains, such as small unmanned aircraft systems (UAS), what is similar in the context of levels of *de minimis* risk for other levels of societal activities. *De* *minimis* risk is useful in establishing safety standards for missions such as with small UAS. Current PRA methodologies do not take societal safety-related benefits into account. Missions such as those with small UAS operations can increase safety with respect to societal need and thus such societal benefit should be considered as part of the analyses. Also, in some cases industry should be responsible for quantitative risk assessment activities or should be able to obtain insurance in lieu of having a separate risk analysis.

As a snapshot, the 2018 NASEM report highlighted:

- Consider broader societal benefits in addition to risk when conducting safety assessments.
- Do not simply treat risk as the single probability: consider risk as a multivariate measure.
- Performance requirements should be commensurate with risk and backed by performance-based standards.
- Consider new institutional mechanisms for conducting, or delegating, risk analysis.

The 2018 NASEM report suggested that:

- Rules, regulations, and restrictions should be commensurate with the risk posed by specific operations.
- Potential safety risks of operations in domains such as robotics and drones primarily include collisions with other aircraft and injury to people and property on the ground. Operations can reduce safety risks by replacing activities that put people at risk with lower risk ones.

Thus, improved measures for assessing risk could be considered such as applicable to background risks that people experience daily. Concepts from *de minimis* risk should inform the process of assessing acceptable levels of risk posed by such technology, especially with measurable and known economic and safety benefits to society (e.g., inspection of critical infrastructure that pose tangible danger to human inspectors, humanitarian delivery of medicines and other lifesaving cargo to rural areas or hard to reach areas, emergency response, and agricultural sensing leading to reduction in use of pesticides and other chemicals). These benefits may outweigh any risks added by autonomous system operations.

Opportunities to increase the safety of operations through increased autonomy should not be missed due to a lack of accepted risk assessment methods. In addition, risk characterization should include multivariate measures with co-variates such as the mission type, characteristics of the autonomous system (e.g., weight) and other environment variables.

Concerns related to the teaming of humans and machines can be reflected in the risk analysis methods. No broad-brush statements can be reliably made about the role of the human and machine technologies. Instead, those design variables that determine system sensitivity to likely machine failures, and to foreseeable inadvertent slips and mistakes by humans, can be accounted for within each system. Further, risk analysis, by examining how the human-machine team interacts, can better capture how the autonomous system will detect and resolve hazards arising within the team. Such a risk analysis would determine the extent to which humans and machine technologies are able to coordinate to resolve hazards arising in the broader operational environment.

Accepting risk is far easier when the risk is well quantified by relevant empirical data. Uncertain risk does not equate to high risk, however. By accepting the uncertain risk associated with a new technology, with reasonable mitigations, one can obtain the data needed to better quantify that risk. As the uncertainty diminishes, one can remove or augment the mitigations as appropriate.

Integration of sensors and analytics present an opportunity to learn and test new models for better data collection and analysis with the aim of improving overall safety. When computational models are used, model prediction uncertainties are not always being calculated and no distinction is being made to distinguish between uncertainties due to lack of knowledge and those due to natural variability of the data.

Recommendations

What follows are some recommendations demonstrating quantitative risk assessment and other strategies.

- Where operational data are insufficient to credibly estimate likelihood and severity components of risk, the analysts should use a comparative risk analysis approach to compare proposed operations to comparable existing or *de minimis* levels of risk.
- Risk level and risk mitigation strategies should consider not only proximal ones but also third-party risks.
- Benefits to society of the new technology need to be explicitly defined, and when possible, quantified. Such definitions can support systematic and purposeful trade offs.
- Analysts should identify classes of operations where the level of additional risk is expected to be so low that it is appropriate to base approval on requiring insurance in lieu of having a separate risk analysis.

References

Federal Aviation Administration 2020. Order 8000.369C - Safety Management Systems, accessed 15 March, 2024. https://www.faa.gov/documentLibrary/media/Order/Order_8000.369C.pdf

Federal Aviation Administration 2023. Order 8040.4C - Safety Risk Management Policy, accessed 15 March, 2024. https://www.faa.gov/documentLibrary/media/Order/FAA_Order_8040.4C.pdf

National Academies of Sciences, Engineering, and Medicine 2018. Assessing the Risks of Integrating Unmanned Aircraft Systems (UAS) into the National Airspace System. Washington, DC: The National Academies Press.