



Trustworthiness in Industrial System Design

Author:

Marcellus Buchheit

President and CEO

Wibu-Systems USA Inc.

mabu@wibu.com

INTRODUCTION

Trustworthiness in the context of an industrial system is a relatively new term intended to provide a better understanding of the meaning of trust in such a system and how this trust can be approached by the operational user as well as the planner and designer of the system. In general, the definition of trustworthiness by the Industrial Internet Consortium (IIC) is quite abstract and academic and of little help for the operation, planning or design of an industrial system. For example, the direct implementation of the five characteristics of trustworthiness into a concrete system is difficult or even impossible because these characteristics interact with each other and do not permit an isolated implementation of each. The five characteristics of trustworthiness are safety, security, privacy, reliability and resilience.

THE LANDSCAPE OF INDUSTRIAL SYSTEMS

Industrial systems can be very different in purpose, usage and size. Examples of industrial systems are:

- A power plant to create electricity, based on natural resources (water, wind, solar) or by consuming fossil resources (coal, gas, oil, uranium, etc.)
- A hospital to treat the health of sick patients
- An urban transportation system, under or above ground on rails, to carry passengers or goods from one location to another

- A refinery which converts specific fossil resources into specific elements, e.g. converting crude oil into heating oil and gasoline
- A commercial airplane to transport passengers from one airport to another
- An off-shore oil rig to drill and harvest crude oil

The design of such systems is extremely complex and requires highly specialized designers and engineers: Even someone who has skills to harvest crude oil by designing oil rigs cannot use such skills for processing the crude oil and operating an oil refinery. And at first glance, it is difficult to see what a hospital and a power plant have in common beyond that both are industrial systems.

However, these industrial systems share one important common element, and that is a deep-rooted trust between the various stakeholders:

- The owners, investors and operational users trust that these systems work as specified, are profitable and flawless during their expected lifetime.
- Neighbors, customers and employees trust that the systems are safe and do not threaten their health or pollute the environment.
- The government trusts that laws and regulations are fulfilled: e.g., patient privacy standards in a hospital, clean-air directives in a fossil power plant or transportation safety in an urban transportation system.

One challenge is to fulfill this trust during the design and the operation of the industrial

system: Trust is a human trait and hard to explain as an output of industrial design principles. That is why trustworthiness is so important: It bridges the gap between design and trust. And it works for all types of industrial systems: Even if the design and operation of a system are very different, the principles of trustworthiness are always the same.

A BETTER WAY TO IMPLEMENT A SYSTEM EVERYONE CAN TRUST

- In the beginning, most designs were not reliable. Stakeholders who invested in the systems, were disappointed with lost profits because each failure stopped production output.
- Over time, reliability of the system and its components were improved and stakeholders began to address the resilience characteristic of trustworthiness, e.g., making the system more robust against unexpected disruptions, such as fire, but also against

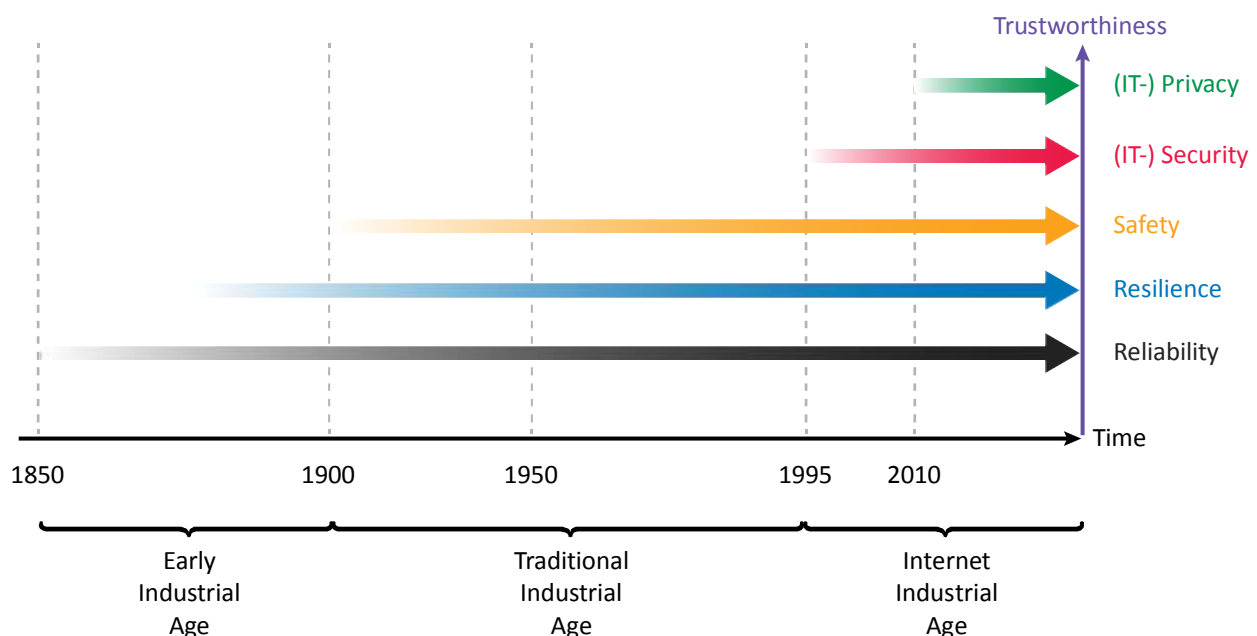


Figure 1: The Evolution of Trustworthiness in Industrial Systems

The five characteristics of trustworthiness are not new in the history of industrial design. Figure 1 shows the history of these characteristics: They were introduced at different stages during the progression of the industrial revolution.

For example, think about the evolution of a steel plant from the mid-19th century up to today and you will see:

- natural catastrophes such as storms, flooding or earthquakes.
- With the increasing power of unions and the influence of government, especially in democracies enforced by voters, safety issues came to the forefront around the end of the 19th and beginning of the 20th century. The early focus was on employee safety and later expanded

to the protection of the community and nature.

- With the availability of internet connectivity, industrial systems were able to access websites and exchange information via email. This quickly raised the risk of hacker attacks and, subsequently, the requirement for active IT security beyond the physical (or traditional) security, such as fencing around a plant or surveillance by security guards, which were the mainstay of the oldest industrial systems.
- With storage of more and more personal data in industrial systems, privacy has become a key concern and privacy regulations have become an important factor in the industrial environment, mainly focused on IT and electronic data. There are industrial system designers who think that privacy does not affect most industrial systems beyond specific exceptions like hospitals. However, the new General Data Protection Directive (GDPR) ¹ regulation of the European Community, for example, clearly specifies that privacy also addresses employees, and because nearly all industrial systems require employees to operate equipment, privacy is an integral part of industrial design.

The time coordinates in Figure 1 are not accurate for all systems. But interestingly enough, the relative introduction

progression was similar for automobiles and airplanes: Reliable automobiles were first available around 1910, but safety features were not incorporated until 1930²; reliable commercial airplanes were available in 1930 but the demand for and implementation of safety features only began in the 1950s.

¹ <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&qid=1473816357502&from=en> and <http://data.europa.eu/eli/reg/2016/679/oj>

² History of Car Safety: <https://www.crashtest.org/history-car-safety/> and <https://www.theaa.com/breakdown-cover/advice/evolution-of-car-safety-features>

TRUSTWORTHINESS AND ITS APPROACH OF COMPLETENESS

Characteristic, the model's completeness is proven. The Trustworthiness Target Model in Figure 2 demonstrates this graphically: There are four quadrants of targets which

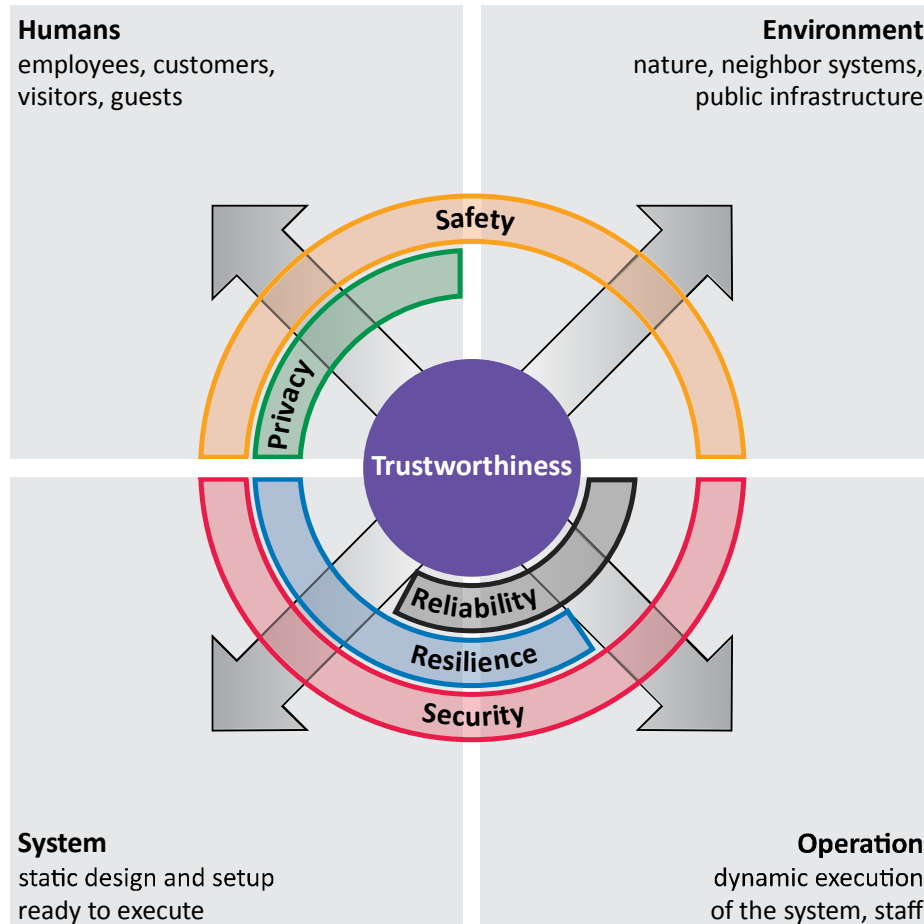


Figure 2: The Trustworthiness Target Model

While most experts agree that the five trustworthiness characteristics and their interaction are an important goal for any industrial system design, there are ongoing discussions about whether a design which fulfills all requirements of trustworthiness can be automatically trusted by all parties. One way to answer the question is to analyze how these five characteristics address the specific targets which they protect. If the list of targets is complete and every target has at least one assigned Trustworthy

require a specific protection from harm:

- **Humans** (top left quadrant) represent not only customers (like patients in a hospital) but also employees, visitors or guests. It is obvious that *privacy* will protect this quadrant. But *safety* is also responsible to shield humans from harm. *Security, reliability and resilience of the system as part of trustworthiness, however, have no direct relation to this quadrant.*

- **Environment** (top right quadrant) is exclusively protected by *safety*. It includes any natural aspects that are accessed by the system (e.g., pollution of air or water in nature), but also private neighborhoods and public infrastructure. No other trustworthiness characteristic directly addresses this quadrant.
- The **System** (bottom left quadrant) describes only the static system, including installed software and operational data, but not the operation itself. *Security* is responsible for its protection; *resilience* and to some degree *reliability* also protect the system against damage or loss of components, e.g., by fire or theft.
- Finally the system in **Operation** is mostly shielded by *security*, *reliability* and partially *resilience*. The operational part of the system also includes the staff, e.g., being protected by security against human threats from outside.

Employees are targeted in the *Humans* as well as in the *Operation* quadrant which may sound unusual. But, for example, every employee knows exactly when he or she has their yearly review meeting with the boss: The employee wears one hat for the personal expectation of receiving higher salary and benefits and another hat as a staff member agreeing to work with higher efficiency and better interaction with the rest of the team.

The complete vision of trustworthiness can be seen in these four quadrants: All important elements are protected. This model also shows that the five

trustworthiness characteristics have sharp boundaries between their protected targets. This makes it easier to understand the design focus around each one of the five characteristics.

The reader may be confused by these sharp boundaries of trustworthiness characteristics between the four quadrants. For example: are not resilience functions typically installed in a system to prevent a disaster for humans and environment in case of a major system malfunction? The short answer is that such resilience functions are used to establish safety functions that ultimately protect the humans and environment; the resilience functions themselves do not provide the protections. The general answer will be provided by the concept of Trustworthiness Methods, introduced later in this article.

It is not possible to redirect the arrows in Figure 2 by 180 degree to ask the question “Who is threatening the trustworthiness of a system?” Simply stated, every member of the four quadrants threatens any of the five characteristics of trustworthiness: e.g., humans by making errors or attacks, the environment by disturbances and the system or operation by faults.

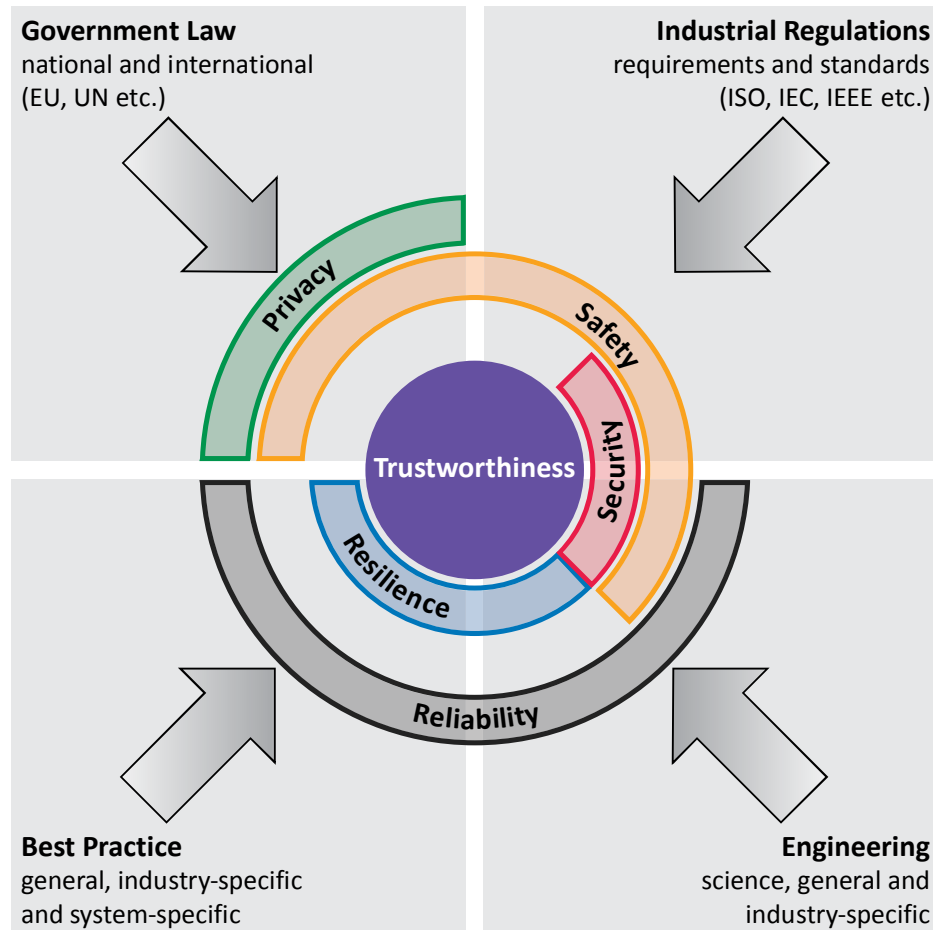


Figure 3: The Trustworthiness Foundation Model

It is still possible to redirect the arrows, but the definition of the quadrants needs to be modified, as shown in Figure 3. Quadrants are fundamental sources of knowledge and directives which influence trustworthiness. And again, the five trustworthiness characteristics can be drawn with sharp boundaries between the quadrants:

- **Privacy** is heavily defined by government law. The European community started with GDPR; it is expected that other countries will follow very soon with their own laws.
- Similarly, **safety** is defined more deeply in government law, not just around

consumer products such as automobiles, but also environment-critical systems like nuclear plants or oil refineries. On the other hand, industrial associations are providing additional regulations and policies for their specific industries, which are documented in standards from international organizations such as ISO, IEC or IEEE to define safety regulations.

- **Security** is affected again by such standards but also by *engineering* knowledge. Such knowledge may be general (e.g., IT and network security) but also limited in their usage for specific industrial branches only.

- **Reliability** is also addressed by such engineering knowledge but additionally by *best practice* of an industrial branch and probably even inside a specific system.
- **Resilience**, similar to reliability, has its foundation in best practice and engineering. However, from the educational perspective, resilience in general is less engineered than reliability, which is why the main foundation (the rim of resilience fills the entire quadrant) is best practice and not engineering.

To demonstrate that the boundaries of the five characteristics are as sharp as shown in the quadrants, we can test the opposite and see that:

- Even with reliability and resilience as the oldest characteristics in the industrial system design, there are very few government laws or standards focused on these two areas. They are both demanded by the stakeholders of an industrial system and fulfilled by engineering principles and best practice.
- Safety and privacy on the other hand are mostly government enforced or demanded in standards, so there is little foundation from best practice and engineering. Of course, safety equipment and future privacy functions will be designed using engineering, but this is an implementation rather than a foundation for these two characteristics.
- Finally, security is not a target of government law, at least not today. And it would be a bad idea to implement and operate security by best practice: The

risk that a security vulnerability could be opened by some incomplete best practice would be high.

Similar to the Target Model, the Foundation Model's four quadrants describe all sources of knowledge. These sources are well addressed by specific trustworthiness characteristics and represent more evidence of the completeness of trustworthiness.

The boundaries of the Trustworthiness Characteristics in the Foundation Model describe the original historical motivation for these characteristics and it can be expected that the related sectors will become wider in the future. For example, privacy is likely to be a future target of industrial regulations and engineering.

Of course, there are other important design principles for an industrial system, examples of which include usability, efficiency or flexibility: They are not part of trustworthiness and they are not part of trust that the system works as expected. These principles are partially affected by trustworthiness but the analysis of this interaction is outside of the scope of this article.

TRUSTWORTHINESS METHODS

The first challenge of using trustworthiness in system design is that none of the trustworthiness characteristics can be implemented as a separate technology and that the trustworthiness of an industrial system cannot be implemented by just combining such technologies: The characteristics may support or block each

other; a simple combination results in new challenges.

The solution is to take the system design away from the system characteristics and move to methods which are assigned to the system characteristics. In traditional system design such methods had been used for a long time but were not classified by the Trustworthiness Characteristics. And this classification can be extended by other attributes.

Definition: A **Trustworthiness Method** is defined as a component, tool, technology, software application, an operational procedure or a management directive which is assigned to at least one trustworthiness characteristic. Such methods are named as *Trustworthiness Safety Method*, *Trustworthiness Resilience Method*, etc. If a method is assigned to several trustworthiness characteristics, the list of characteristics is separated with a slash, e.g., *Trustworthiness Security/Privacy Method*.

The definition of such a method is intentionally as broad as possible as only the assignment to one or more trustworthiness characteristics is key.

Examples of Trustworthiness Methods are:

- *Fire extinguisher*: a tool and a Trustworthiness Safety Method.
- *CO₂ fire suppression system*³: a tool and a Trustworthiness Resilience Method (the main purpose is to protect the

system not the environment or humans; CO₂ is indeed dangerous for humans).

- *Network firewall*: a tool and a Trustworthiness Security Method.
- *Melt-resistant steel*: technology and a Trustworthiness Resilience Method.
- *Windmill Restart*: operational procedure for airplanes during an engine flameout and a Trustworthiness Resilience Method⁴.
- *Electric motor brush replacement*: operational procedure and a Trustworthiness Reliability Method.
- *Brushless motor*: technology and a Trustworthiness Reliability Method.
- *Encryption of all social security numbers on servers*: management directive and a Trustworthiness Privacy Method.

Examples of Trustworthy Methods assigned to several trustworthiness characteristics are:

- *Fire-resistant plastic*: technology and a Trustworthiness Safety/Resilience Method: it prevents a fire from spreading and endangering humans (safety) but also prevents the system itself from damage (resilience).
- *Using encrypted hard disks*: management directive and a Trustworthiness Security/Privacy Method.

Most of these Trustworthiness Methods for industrial systems have existed for many years. The only novelty being the assignment

³ Gaseous fire suppression, https://en.wikipedia.org/wiki/Gaseous_fire_suppression and Carbon dioxide, https://en.wikipedia.org/wiki/Carbon_dioxide

⁴ Flameout, <https://en.wikipedia.org/wiki/Flameout>

to one or more of the trustworthiness characteristics and the addition of a new name.

CLASSIFICATION OF TRUSTWORTHINESS METHODS

Beyond the assignment to one or more trustworthiness characteristics, Trustworthiness Methods can be classified in other directions:

Definition: A Trustworthiness Method can be **essential** or **supportive**. The *essential* attribute means that dropping of this Trustworthiness Method leads to a loss of the assigned trustworthiness characteristic in the specific context. In contrast, a *supportive* Trustworthiness Method increases the trustworthiness of one or more of the other essential methods in the same context.

Examples:

- The network firewall in an internet/LAN router is essential for security. Disabling this firewall would lead to instant loss of security in the context of internet access protection.
- A VPN system in an internet/LAN router is essential for security in the context of communication across the internet. But it is also supportive for the internet access protection because any non-VPN access by authorized remote access clients can be dropped, requiring that hackers have difficulty in obtaining VPN access. But a temporary disabling of the VPN access would not result in a loss of security in the internet access protection context.

- A fire alarm sensor is an essential Trustworthiness Safety Method. Disabling it would go against any fire alarm legal requirements and industrial regulations.
- A video surveillance system with automatic picture evaluation could also detect open fires and send an additional alarm, which makes this system supportive. But the usage does not follow official requirements and it is not guaranteed to work in all conditions of a fire. That is why it is not essential. Shutting off this surveillance system would essentially drop the physical security of the system but not the fire safety system.

Another classification for Trustworthiness Methods is the location in the system status. The meaning of system status in the context of trustworthiness is explained in the next section. A Trustworthiness Method is originally designed for one specific status but can also be useful in other status locations. Removing or modifying a Trustworthiness Method for one status could lead to unexpected consequences for another status if this relationship is not defined, which leads to another classification:

Definition: A Trustworthiness Method is **primary** for a specific system status if it is originally designed for this location. A Trustworthiness Method is **secondary** for a specific system status if it useful for this status but primary for another system status.

THE TRUSTWORTHY SYSTEM STATUS

The Trustworthy System Status defines the health of an existing system from *normal* to *ruined* as the result of specific levels of loss of functionality. Only in the *normal* status does the system work as specified. In the next sections we will delve deeper into this status definition, ending with a universal Trustworthy System Status Model (TSSM).

Ideal: A System with No Threats

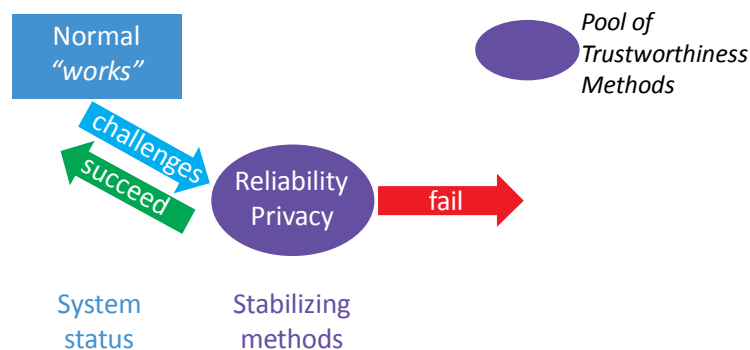


Figure 4: Trustworthiness in a system with no Incidents

The Trustworthy System *normal* meets everyone's expectations on how the system should work and everyone has full trust in this system. As long as the system is not a target of threats this *normal* status could be permanent. Of course, a system without threats is purely theoretical, but is a good starting point to understand the Trustworthy System Status.

Even without threats, Trustworthy Methods are necessary: Every system needs maintenance and every system has to fulfill privacy requirements. The methods are frequently challenged by the system as shown in Figure 4: The specific methods assigned to reliability and privacy ideally reject the challenge and the *normal* system

status is established again. Examples for such Trustworthy Methods are:

- A combustion engine needs frequent oil changes.
- Standard software products need frequent updates (service packs).
- New regulations and laws around privacy must be reviewed and Trustworthy Methods around privacy most likely need updates or additional installations.

In Figure 4, the purple circle contains all types of Trustworthy Methods which are necessary to keep the Trustworthy System Status normal as long as possible.

If a Trustworthy Method was forgotten or does not work as expected then the challenge cannot be rejected and the method fails (red arrow in Figure 4). We will see in the next sections what

happens in that case.

Defending the System Against Incidents

After this theoretical but core system design is finished, all potential threats must be addressed. In the spirit of the definition of trustworthiness such threats can come from outside, e.g., a hurricane, loss of power or a hacker attack, or from inside, e.g., an overheated motor or a design error which results in a failed system status.

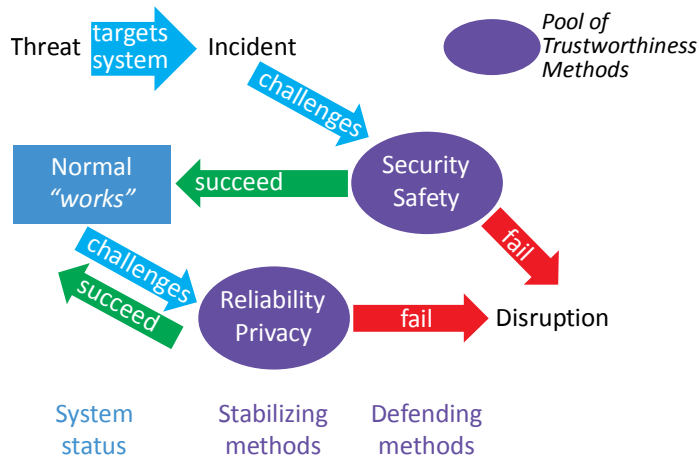


Figure 5: Trustworthiness in normal system status receiving incidents

A threat in general is not a problem in and of itself. For example, every electric motor has the principle threat of overheating and every internet access the threat of a hacker attack. But only if the threat actually reaches the system, is it relevant. In this case the threat created an incident, as shown in Figure 5. Now Trustworthiness Mechanisms which are assigned to security and/or safety are trying to reject this incident. For example, a safety method reduces the speed of the overheated motor so it can cool down. Or the firewall in the router blocks the hacker attack as a security method. If protection is successful, the system status returns back to normal. If the threat cannot be prevented – either because the Trustworthiness Mechanisms are not working as expected or an oversight by design failure – the system status switches from normal to *disrupted*.

Disrupted Systems

A disrupted system is not necessarily a big problem. The Trustworthy System Status just defines this as a condition that the system is outside the *normal* status and needs some individual handling to be

brought back to *normal*. For example, an airplane engine flame-out situation would cause the captain to react by bringing the airplane to a lower altitude so he can try a windmill Restart. After that maneuver the pilot needs to check the entire system, e.g., to find out why the engine flamed out, bring the airplane back to the original altitude and declare the problem as *solved*, and thus change the status back to *normal*. Figure 6 demonstrates this case: The pilot's action to bring the airplane to a lower altitude is a Trustworthiness Safety Method, reaching a safe status of *disrupted*. The Windmill Restart is a Trustworthiness Resilience

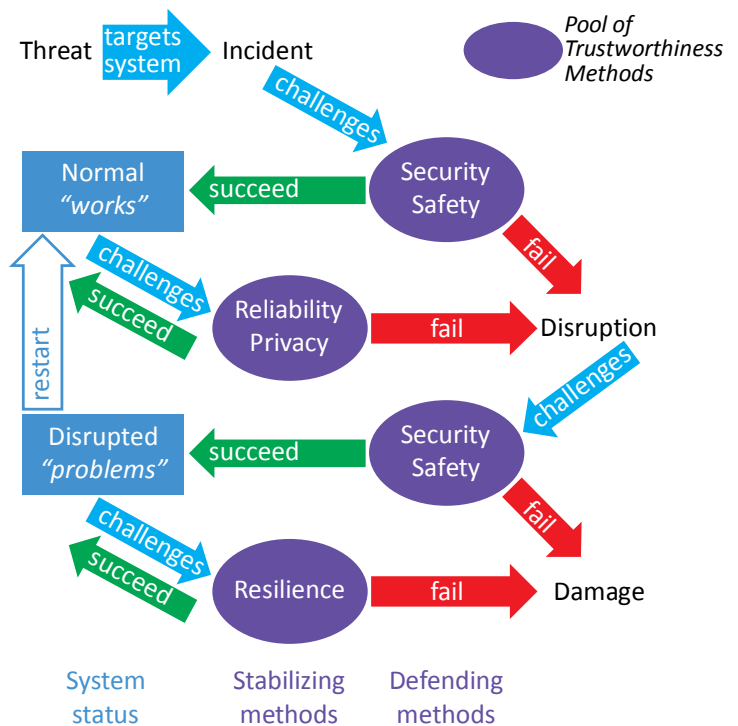


Figure 6: Trustworthiness in normal and disrupted system status

Method. If one of these methods fails, the *disrupted* status cannot be continued, and the system status moves to *damaged* (because now one of the engines cannot be

started again – an issue which needs deeper analysis and likely repair after a safe emergency landing). Of course, there are industries, such as nuclear plants, for example, which must take disrupted systems seriously while analyzing the reason for the disruption and modifying Trustworthiness Reliability, Security and Safety Methods before the system restart back to normal is possible. For other industries it is good practice to document disruptions and also take precautions and make specific enhancements to prevent this disruption in the future.

The interesting thing about status is the symmetry: *Defending Methods*, assigned to security and safety, try to protect the current system status from incidents to avoid latter failures, e.g., from *normal* to *disrupted* or

from *disrupted* to *damaged*. *Stabilizing Methods* on the other hand try to defend challenges which are coming from the current status. Furthermore, Trustworthiness Methods, assigned to reliability or privacy, are replaced by methods assigned to resilience as soon as the normal Trustworthy System status leaves. This switch is a result of the original definitions of reliability and resilience: All methods, assigned to reliability, target well-known issues inside the normal operation of the system. As soon as the normal status moves to the disruption stage, we reach the unexpected status of the system. Now methods assigned to resilience take over to stabilize the current status.

THE TRUSTWORTHY SYSTEM STATUS MODEL (TSSM)

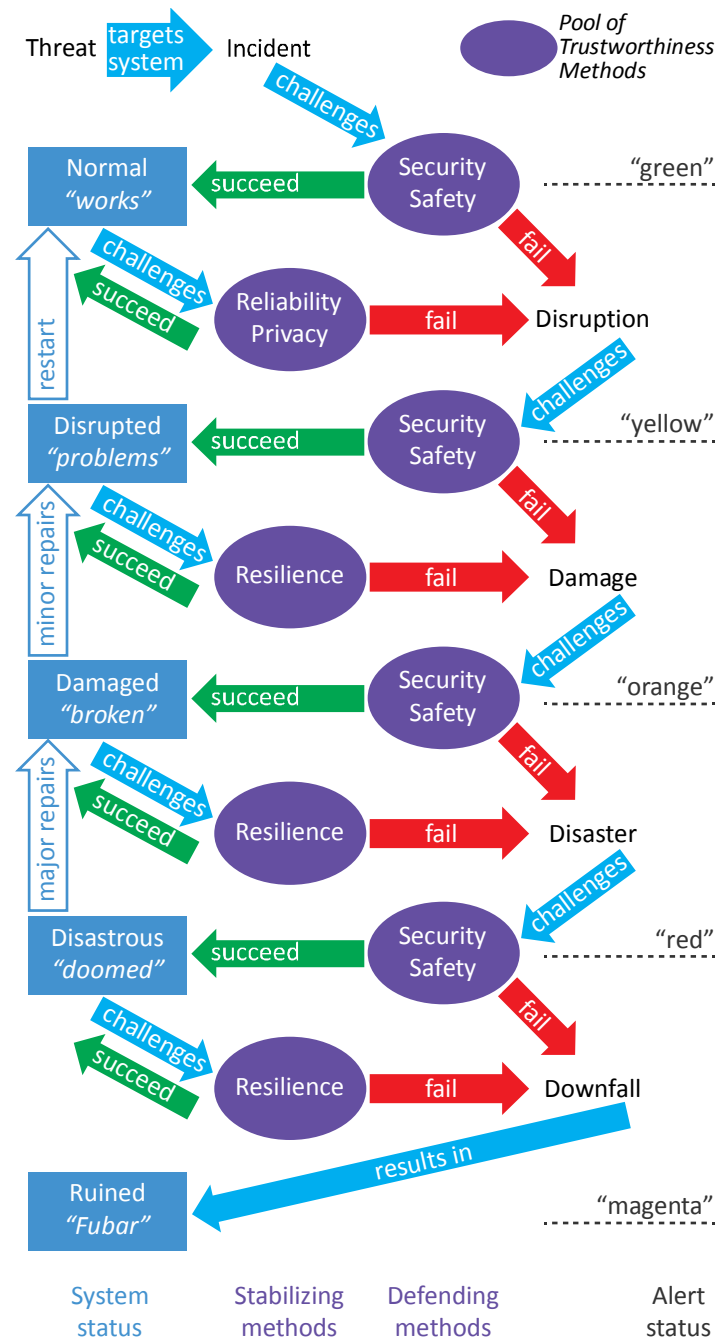


Figure 7: Figure 7: Trustworthy System Status Model (TSSM)

The schema from *normal* to *damage* can be extended to further statutes which bring the system into a growingly fatal situation. This

results in the Trustworthy System Status Model (TSSM) shown graphically in Figure 7.

As shown, there are no more changes in using Trustworthiness Methods in the subsequent status – everything is based on methods assigned to safety, security and resilience. Traditional alert colors are used to demonstrate status: *green* for *normal*, *yellow* for *disrupted*, *orange* for *damaged*, *red* for *disastrous* and *magenta* for *ruined*. The graphic also shows the required effort to move from a lower system status to a higher one. Principally, such a status change could also make jumps – for example from *damaged* to *normal*; to keep this graphic simple such practical options were not added. To have a better understanding of the lower status values I continue my example with the flamed-out engine of an airplane. Assuming the Trustworthiness

Method of bringing the airplane to a lower altitude or that the windmill Restart fails, the status would stay *damaged*. In this case, the other engine would probably flame out too, e.g., because the airplane ran out of fuel. Now the status falls to *disastrous*. If the pilot is able to succeed with the Trustworthiness Method of an emergency landing, the status will stay as *disastrous* and the airplane would probably fly again after significant repair. Otherwise, the plane will crash and end as *ruined* making it clear that there is no way back to *normal*.

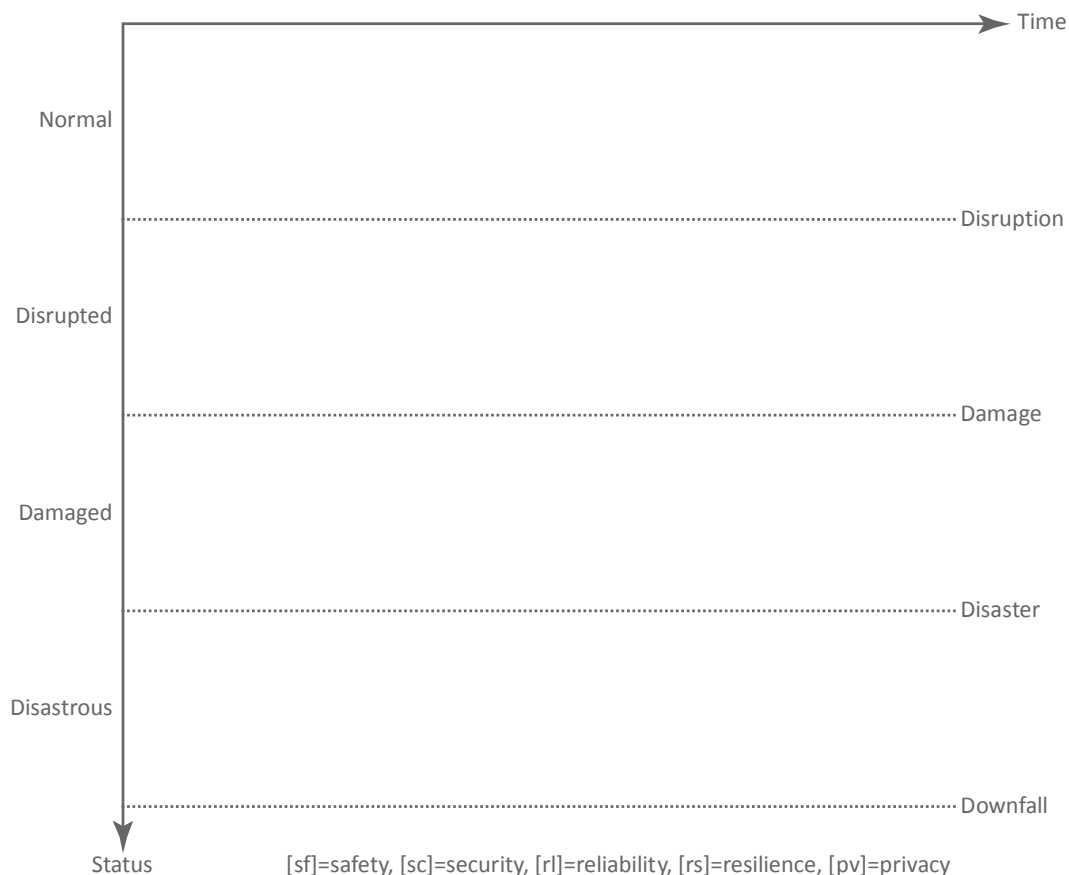


Figure 8: TSSM planning table

SYSTEMATIC USAGE OF TSSM INTO THE SYSTEM DESIGN

The TSSM graphic can be used as schema to plan systems and also to describe expected or unexpected status changes around the Trustworthiness System Status. Figure 8 shows the empty schema: With any anticipated or unexpected status change, the method to defend or stabilize (see figure 7) can be entered with their *succeeded* or *failed* arrows, latter crossing one of the dotted lines from up to down. And also *restart* or *repair* methods can be drawn as

arrows, crossing one or more lines from down to up.

EXAMPLES OF TSSM USAGE

Figure 9 shows an example of the TSSM planning table with a well-known IT problem: Hard disks are not reliable so we simulate the threat *Hard disk may break*. Within the normal system status this will be addressed by a RAID 1/5/10 system⁵, frequent checking of the disks, replacing the damaged disks and doing parallel frequent backups.

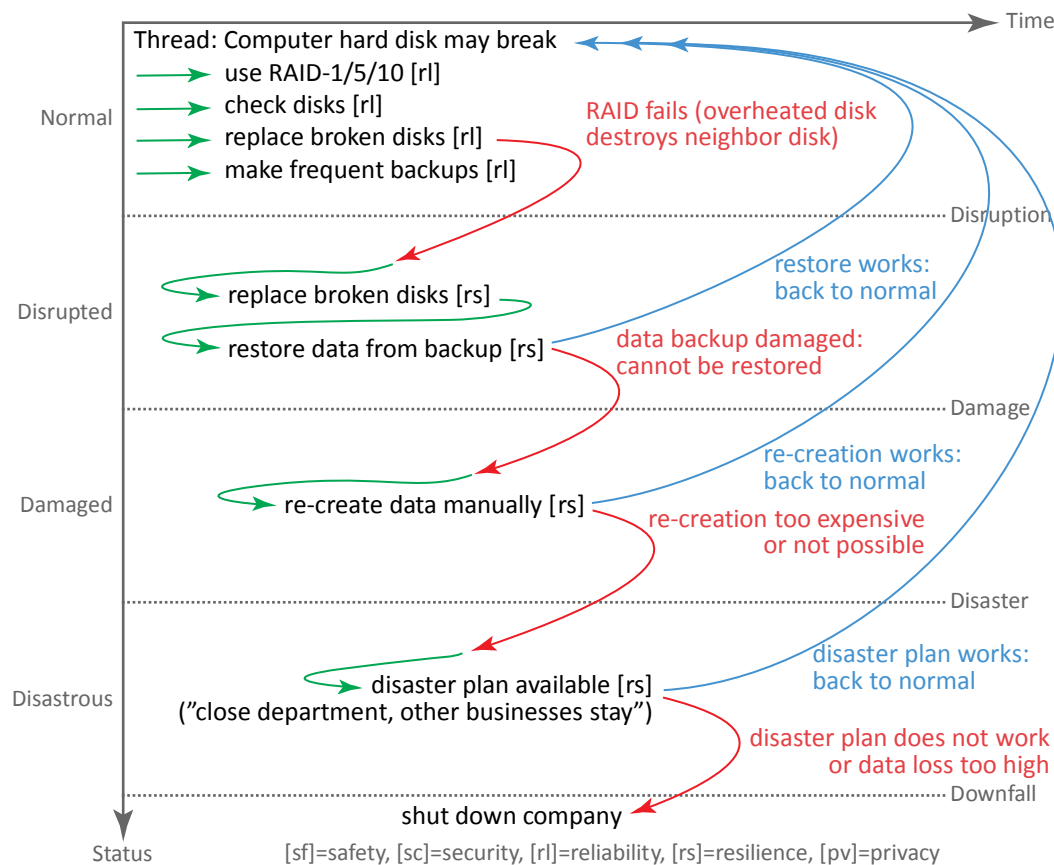


Figure 9: Usage of TSSM planning table to address an IT problem

⁵ RAID systems: <https://en.wikipedia.org/wiki/RAID>

An unexpected hardware error brings the system to the status of disruption. If the restore from the backup fails and the recovery of the data is too expensive or just not possible, the whole company may end in a disaster.

CLASSIFICATION OF TRUSTWORTHINESS METHODS INSIDE TSSM

The TSSM provides another classification of the Trustworthiness Methods: The location of the specific TSSM status:

- Primary Trustworthiness Reliability or Privacy Methods are designed and used around the *normal* status.
- Primary Trustworthiness Resilience Methods are designed and used in the time after the system has left the *normal* status.
- Primary Trustworthiness Safety or Security Methods can be designed and used in any status.

All these methods are *primary* (see definition in the section above): They were originally introduced to support trustworthiness at a specific TSSM status. Of course, they can also support any other TSSM status *secondarily*. For example, a protection wall between fire-critical areas in a plant was originally introduced to prevent a small fire from spreading from one area to another, resulting in a large plant-wide fire. In the TSSM, such a protection wall would be defined as a Trustworthiness Resilience Method to defend the *damaged* status, preventing moving into the disastrous status. But this wall could also be used in the *normal* status as a Trustworthiness Safety

Method, preventing dangerous air pollution from being transferred from one plant area to another. And, at the same time, act as a Trustworthiness Security Method in all statuses, preventing unauthorized people from moving from one plant area to another. All these additional Trustworthiness Methods are *secondary*.

In general Trustworthiness Methods, primarily introduced for the *normal* status, are still valid in the other statuses and act there as secondary. This also answers the question of missing Trustworthiness Privacy Methods in the TSSM beyond the *normal* status: This does not mean that after any disruption all privacy protection is gone. Instead most Trustworthiness Privacy Methods introduced for the *normal* status continue to exist as secondary. However, it would be quite unusual to introduce a new primary Trustworthiness Privacy Method just for the *disrupted* status without purpose for the *normal* status.

SUMMARY

Trustworthiness is not just an abstract term to better understand trust in industrial systems. It can also be practically used in designing such systems. By introducing Trustworthiness Methods with their different classification, it is easier for designers to understand how trustworthiness characteristics can be used to design stable, trustful systems. The Trustworthy System Status Model (TSSM) helps designers to plan a system beyond the normal status and proactively prevent, by using specific Trustworthiness Methods, a system that has reached disrupted status

from slipping into a damaged or disastrous status or even permanently lost.

This article introduces Trustworthiness Methods and TSSM publicly for the first time. The author hopes for critical feedback

from all readers in order to enhance and refine the explained models, providing a future trustworthiness model that is highly usable for practical design and operations of industrial systems.

➤ Return to [IIC Journal of Innovation landing page](#) for more articles and past editions.

The views expressed in the *IIC Journal of Innovation* are the contributing authors' views and do not necessarily represent the views of their respective employers nor those of the Industrial Internet Consortium.

© 2018 The Industrial Internet Consortium logo is a registered trademark of Object Management Group®. Other logos, products and company names referenced in this publication are property of their respective companies.