Towards an Ethical Security Platform

Scott Cadzow(✉)

Cadzow Communications Consulting Ltd., Sawbridgeworth, UK
scott@cadzow.com

Abstract. The world of **I**nformation & **C**ommunications **T**echnology (**ICT**) security has been dominated by the **C**onfidentiality **I**ntegrity **A**vailability (**CIA**) paradigm for several decades now and has been very effective in countering relatively simple document based security threats of masquerade, exposure of confidential data, and verification of integrity. Unfortunately real world security problems are not discrete or document based but are complex multi-domain, multi-value ones. In such environments the conventional CIA paradigm is no longer the ideal fit and in particular as we become more reliant on ICT for living support then hard security in the context of CIA needs to be reconsidered. This means taking into account issues that are traditionally "soft" such as Ethics and Dignity and making them "hard" and developing solutions that allow us to treat them. Our starting position is that humans design, operate and are the net beneficiaries of most systems. However humans are fallible and make mistakes. At the same time humans are adaptable and resourceful in both designing systems and correcting them when they go wrong. In contrast machines have in the main been designed to follow rules and are often constrained to produce the same output for the same input over and over again. As we move towards autonomous and intelligent machines the older models of ICT and ICT security based on the CIA paradigm, or deterministic code execution become more and more challenged. Into this mix we then bring a requirement for making ethical decisions.

Keywords: Security · Safety · Ethics · Artificial Intelligence
Machine Learning

1 Introduction - Why Ethics?

Ethical decisions require that different outputs arise from apparently identical inputs as the wider context for the decision has changed. Adaptive machines already appear to have made the switch from deterministic code and the rise of Artificial Intelligence will hasten this switch. The primary concern we ought to have in the long term of AI and M2M is that whilst humans make ethical decisions almost automatically as we move towards an increasingly machine led society those aspects of dignity, ethics and security which are managed by humans will be addressed by machines. The aim of this paper is to give an overview of the state of the art in security standardisation in machine to machine and IoT systems, for the use cases of eHealth and autonomous transport systems, in order to outline the new ethics and security challenges of the machine led

M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 523–533, 2018.
https://doi.org/10.1007/978-3-319-91238-7_41

society. This will consider progress being made in standards towards the ideal of each of a Secure and Privacy Preserving Turing Machine and of an Ethical Turing Machine.

2 Human Fallibility

We start with a simple assertion: Humans design, operate and are the net beneficiaries of most systems. We can also assert as a consequence that humans are the net losers when systems go wrong. If that failure is in the security systems trust in the system can disappear.

Humans are fallible and make mistakes. It is also essential to recognise that humans are adaptable and resourceful in both designing systems and correcting them when they go wrong. These characteristics mean that humans can be both the strongest and the weakest link in system security. It also means that there is an incentive to manage the human element in systems such that those systems work well (functionality matches the requirement), efficiently (don't overuse resources), safely and securely. Thus human centric design, even for mostly machine based systems, is essential. However as we adopt more and more machines as human proxies we need to provide some of our intelligence into the systems but in doing so we need to be aware that the intelligence we are offering in machine systems will be different from that of humans. In part this is because the intelligence is by its nature artificial. The consequence may be that we actually design in fallibility.

The purpose of this paper is to mark those elements of the connected world and the publicised attacks on it, and to identify steps that security engineers should be taking to minimise the concerns raised. Addressing the fear of the threat model, promoting why good design works, relegating the "movie plot" threats to the fiction they belong in. The existing security design paradigms are those of "Secure by default" and "Privacy by design". It is not suggested that either of these paradigms is complete and that every product is both secure by default, and privacy protecting (privacy preserving) by design, however even when privacy is protected and security is assured the need for systems to act ethically and to treat their affected users with dignity needs to be assured too. The role of ethics—doing the right thing—in design is not yet clear as it is also not clear in real life. However as more and more decision making is moved into the machine world the need for machines and systems of machines to make the right decision is going to arise more and more. The consideration of dignity is perhaps even harder to quantify but again in machines interacting with humans there is often a need to treat the recipient with a certain degree of dignity, and furthermore to allow the human actor to hold their dignity intact.

The path to Artificial Intelligence, and to Machine Learning (the implied role of both is to gain wisdom in the use of data), is obviously key in this mode of development. The Ethics problem lies at the very top of the data transformation tree shown in Fig. 1. However much of our technology lies at the very lowest layers (networking, databases or data collections, some use of the semantic web).

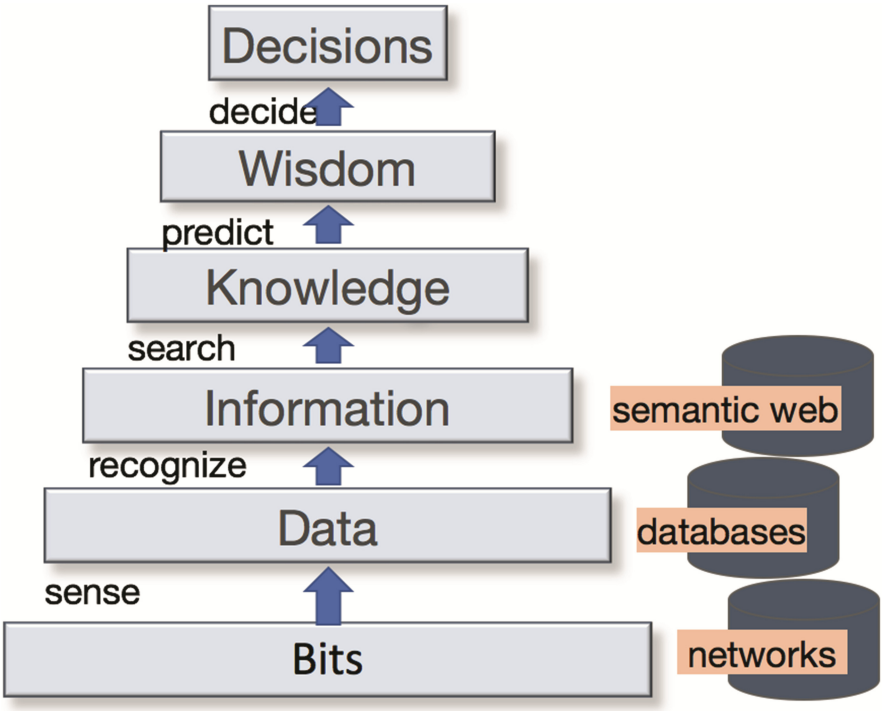

Fig. 1. Data transformation hierarchy from bits to decision

In terms of information processing experience it may be suggested that for the bulk of systems we lie somewhat to the left of the expertise scale (i.e., closer to Random decision making than at Expert decision making) (Fig. 2).

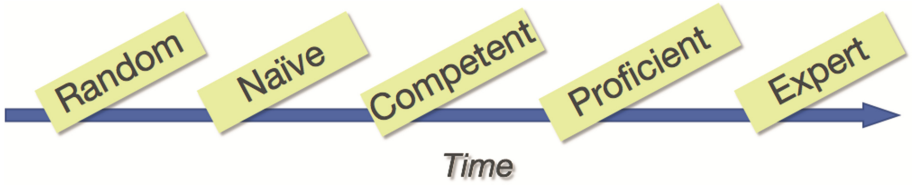

Fig. 2. The time dimension of learning

In looking to use cases there are two very obvious areas where machine ethics will be critical. In the domain of Intelligent Transport Systems (ITS) the operation of semi and fully autonomous vehicles will be increasingly divorced from human control, thus at the point when a crash is inevitable the vehicle has to be able to react in a way that minimises injury to both the occupants and to anyone or anything in the local area. This introduces the classical trolley problem at the ethical decision point: one decision may kill n people, the alternative will definitely kill 1 person. What is the right choice to

make? Obviously neither answer is right. There is no rationale for the vehicle to disavow itself of all responsibility and pass control to the local human so the ethical processing has to be built into the machine. In practice the existence of ethical decision points are most visible in hindsight as the consequences of decisions are often out of sight at the decision point.

The second critical domain is that of health—the classical source of the Hippocratic Oath which is often simplified down to "do not harm".

3 Intelligent Gaming and Game Theory in Machine Ethics

Ethical decisions are often both time critical and time variant. What is "right" in one context may be "wrong" in another context, where context may include the players, the time, the location or any other variable? An ethical problem often needs solved at the time it arises—there can be no delay without the problem resolving itself or any solution being invalid. Thus the trolley problem is one oft cited example (see Fig. 3).

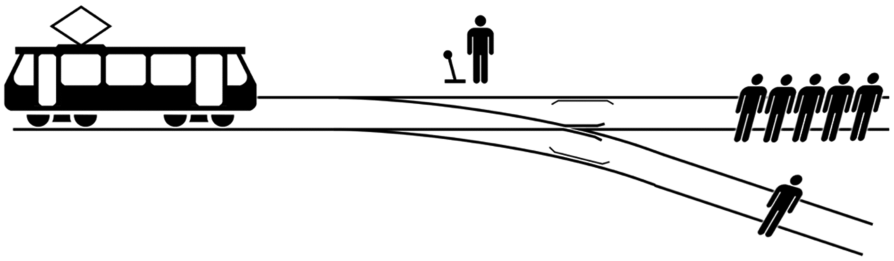

Fig. 3. The Trolley Dilemma (from McGeddon - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=52237245)

The problem is often phrased as a runaway train carriage at speed whilst ahead, on the track, there are five people tied up and unable to move. The train is headed straight for them. You are standing some distance off in the train yard, next to a lever controlled junction. If you pull this lever, the train will switch to a different set of tracks. However, you notice that there is one person on the side track. You have two options:

1. Do nothing, and the trolley kills the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person.

Which is the most ethical choice? If the choice is to be made by a machine how is the machine programmed? There is no correct choice of course and that is a problem of ethics—the right answer is almost wholly contextual and the deciding actor has limited perspective so can only see the 5 versus 1 conundrum. It is kind of assumed that all alternative avenues have either been tried and failed or are simply not available. How do you win and kill nobody? You can't without changing the problem and modifying the ethical argument.

An alternative view is that presented by the classical prisoner's dilemma but for the general case of co-operation. In moving away from the binary choice in the trolley

dilemma the number of actors involved can be expanded such that actors can collude to define the ethically preferable outcome. In the trolley dilemma for example can the trolley itself become involved in the decision? Can it take actions that alter the set of possible outcomes? If we take the prisoner's dilemma where the temptation payoff (T) is greater than the Reward payoff (R) which is greater than the Sucker payoff (S) and which is greater than the Punishment payoff (P) we want to be able to get the actors to work in such a way that with or without collusion they always choose to receive R on the assumption that mutually beneficial strategies are better over the long term.

Game theory is suggested as one way in which ethical issues can be considered. However in order to make such tools work effectively there are a number of pre-conditions that need to be met. The assertion of this paper is that many of the pre-conditions require a commitment to standards to assure interoperability and this is explored more below. It is further contended that the AI that will underpin real time application of game theory itself needs to be standardised at least at the level at which AI systems can interconnect.

4 The Role of Standards

Standards are at the root of sharing a common syntactical and semantic understanding of our world. This is as true for security as it is for any other domain and has to be embraced.

The more flexible a device is the more likely it is to be attacked by exploiting its flexibility. We can also assert that the less flexible a device is it is less able to react to a threat by allowing itself to be modified.

The use of the Johari Window [JOHARI] to identify issues is of interest here (using the phrasing of Rumsfeld).

	Known to self	Not known to self
Known to others	Known knowns - BOX 1	Unknown knowns - BOX 2
Not known to others	Known unknowns - BOX 3	Unknown unknowns - BOX 4

The human problem is that the final window, the unknown unknowns, is the one that gives rise to most fear but it is the one that is not reasonable (see movie plot threats below). The target of security designers is to maximise the size of box 1 and to minimise the relative size of each of box 2 and box 3. In so doing the scope for box 4 to be of unrestrained size is hopefully minimised (it can never be of zero size).

We can consider the effect of each "box" on the spread of fear:

BOX 1: Knowledge of an attack is public knowledge and resources can be brought to bear to counter the fear by determining an effective countermeasure

BOX 2: The outside world is aware of a vulnerability in your system and will distrust any claim you make if you do not address this blind spot

BOX 3: The outside world is unaware of your knowledge and cannot make a reasonable assessment of the impact of any attack in this domain and the countermeasures applied to counter it
BOX 4: The stuff you can do nothing about as e.g., far as you know nothing exists here.

In the security domain we can achieve our goals both technically and procedurally. This also has to be backed up by a series of non-system deterrents that may include the criminalisation under law of the attack and a sufficient judiciary penalty (e.g., interment, financial penalty) with adequate law enforcement resources to capture and prosecute the perpetrator. This also requires proper identification of the perpetrator as traditionally security is considered as attacked by *threat agents,* entities that adversely act on the system. However in many cases there is a need distinguish between the threat source and the threat actor even if the end result in terms of technical countermeasures will be much the same, although some aspects of policy and access to non-system deterrents will differ. A *threat source* is a person or organisation that desires to breach security and ultimately will benefit from a compromise in some way (e.g., nation state, criminal organisation, activist) and who is in a position to recruit, influence or coerce a threat actor to mount an attack on their behalf. A *Threat Actor* is a person, or group of persons, who actually performs the attack (e.g., hackers, script kiddy, insider (e.g., employee), physical intruders). In using botnets of course the coerced actor is a machine and its recruiter may itself be machine. This requires a great deal of work to eliminate the innocent threat actor and to determine the threat source.

The technical domain of security is often described in terms of the CIA paradigm (Confidentiality Integrity Availability) wherein security capabilities are selected from the CIA paradigm to counter risk to the system from a number of forms of cyber attack. The common model is to consider security in broad terms as determination of the triplet {threat, security-dimension, countermeasure} leading to a triple such as {interception, confidentiality, encryption} being formed. The threat in this example being interception which risks the confidentiality of communication, and to which the recommended countermeasure (protection measure) is encryption.

The very broad view is thus that security functions are there to protect user content from eavesdropping (using encryption) and networks from fraud (authentication and key management services to prevent masquerade and manipulation attacks). What security standards cannot do is give a guarantee of safety, or give assurance of the more ephemeral definitions of security that dwell on human emotional responses to being free from harm. Technical security measures give hard and fast assurance that, for example, the contents of an encrypted file cannot, ever, be seen by somebody without the key to decrypt it. So just as you don't lock your house then hang the key next to the door in open view you have to take precautions to prevent the key getting into the wrong hands. The French mathematician Kerchoff has stated "A crypto system should be secure even if everything about the system, except the key, is public knowledge". In very crude terms the mathematics of security, cryptography, provides us with a complicated set of locks and just as in choosing where to lock up a building or a car we need to apply locks to a technical system with the same degree of care. Quite simply we don't need to bother installing a lock on door if we have an open window next to it—the attacker will ignore

the locked door and enter the house through the open window. Similarly for a cyber system if crypto locks are put in the wrong place the attacker will bypass them.

It may be argued that common sense has to apply in security planning but the problem is that often common sense is inhibited by unrealistic threats such as the movie plot scenarios discussed below.

Standards are peer reviewed and have a primary role in giving assurance of interoperability. Opening up the threat model and the threats you anticipate, moving everything you can into box 1, in a format that is readily exchangeable and understandable is key. The corollary of the above is that if we do not embrace a standards view we cannot share knowledge effectively and that means we grow our box 2, 3, 4 visions of the world and with lack of knowledge of what is going on the ability of fear to grow and unfounded movie plot threats to appear real gets ever larger.

5 Standards and Interoperability

Let us take health as a use case for the role of standards in achieving interoperability. When a patient presents with a problem the diagnostic tools and methods, the means to describe the outcome of the diagnosis, the resulting treatment and so on, have to be sharable with the wider health system. This core requirement arises from acceptance that more than one health professional will be involved. If this is true they need to discuss the patient, they need to do that in confidence, and they need to be accountable for their actions which need to be recorded. Some diseases are "notifiable" and, again, to meet the requirement records have to be kept and shared. When travelling a person may enter a country with an endemic health issue (malaria say) and require immunisation or medication before, during and following the visit. Sharing knowledge of the local environment and any endemic health issues requires that the reporting and receiving entities share understanding.

Shared understanding and the sharing of data necessary to achieve it is the essence of interoperability. A unified set of interoperability requirements addresses syntax, semantics, base language, and the fairly obvious areas of mechanical, electrical and radio interoperability.

Syntax derives from the Greek word meaning ordering and arrangement. The sentence structure of subject-verb-object is a simple example of syntax, and generally in formal language syntax is the set of rules that allows a well formed expression to be formed from a fundamental set of symbols. In computing science syntax refers to the normative structure of data. In order to achieve syntactic interoperability there has to be a shared understanding of the symbol set and of the ordering of symbols. In any language the dictionary of symbols is restricted, thus in general a verb should not be misconstrued as a noun for example (although there are particularly glaring examples of misuse that have become normal use, e.g., the use of "medal" as a verb wherein the conventional text "He won a medal" has now been abused as "He medalled"). In the context of eHealth standardisation a formally defined message transfer syntax should be considered as the baseline for interoperability.

Syntax cannot convey meaning and this is where semantics is introduced. Semantics derives meaning from syntactically correct statements. Semantic understanding itself is dependent on both pragmatics and context. Thus a statement such as "Patient-X has a heart-rate of 150 bpm" may be syntactically correct but has no practical role without understanding the context. Thus a heart-rate of 150 bpm for a 50-year old male riding a bike at 15 km/h up a 10% hill is probably not a health concern, but the same value when the same 50 year old male is at rest (and has been at rest for 60 min) is very likely a serious health concern. There are a number of ways of exchanging semantic information although the success is dependent on structuring data to optimise the availability of semantic content and the transfer of contextual knowledge (although the transfer of pragmatics is less clear).

Underpinning the requirements for both syntactic and semantic interoperability is the further requirement of a common language. From the eHealth world it has become clear that in spite of a number of European agreements on implementation of a digital plan for Europe in which the early creation of 'e-health' was eagerly expected the uneven development of the digital infrastructure has in practice made for differing levels of initiative and success across the member states. These led to a confusing vocabulary of terms and definitions used by e-health actors and politicians alike. The meaning of the term e-health has been confused with 'tele-health' which in turn is confused with 'm-health' 'Telemedicine,' a term widely used in the USA has been rejected in Europe in favour of 'tele-health.' There is general agreement that for these terms to be effective we need to redefine them in their practical context. Without an agreed glossary of terms, it will be hard to improve semantic interoperability—a corner stone for the effective building of e-health systems. The vocabulary is not extensive but at present it fails to address the need for clarity in exchange of information in the provision of medical services.

Standards therefore enable and assert interoperability on the understanding that:

$$\text{Interoperability} = \textit{Semantics} \cup \textit{Syntax} \cup \textit{Language} \cup \textit{Mechanics}$$

Quite simply if any of the elements is missing then interoperability cannot be guaranteed. However we do tend to layer standards on top of one another, and alongside each other, and wind them through each other. The end result unfortunately can confuse almost as much as enlighten and unfortunately the solution of developing another standard to declutter the mess often ends up with just another standard in the mess.

In the security domain understanding that we need interoperability is considered the default but simply achieving interoperability is a necessary but insufficient metric for making any claim for security. As has been noted above the technical domain of security is often described in terms of the CIA paradigm (Confidentiality Integrity Availability) wherein security capabilities are selected from the CIA paradigm to counter risk to the system from a number of forms of cyber attack. The common model is to consider security in broad terms as determination of the triplet {threat, security-dimension, countermeasure} leading to a triple such as {interception, confidentiality, encryption} being formed. The threat in this example being interception which risks the confidentiality of communication, and to which the recommended countermeasure (protection measure) is encryption.

The very broad view is thus that security functions are there to protect user content from eavesdropping (using encryption) and networks from fraud (authentication and key management services to prevent masquerade and manipulation attacks). Technical security, particularly cryptographic security has on occasion climbed the ivory tower away from its core business of making everyday things simply secure.

6 Movie Plot Threats

Bruce Schneier has defined movie plot threats as "*… a scary-threat story that would make a great movie, but is much too specific to build security policies around*"[1] and rather unfortunately a lot of the real world security has been in response to exactly these kind of threats. Why? The un-researched and unproven answer is that movie plots are easy to grasp and they tend to be wrapped up for the good at the end.

The practical concerns regarding security and the threats they involve is that they are somewhat insidious, like dripping water they build up over time to radically change the landscape of our environment.

Taking Schneier's premise that our imaginations run wild with detailed and specific threats it is clear that if a story exists that anthrax is being spread from crop dusters over a city, or that terrorists are contaminating the milk supply or any other part of the food chain, that action has to be taken to ground all crop dusters, or to destroy all the milk. As we can make psychological sense of such stories and extend them by a little application of imagination it is possible to see shoes as threats, or liquids as threats. So whilst Richard Reid[2] was not successful and there is no evidence to suggest that a group of terrorists were planning to mix a liquid explosive from "innocent" bottles of liquid, the impact is that due to the advertised concerns the policy response is to address the public fears. Thus we have shoe inspections and restrictions on carrying liquids onto planes. This form of movie theatre scenario and the response ultimately diverts funds and expertise from identifying the root of many of the issues.

Again taking Schneier's premise the problem with movie plot scenarios is that fashions change over time and if security policy is movie plot driven then it becomes a fashion item. The vast bulk of security protection requires a great deal of intelligence gathering, detail analysis of the data and the proposal of targeted counter measures. Very simply by reacting to movie plots the real societal threats are at risk of being ignored through misdirection.

Movie plot derived security policy only works when the movie plot becomes real. If we built out bus network on the assumptions behind Speed we'd need to build bus stops for ingress and egress that are essentially moving pavements that don't allow for the bus to ever slow down, and we'd need to be able to refuel and change drives also without slowing the bus. It'd be a massive waste of money and effort if the attackers did a Speed scenario on the tram or train network or didn't attack at all.

A real problem is that for those making security policy, and for those implementing the countermeasures, they will always be judged in hindsight. If the next attack targets

[1] https://www.schneier.com/blog/archives/2014/04/seventh_movie-p.html.
[2] https://en.wikipedia.org/wiki/Richard_Reid => The "shoe bomber".

the connected vehicle through the V2I network, we'll demand to know why more wasn't done to protect the connected vehicle. If it targets schoolchildren by attacking the exam results data, we'll demand to know why that threat was ignored. The answer "we didn't know…" or "we hadn't considered this…" is not acceptable.

The attractiveness of movie plot scenarios is probably hard to ignore—they give a focus to both the threat and the countermeasures. In addition we need to consider the role of Chinese Whispers[3] in extending a simple story over time.

We can imagine dangers of believing the end point of a Chinese Whispers game:

- Novocomstat has missile launch capability
- Novocomstat has launched a missile
- Novocomstat has launched a bio weapon
- Novocomstat has launched a bio weapon at Neighbourstat
- Neighbourstat is under attack
- Neighbourstat is an ally and we need to defend them
- We're at war with Novocomstat because they've attacked with the nuclear option

As security engineers the guideline is to never react without proof. Quite simply acting on the first of these Chinese Whispers is unwarranted, and acting on the 6th is unwarranted unless all the prior statements have been rigorously verified, quantified and assessed. The various risk management and analysis approaches that exist (there are many) all come together by quantifying the impact of an attack and its likelihood. In recent work in this field in ETSI the role of motivation as well as capability in assessing risk has been re-assessed and now added to the method [E-TVRA]. The aim in understanding where to apply countermeasures to perceived risk requires analysis. That analysis requires expertise and knowledge to perform. In the approach defined by ETSI in TS 102 165-1 this means being able to quantify many aspects of carrying out a technical threat including the time required, the knowledge of the system required, the access to the system, the nature of the attack tools and so forth.

7 Where to Go?

How do you get rid of fear and get acceptance of the threat model? Shared knowledge, shared understanding and willingness to educate each other about what we know and what we may not know. This is the only real way forward. This result is close to zero in boxes 2 and 3 and a bounteous box 1.

8 Conclusions

As stated in Sect. 6 of this paper the approach to getting rid of fear and get acceptance of the threat model is in the wider acceptance of shared knowledge, shared understanding and willingness to educate each other about what we know and what we may not know.

[3] https://en.wikipedia.org/wiki/Chinese_whispers => A parlour game that passes a message round introducing subtle changes in meaning with each re-telling.

The role of standards in giving assurance of interoperability as the key to a solution where more than one stakeholder is involved is difficult to argue against. The nature of the standard is unimportant—it simply has to be accepted by the stakeholders. If the stakeholders are global and largely unknown then an internationally accepted standard is most likely to be the way forward. If, however, the stakeholders are members of a small local team the standard could be as simple as a set of guidance notes maintained on a shared file.

Spreading of fear through a combination of movie plot threats and Chinese Whispers is an inevitable consequence of human curiosity and imagination.

Standards are at the root of sharing a common syntactical and semantic understanding of our world. This is as true for security as it is for any other domain and has to be embraced.

Acknowledgements. Contributions made by the author in development of this paper have in part been supported by EU projects i-locate (grant number 621040), SUNSHINE (grant number 325161) and UNCAP (grant number 643555).

References

[E-TVRA] ETSI TS 102 165-1. https://portal.etsi.org/webapp/WorkProgram/SimpleSearch/QueryForm.asp

[JOHARI] Luft, J., Ingham, H.: The Johari window, a graphic model of interpersonal awareness. In: Proceedings of the Western Training Laboratory in Group Development, Los Angeles (1955)