THE RISK OF TRUST IN SAFETY: CONFIDENCE GAMES

Massimo Felici

School of Informatics, The University of Edinburgh, Edinburgh EH9 3JZ, UK
http://homepages.inf.ed.ac.uk/mfelici/

Keywords: Arguments, Risk, Confidence, Trust, Games.

Abstract

This paper investigates the relationship between confidence, trust and risk. It proposes a game theoretical representation of confidence (trust). Confidence games explain to some extent similarities (differences) between confidence and trust. Moreover, the link between confidence and trust, in terms of games, allows a characterization of risk with respect to confidence.

1 Introduction

This paper is concerned with understanding the relationship between *confidence*, *trust* and *risk*. Research and practice in safety-critical systems emphasize the relationship between safety and risk. The understanding of the relationship between safety and risk allowed the development of risk assessment and management methodologies and their integration in industry standards (e.g., IEC 61508, etc.), concepts (e.g., ALARP, etc.) and practices (e.g., certification, construction of safety cases, etc.). Unfortunately, despite the progress in understanding the relationship between safety and risk, there is often a lack of *confidence* in safety argumentations – *How to trust system safety? How much trust in safety?* The relationship between trust and safety has been investigated, to a certain extent, in those application domains (e.g., ATM) in which it appears how a lack of trust (or misplaced trust in automation) affects overall safety (performances). This is the case in which safety-critical systems (functionalities) support monitoring (human) activities. Intuitively, a lack of trust exposes organizations to reduced (safety) performances as well as to an increased risk of *failures*. Therefore, it is necessary further to investigate the relationship between trust and risk, hence, understanding about how confidence, trust and risk relate each other.

This paper draws on relevant literature reporting some *counterintuitive* situations [1][2][9]. For instance, there could be cases in which safety arguments need to change, or evolve, due to emerging knowledge (e.g., safety issues) [4][5][6]. Moreover, subtle contingencies emerge in structuring safety cases. On the other hand, emerging knowledge, or *unforeseen* interactions (e.g., system usages, system interactions, system dependencies, etc.), may controversially reduce our confidence in safety argumentations. These situations highlight contingencies in the construction of safety arguments. This paper investigates the opportunity to characterize construction processes of safety arguments as *trust games*. Trust games allow a characterization of the relationship between risk and trust. The underlying idea is to devise games that support the decision-making involved in construction processes of safety arguments. *How does emerging knowledge affect trust?* This paper investigates the relationship between risk and trust. It questions processes of arguing safety. It analyzes how trust games capture safety argumentation processes. It analyzes similarities between confidence and trust. The identified similarities allow the extensions of trust games to confidence too, hence, *confidence games*.

This paper argues a link between confidence and trust. Probabilistically, confidence and trust behave similarly. This clarifies a relationship between confidence and trust, although it exposes the limitations of probabilistic confidence. On the other hand, it provides a characterization of risk in confidence games. The paper is structured as follows. Section 2 briefly describes the results of relevant work, which defines and characterizes confidence in diverse arguments. Section 3 draws the similarities between confidence and trust. It defines and introduces confidence games. Section 4 draws some conclusions.

2 Confidence in Diverse Arguments

This section draws over relevant work on confidence in diverse arguments [1][2][9]. The work in [1] points out contingencies in the compositions of diverse arguments. The combination of diverse arguments, e.g., a testing leg argument and a verification leg argument, intuitively, intends to provide a stronger support for the overall claim (e.g., reliability, safety or dependability claim) than the individual arguments themselves. The more supporting arguments, the stronger is the claim. Unfortunately, the combination of diverse arguments could give rise to unforeseen interactions (e.g., dependencies) between the arguments. In practice, there are cases for which the combinations of diverse arguments are problematic. Issues arising may affect the overall confidence in the arguments or the claims. The statistical interpretation of confidence is *"the probability that a claim is true"* [9]. The study in [9] of the argument structures highlights counterintuitive cases for which confidence decreases despite supporting arguments.

A BBN (Bayesian Belif Network) model topology in [9] captures the structure of the combined two-legged arguments. The model for the two-legged arguments consists of six variables: *S*, the system's unknown true probability of failure on demand (*pfd*); *Z*, the system specification's correctness; *V*, the verification outcome; *O*, the testing oracle's correctness; *T*, the system test results; and *C*, the acceptance of the final claim. The BBN model allows the theoretical probabilistic analysis of *conditional independences* among the statistical variables. The model, taking into account some simplifying and conservative assumptions, allows the statistical characterization of *confidence* and *doubt* (as complementary to confidence) functions, respectively. The expressions for confidence and doubt depend on the *unconditional joint distribution* between *Z*, the correctness' specification, and *O*, the correctness' oracle. Therefore, the analytical simulations of the functions depend on the correlation (i.e., covariance) between *Z* and *O*. The simulations of the doubt functions, i.e., the doubt functions for the two-legged arguments and the single arguments, highlight different behaviours with respect to the correlation. The study in [9] observes that:

- The correlation between *Z* and *O* is irrelevant for the verification-only argument. This results in a constant doubt over an increasing correlation. This is due to the fact that the correlation between *Z* and *O* is irrelevant for the verification argument. It only depends on the prior marginal distribution of *Z*.
- The unconditional joint distribution affects the level of doubt. In particular, the doubt increases (the confidence decreases, respectively) as the prior marginal probabilities of the correctness of *O* and *Z* decrease.
- The convenience of the different arguments (i.e., two-legged or single arguments) depends on the correlation between *Z* and *O*. The correlation affects the ordering of the different doubt functions.

The analytical model, moreover, points out counterexamples for both single arguments and two-legged arguments [9]. In particular, there are cases for which it is possible to obtain non-supportive testing and verification arguments, respectively. The former corresponds to those cases for which an increasing number of fault-free testing increases the doubt on the oracle's correctness. The latter corresponds to a strong believe that the specification is incorrect – "*many failure-free test cases […] increase our mistrust in the oracle, increase our mistrust in the specification, and thus increase our mistrust in the pfd"* [9]. Similarly, under conservative assumptions for the verification argument, the successful verification provides "*stronger support for the incorrectness than the correctness of the specification Z"* [9].

The more complex and intriguing cases than the single arguments are those for which adding further evidence (that is, adding a supportive leg) produces a decrease in confidence. For instance, in some cases, adding a supporting verification argument to a testing one may decrease our confidence over the (dependability) claim. This is due to the belief that the specification is incorrect. Hence, the testing oracle is incorrect too – "*The verification leg is supportive when we have no testing evidence […] testing can undermine the contribution that the verification leg makes to the overall confidence in the dependability claim when both argument legs are present"* [9]. The results about confidence in multi-legged arguments stress the need to understand subtle interaction (or dependency) mechanisms between diverse arguments. For instance, it is necessary further to understand both the underlying argument structures and processes. The understating of both argument structures and construction processes stresses strategies and policies for dependability analyses [3][8].

3 Confidence Games

This section analyzes similarities between confidence, as the probability that a claim is true, and trust. A review of trust highlights diverse accounts [3]. Confidence and trust are complex concepts, which exhibit subtle contingencies. Intuitively, confidence and trust are similar concepts and exhibit similar behaviours. For instance, both confidence and trust relate to knowledge, although they may have different relationships. They change over time due to arising knowledge (e.g., further evidence or new information). Unfortunately, their relationships have been yet little investigated and understood. This section takes into account a game theoretical characterization of the notion of trust. This characterization points out, to some extent, similarities between confidence and trust. The underlying problem is whether it is possible to find an alternative representation of confidence (trust) that allows a characterization of the relationship between confidence and trust.

Trust has been extensively studied in diverse domains [3]. Intuitively, trust is a relationship between different entities or peers. This relationship involves diverse interactions (e.g., interactions between peers, systems, humans, etc.), which affect trust over time according to previous experiences (e.g., interactions, cooperation behaviours, etc.). It is often easy to lose trust in something or someone and takes time to regain or re-establish trust. Theoretical games are common characterizations of interactions. For instance, *Prisoner's Dilemma* (PD) games have been extensively used in various domains (in particular, finance) for the characterization of trust. In order to overcome some practical limitations (e.g., knowledge distribution, risk perception, etc.), trust games provide a better characterization of trust than classical PD games [4].

In order to highlight similarities between confidence and trust, this section introduces a game characterization of trust (confidence). The analytical model in [9] highlights how confidence (doubt) functions depend on the correlation between the correctness of the specification *Z* and the testing oracle *O*. It is possible to highlight similar results in terms of (trust) games. This section preliminary defines a theoretical game between the specification *Z* and the oracle *O*. Figure 1 shows a sample payoff matrix for a game between the specification *Z* and the oracle *O*. The matrix captures the combination of all possible cases of Oracle and Specification being *Correct* (C) or *Incorrect* (I), respectively. Correct and Incorrect correspond to *Collaborate* and *Defeat* in PD (trust) games, respectively. This correspondence is possible because

of the underlying assumptions in the analytical characterization of confidence [9].

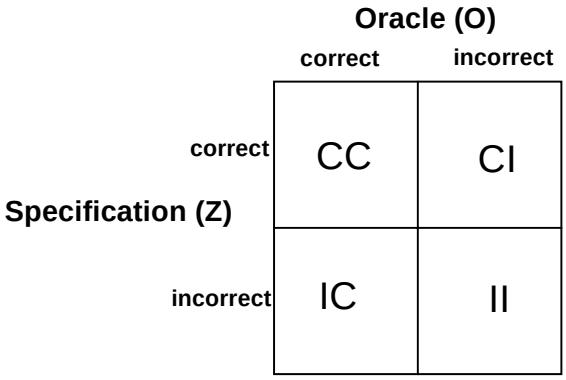

Figure 1: Sample Payoff Matrix.

The correlation between *Z* and *O* depends on the distributional assumptions, in particular, on the unconditional joint distribution of the different possible cases [9]. Similarly, the overall correlation between the correctness of the oracle and the specification depends on the history of previous games. This is the case for iterated games. It is possible to represent iterated games in terms of the previous instances of the games as decision trees. Figure 2 shows a sample decision tree.

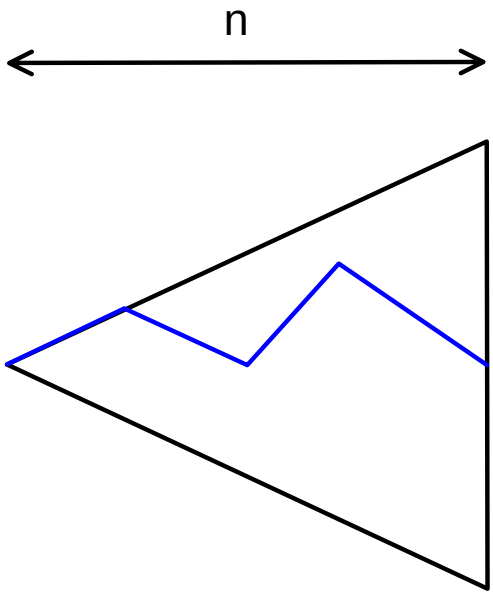

Figure 2: A Decision Tree.

A decision tree captures the *n* interactions (or iterations) between the players, that is, the specification Z and the oracle O. The number of games (interactions) corresponds (at least) to the number of test cases. The study of the doubt functions with respect to the number of test cases corresponds to a game having the same number of iterations. The *unconditional joint distribution* represents the history of the decisions (interactions) between the specification and the oracle. Note that asynchronous games capture also single arguments, that is, those cases that consider only one argument. This corresponds to asynchronous games in which a player decides to withdraw (neither is correct nor incorrect). This stresses the similarity with trust games rather than classical PD games.

Trust is then the cumulative sum of the outcome of each game. Although, the obtained results provide an unbounded index, it is possible to normalize it. Moreover, the important aspect is the monitoring of the index trend (increasing or decreasing). The trend depends of course on the payoffs (which could be negative in trust games). Calibrating the game for a specific domain requires us to identify how oracle and specification' correctness interact each other. It is possible to figure out how confidence games capture those counterintuitive cases of decreasing confidence (i.e., negative payoffs for some combinations). The game theoretical representation of confidence (trust) allows an intuitive characterization of hazard and risk for a decrease in confidence (trust) in dependability claims. The hazards correspond to those sequences (combinations) of events (decisions) that lead to a decreasing level of confidence. The risk is, then, related to the likelihood of these cases. That is, the ratio between the number of cases in which the confidence decreases and the number of all possible combinations. Note that the ratio could be easily calculated by taking into account all possible paths in a decision tree. The initial confidence level depends on domain observations, similarly to the calibrations of probabilistic models.

4 Conclusions

This paper reviews relevant work about confidence in diverse arguments. The paper summarizes and draws over some *counterintuitive* examples for which confidence in diverse arguments decreases. It is possible to characterize confidence (trust) in terms of theoretical games. This characterization allows us to identify a similarity relationship between trust and the probabilistic account of confidence. For such cases, confidence and trust behave similarly. Hence, *the probabilistic account of confidence corresponds to a set of trust games*. On the one hand, trust captures to some extent confidence. On the other hand, confidence falls short to capture trust. The benefit of the proposed game theoretical representation of confidence (trust) is twofold. First, it captures processes of arguing, combining and constructing claims. Second, it allows a characterization of the relationship between confidence, trust and risk. Future work intends to formalize (in terms of rules) and specify (in terms of payoff matrixes) confidence (trust) games. However, the preliminary characterization of confidence (trust) games allows an explanation of the similarities (differences) between confidence and trust.

Acknowledgements

This work was partially supported by the "Interdisciplinary Design and Evaluation of Dependability" (INDEED) project funded by the UK Engineering and Physical Research Council (EPSRC), Grant EP/E001297/1.

References

[1] R. Bloomfield, B. Littlewood, “Multi-legged arguments: the impact of diversity upon confidence in dependability arguments”, Proceedings of the 2003 International Conference on Dependable Systems and Networks (DSN’03), pp. 25-34, (2003).

[2] R. E. Bloomfield, B. Littlewood, D. Wright, “Confidence: its role in dependability cases for risk assessment”, Proceedings of the 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2007), pp. 338-346, (2007).

[3] S. Bottitta, M. Felici. “Understanding and learning trust: A review, characterization and tool”, C. Guedes Soares, and E. Zio (Eds.), Safety and Reliability for Managing Risk, Proceedings of the European Safety and Reliability Conference 2006 (ESREL 2006), Taylor & Francis Group, Volume 2, pp. 1273-1280, (2006).

[4] M. Felici, “Evolutionary Safety Analysis: Motivations from the Air Traffic Management Domain”, R. Winther, B.A. Gran, and G. Dahll (Eds.), Proceedings of the 24th International Conference on Computer Safety, Reliability and Security, SAFECOMP 2005, LNCS 3688, Springer-Verlag, pp. 208-221, (2005).

[5] M. Felici, “Modeling Safety Case Evolution - Examples from the Air Traffic Management Domain”, N. Guelfi and A. Savidis (Eds.), Proceedings of the 2nd International Workshop on Rapid Integration of Software Engineering techniques, RISE 2005, LNCS 3943, Springer-Verlag, pp. 81-96, (2006).

[6] M. Felici, “Capturing emerging complex interactions: Safety analysis in air traffic management”, Reliability Engineering & System Safety (RESS), Elsevier, Volume 91, Issue 12, pp. 1482-1493, (2006).

[7] M. Felici, “Trust Strategies and Policies in Complex Socio-technical Safety-Critical Domains: An Analysis of the Air Traffic Management Domain”, N. Guelfi and D. Buchs (Eds.), Proceedings of the 3rd International Workshop on Rapid Integration of Software Engineering techniques, RISE 2006, LNCS 4401, Springer-Verlag 2007, pp. 51–65, (2007).

[8] M. Felici, “Trust strategies: Motivations from the Air Traffic Management domain”, C. Guedes Soares, and E. Zio (Eds.), Safety and Reliability for Managing Risk, Proceedings of the European Safety and Reliability Conference 2006 (ESREL 2006), Taylor & Francis Group, Volume 3, pp. 1797-1804, (2006).

[9] B. Littlewood, D. Wright. “The Use of Multilegged Arguments to Increase Confidence in Safety Claims for Software-Based Systems: A Study Based on a BBN Analysis of an Idealized Example”, *IEEE Transactions on Software Engineering*, **33**, pp. 347-365, (2007).