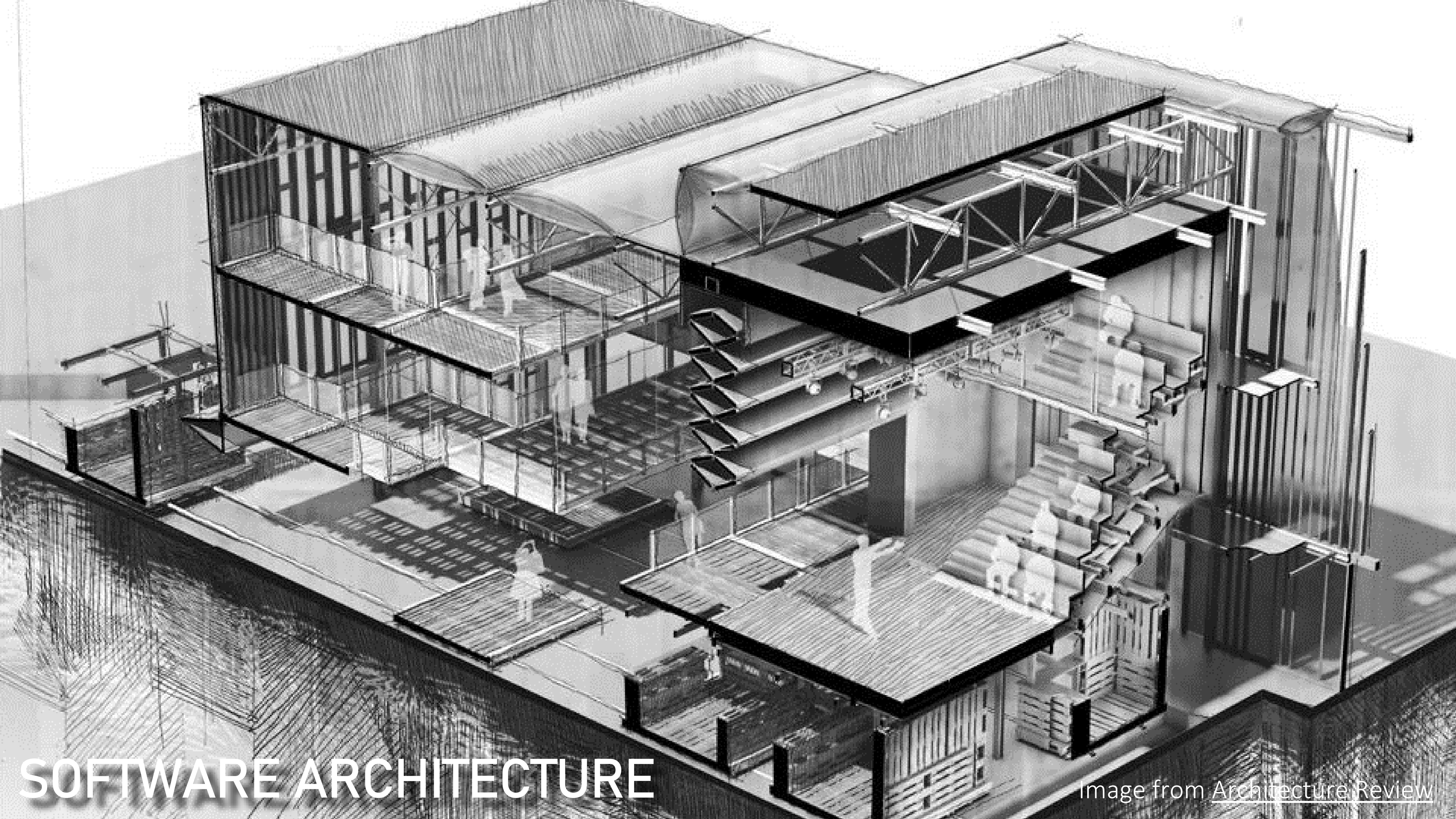
Trust Management in Digital Autonomous Ecosystems

Barbora Buhnova, PV226 Lasaris Seminar, March 2, 2023



SOFTWARE ARCHITECTURE

Image from Architecture Review

SOFTWARE ARCHITECTURE

Image from Crandall Arambula

SOFTWARE ARCHITECTURE

Image from IEEE CSS

AUTONOMOUS ECOSYSTEMS

Image from Parking Network

AUTONOMOUS ECOSYSTEMS

Information Exchange in Coordinated Moves

Image from Parking Network

AUTONOMOUS ECOSYSTEMS

Collision Avoidance with Misbehaving Vehicle

Information Exchange in Coordinated Moves

Image from Parking Network

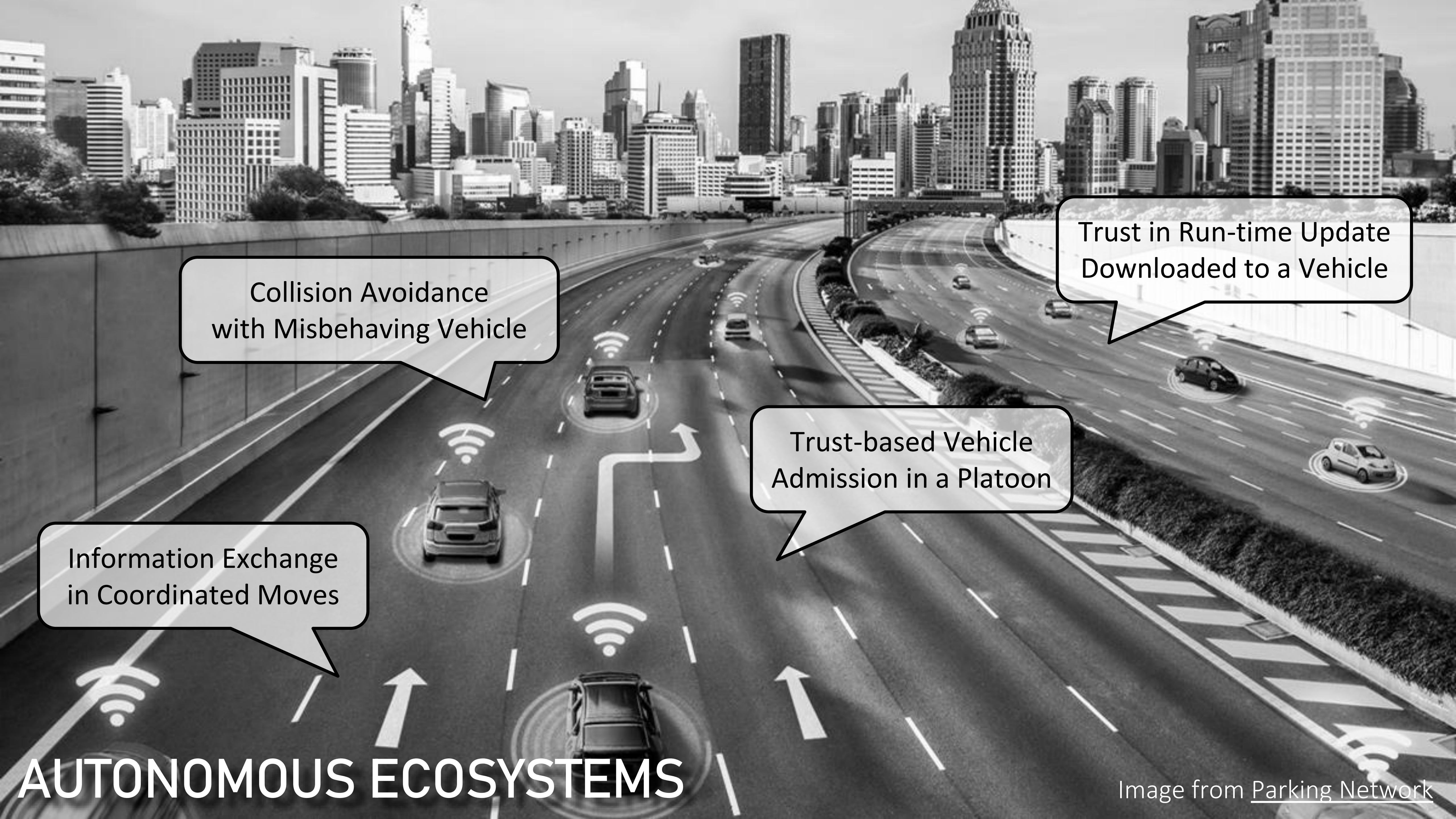
AUTONOMOUS ECOSYSTEMS

Collision Avoidance with Misbehaving Vehicle

Trust-based Vehicle Admission in a Platoon

Information Exchange in Coordinated Moves

Image from Parking Network



AUTONOMOUS ECOSYSTEMS

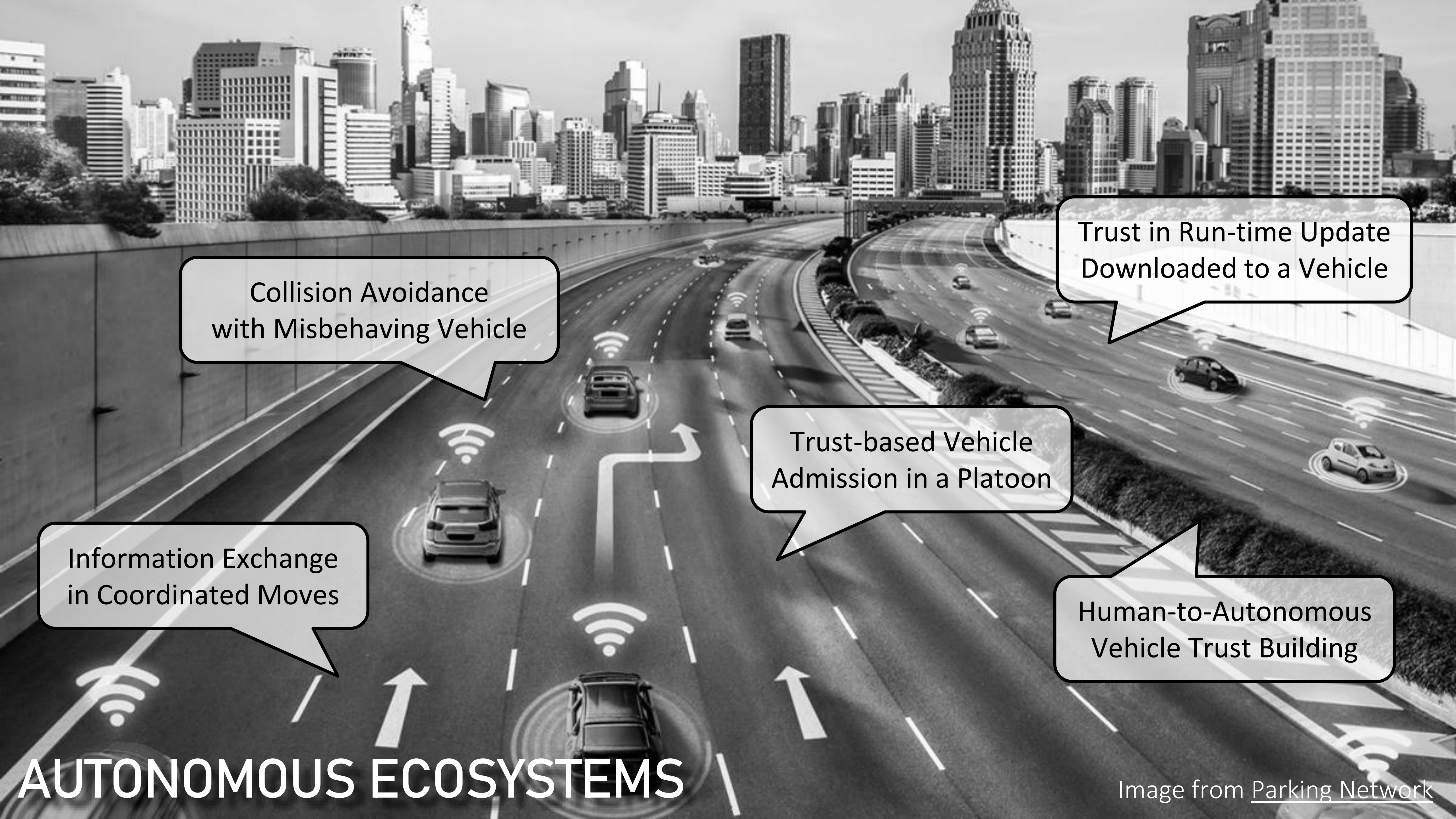
Collision Avoidance with Misbehaving Vehicle

Trust in Run-time Update Downloaded to a Vehicle

Trust-based Vehicle Admission in a Platoon

Information Exchange in Coordinated Moves

Image from Parking Network



Collision Avoidance with Misbehaving Vehicle

Trust in Run-time Update Downloaded to a Vehicle

Trust-based Vehicle Admission in a Platoon

Information Exchange in Coordinated Moves

Human-to-Autonomous Vehicle Trust Building

AUTONOMOUS ECOSYSTEMS

Image from Parking Network

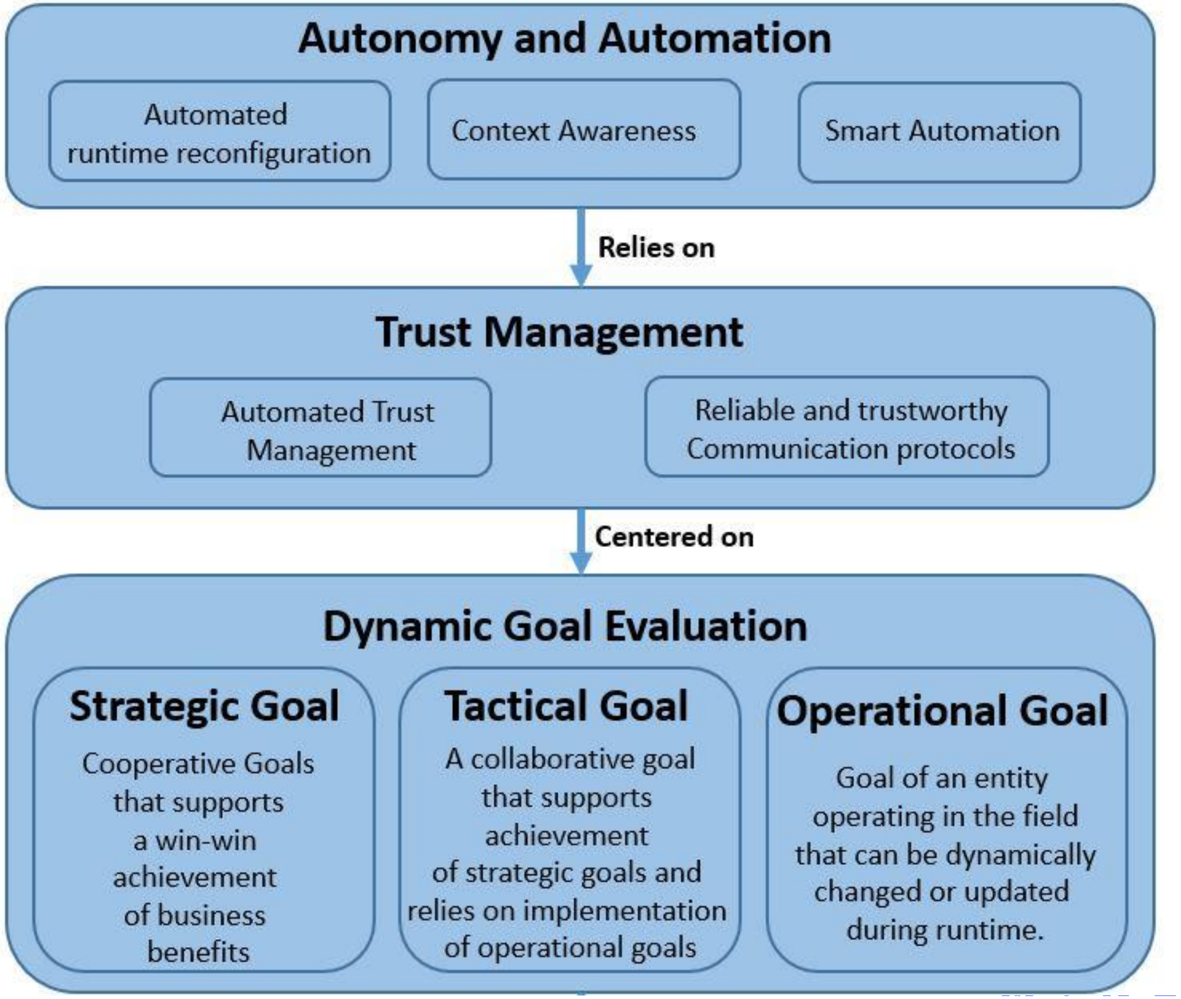
AUTONOMOUS ECOSYSTEMS

Image from WardsAuto

Autonomous Dynamic Ecosystems

- Need to build software systems out of traditional boundaries.
- Transition towards software ecosystems able to support **dynamic, smart and autonomous** features demanded by modern software systems.
- Decentralized interaction leading to **self-organization around mutual goals** and formation of social relationships.

[Ref] Capilla, R., Cioroaica, E., Buhnova, B., and Bosch, J. (2021). On autonomous dynamic software ecosystems. IEEE Transactions on Engineering Management.



Autonomy and Automation

Automated runtime reconfiguration

Context Awareness

Smart Automation

Relies on

Trust Management

Automated Trust Management

Reliable and trustworthy Communication protocols

Centered on

Dynamic Goal Evaluation

Strategic Goal

Cooperative Goals that supports a win-win achievement of business benefits

Tactical Goal

A collaborative goal that supports achievement of strategic goals and relies on implementation of operational goals

Operational Goal

Goal of an entity operating in the field that can be dynamically changed or updated during runtime.

Trustworthiness does NOT guarantee Trust

- Approaches exist to ensure trustworthiness of the individual ecosystem components, via improving their **security**, **reliability**, **availability**, etc.
- Trust is difficult to get addressed via such solutions.
- Trust is a social psychological concept crucial for forming partnerships, it is conceptually a **belief** about a system that is **out of our control**.
- Although the system might **declare its trustworthiness**, this does not give a guarantee that it can be trusted.
- This is an effect of the fact that **malicious objects** can enter the ecosystem with the intention to disrupt the basic functionality of a network for malicious purposes.

Agents with Malicious Intentions

- Malicious objects can enter the ecosystem with **the intention to disrupt** the basic functionality of the ecosystem for malicious purposes.
- This can be done via **causing harm** directly or via **damaging the reputation** of good (well behaving) objects or by increasing the trustworthiness of misbehaving objects.

What if tech progress gets out of our control? Is tech ban a solution?

- Not really. A safe digital ecosystem therefore needs to be **equipped for dealing with the misbehaving objects** (which are capable of jeopardizing the ecosystem functionality) by restricting their services and prioritizing the trustworthy alternatives.

PROTECTING SAFETY in the midst of the TECHNOLOGICAL EVOLUTION

AUTONOMOUS ECOSYSTEMS

Image from Parking Network

Build “Immune Response” within the Ecosystem

- **Learn from our biology** – “immune response” ecosystem

1. We need to be able to **tell the “good” and the “evil” apart**
2. We need the ability to pacify the “evil”

- We need the ability to make the distinction during real-time interaction
  - Because malicious intentions can be hidden behind collaborative intentions (or change)
- Using mechanisms of **trust**
  - Challenging to formalize, influenced by many factors

Trust in Human Ecosystems

- Helps us decide if we enter the interaction, expose our vulnerabilities
  - Or even isolate the other person, report them to law enforcement, spread our mistrust experience
- Trust building is influenced by many factors
  - Inherited trust from our environment, peer opinion, reputation building
  - Our levels of uncertainty and vulnerability
  - Compliance checking with our expectations and tolerance of being off
- Sometimes we enter in the interaction even if we don’t trust
  - If we know the violation will result in holding the violator accountable (law enforcement)
  - When we do not have that much to lose – and there is something to win from the interaction

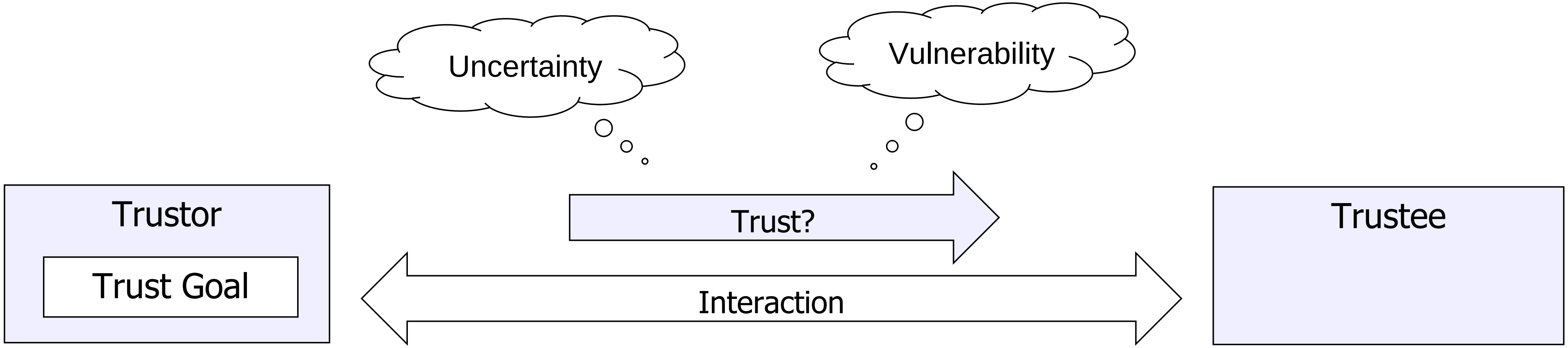
UNDERSTANDING TRUST

What is Trust?

- **Trust in Sociology:** Subjective probability that another party will perform an action that will not hurt my interest under uncertainty or ignorance.
- **Trust in Psychology:** Cognitive learning process obtained from social experiences based on the consequences of trusting behaviors.
- **Trust in Economics:** Expectation upon a risky action under uncertainty and ignorance based on the calculated incentives for the action.
- **Trust in Automation:** Attitude or belief that an agent will help achieve another agent’s goal in a situation characterized by uncertainty and vulnerability.

What is Trust?

The attitude or belief of an agent (trustor) to achieve
a specific goal in interaction with another agent (trustee)
under uncertainty and vulnerability.



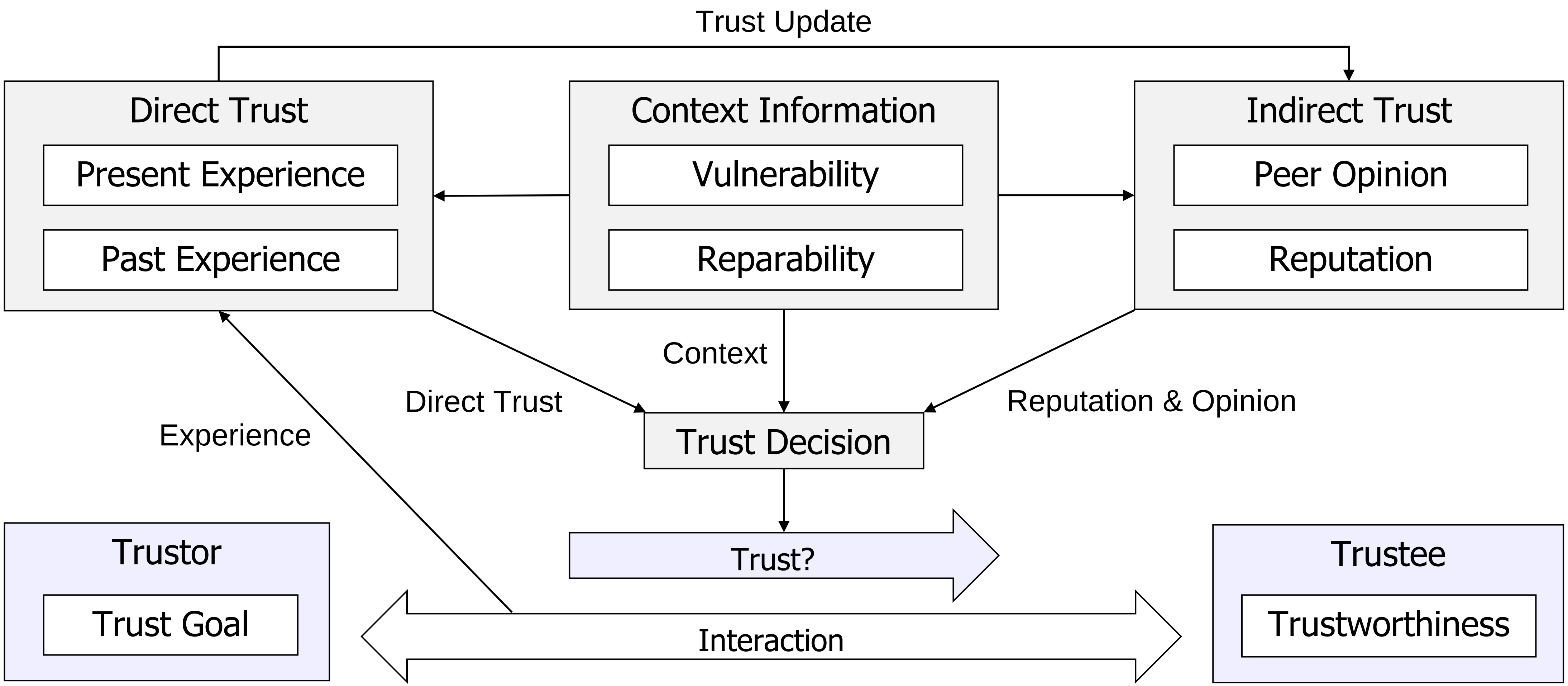
Characteristics of Trust

- **Subjective:** Trust is viewed using the centrality of an agent, wherein the trust is computed based on trustor's observation (i.e., direct trust) as well as the opinion (i.e., feedback or indirect trust) of the other agents.
- **Asymmetric:** Trust is an asymmetric property, i.e., if an agent A trusts another agent B, it does not guarantee that B also trusts A.
- **Transitive:** System agent A is more likely to develop trust towards an agent B if A trust agent C that trusts agent B.

Scope of Trust Evaluation

- **Local:** It represents the trust based on an agent-agent relationship, wherein an agent evaluates the trustworthiness of another agent using local information such as its current observation and past experience.
- **Global:** Global trust is based on the reputation of an agent within the ecosystem, wherein the reputation of each agent might be influenced by the local trust score of each of the other agents in the ecosystem.
- **Context-specific:** Trust of an agent towards another agent varies with context. A trust relation between the agents is usually dynamic and depends on multiple factors such as temporal factors or location.

Trust Management Components



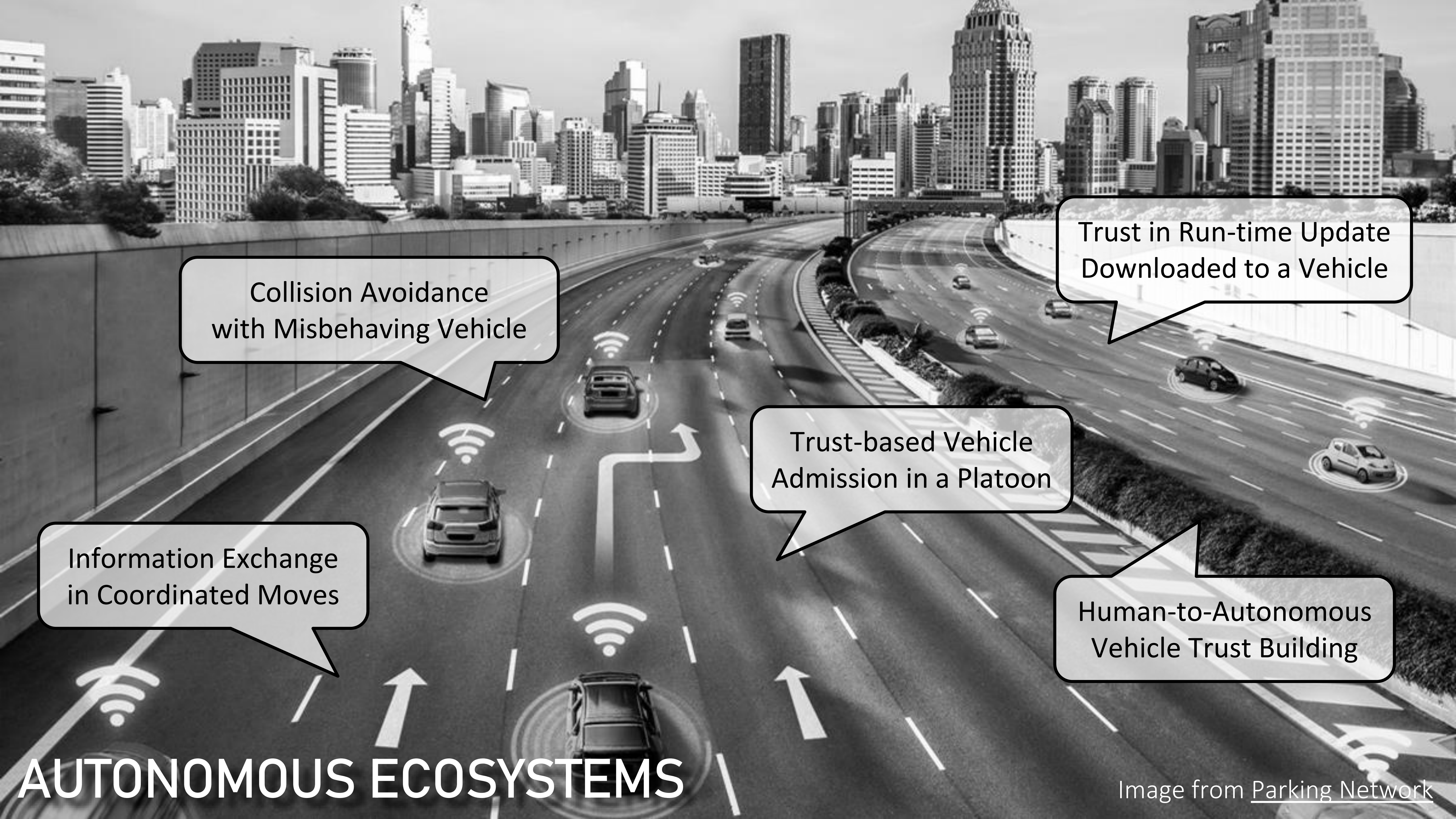
Direct-Trust Evaluation – Trust Metrics

Trust Metrics refer to the features that are chosen and combined in trust computation. These features can refer to:

- **QoS Metrics**, which represent the confidence that an agent is able to offer high quality of the delivered service, e.g. in terms of reliability, availability, security or accuracy.
- **Social Metrics**, which represent the social relationships among ecosystem agents, which can include integrity, benevolence, honesty, friendship, openness, altruism, unselfishness.

SOCIAL METRICS?
What do you mean?

AUTONOMOUS ECOSYSTEMS



Collision Avoidance with Misbehaving Vehicle

Trust in Run-time Update Downloaded to a Vehicle

Trust-based Vehicle Admission in a Platoon

Information Exchange in Coordinated Moves

Human-to-Autonomous Vehicle Trust Building

Image from Parking Network

Research Problems

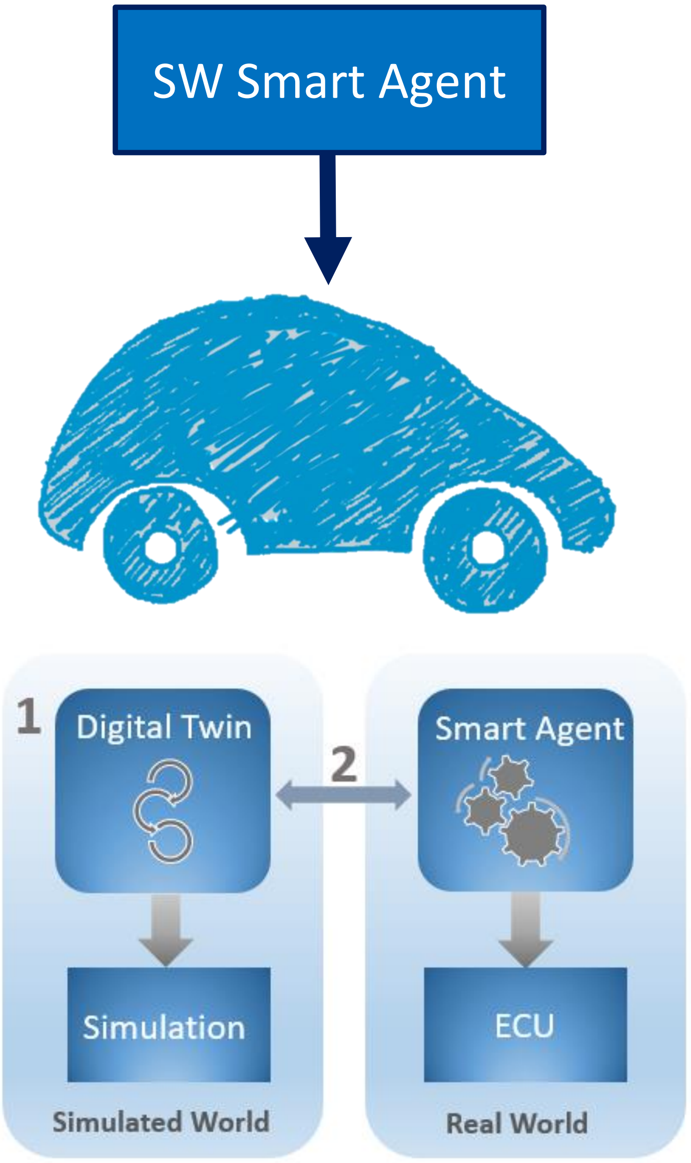
1. **System-to-System Trust (update scenario):** A vehicle is downloading a black-box update at runtime. May I trust that update and give it access to my critical driving functions?
2. **System-to-System Trust (collision avoidance scenario):** Two vehicles approaching each other. May I trust the other vehicle that it does not intend to cause a crash?
3. **Trust-Based Adaptive Safety:** How shall I adapt my safety mechanisms to the level of trust? What if I misdudge trust (false postivies/negatives)?
4. **Trust Management and Governance:** What mechanisms (e.g., incentives, evidence collection, reparation) shall be in place to protect and govern trust values?
5. **Management and Governance for Other Values:** How shall we engineer systems that protect and govern other than trust values (e.g., ethics, fairness, solidarity)?

PROBLEM 1
System-to-System Trust (update scenario)

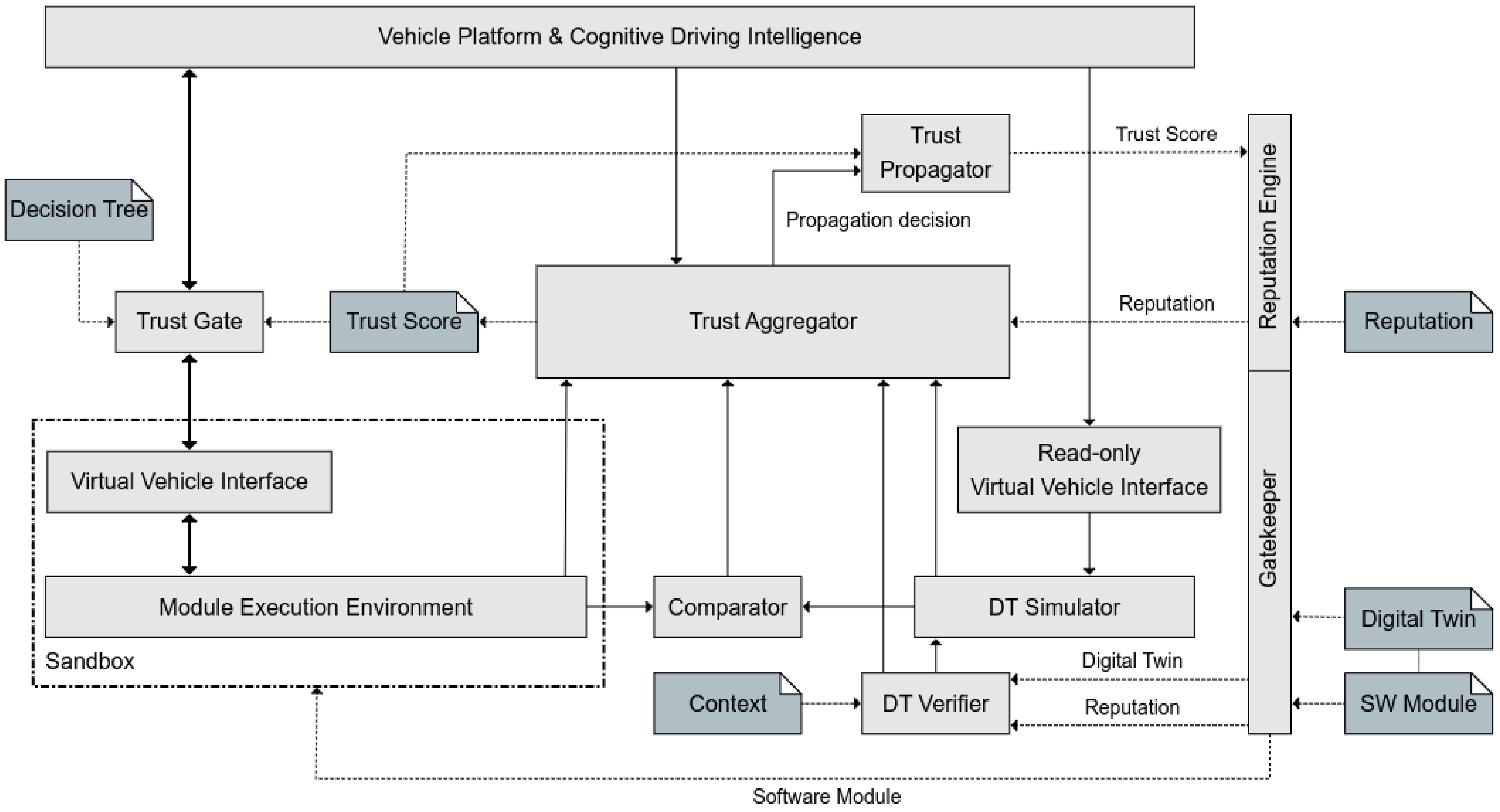
Building Trust in a SW Smart Agent

- Trust of a vehicle into an automated update downloaded to the car.
- DT of the black-box smart agent provided.
- DT run in a simulation environment to get predictive awareness of possible harmful effects.
- A fail-over behavior can be triggered for the system in the real world.

[Ref] Cioroaica, E., Kuhn, T., and Buhnova, B. (2019, May). (Do not) trust in ecosystems. In 2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER) (pp. 9-12). IEEE.


SW Smart Agent

Conceptual Framework of Runtime Trust Evaluation



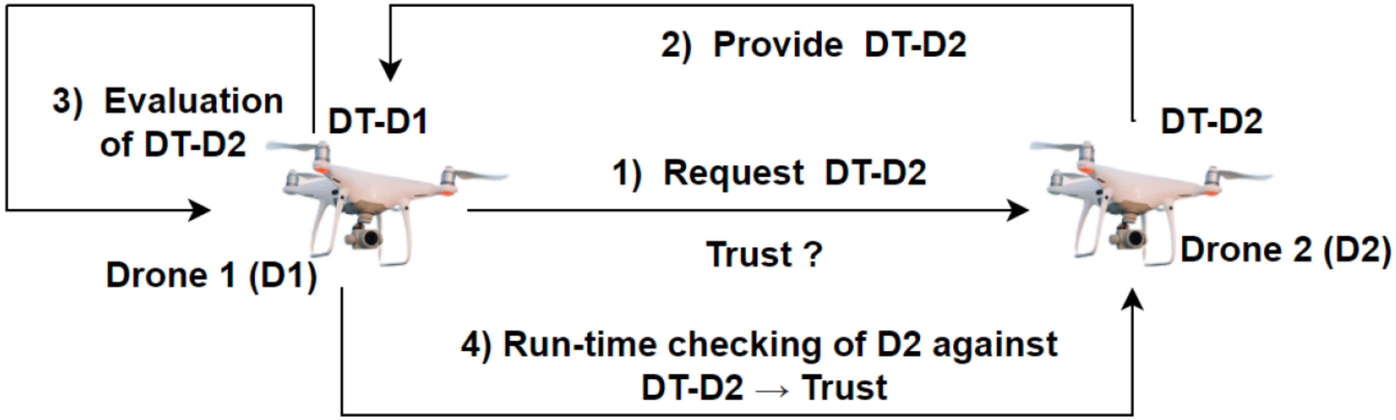
The Role and Effects of Predictive Simulation

- Progress towards **trustworthy and ethical evolution** of dynamic autonomous ecosystems.
- Predictive simulation can lead towards integration of **mechanisms capable of foreseeing the effects of actions** within an ecosystem in order to respond with appropriate ethical reactions.
- Enabling the creation of **predictive awareness**, we can see a move towards **artificial cognitive machines**, which can support the process of ensuring moral operation of AI-controlled systems.

PROBLEM 2
System-to-System Trust (collision avoidance)

Trust Building via Predictive Simulation

- Consider **Drone 1** assessing its level of **trust in Drone 2**, as illustrated below.
- From the point of view of Drone 1, Drone 2 is **a black box** and **out of our control**, with **unknown intentions** (possibly malicious, hidden behind good behaviour).



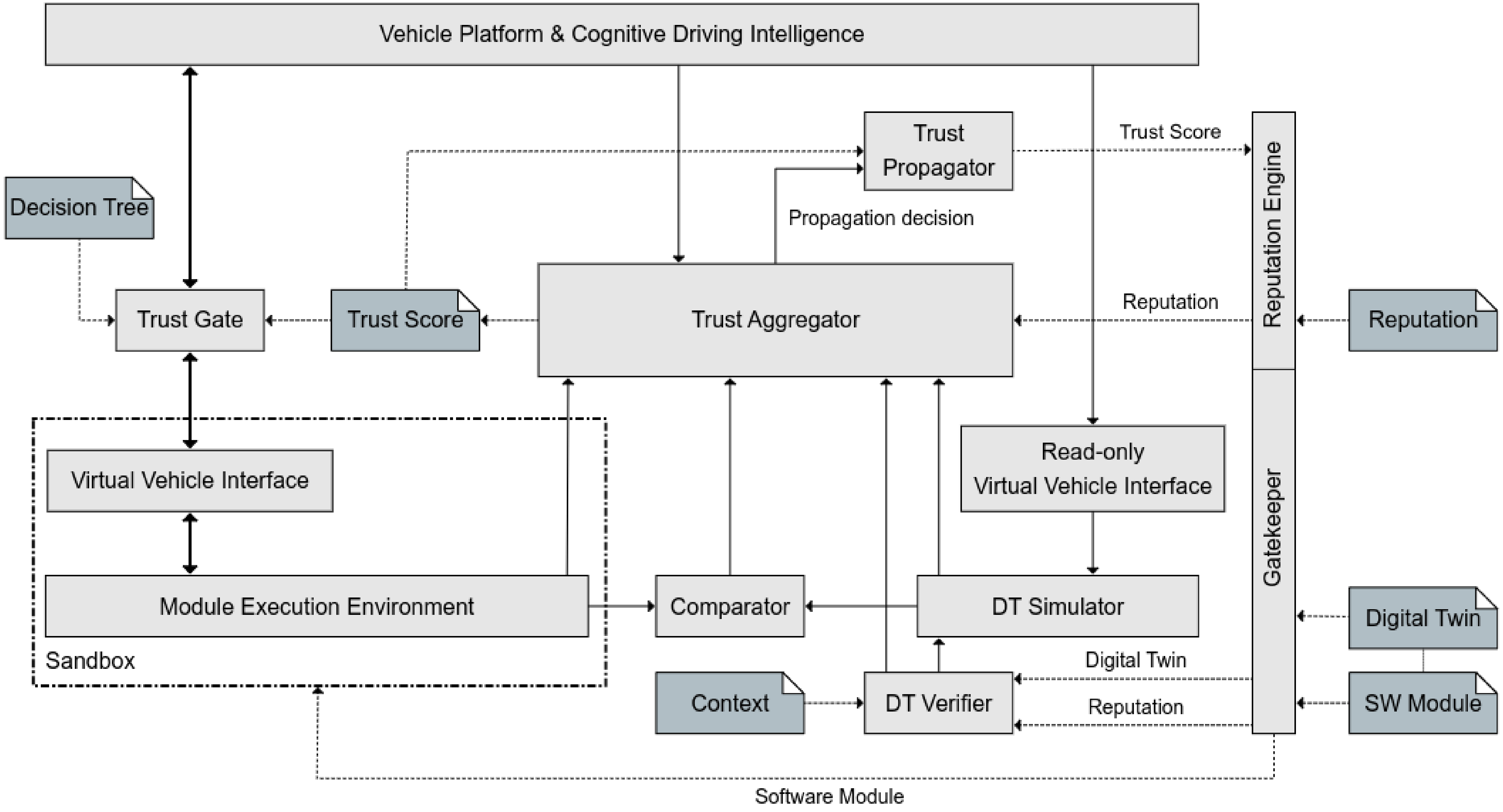
[Ref] Iqbal, D., and Buhnova, B. (2022). Model-based Approach for Building Trust \\ in Autonomous Drones through Digital Twins. In IEEE SMC 2022 International Conference on Systems, Man, and Cybernetics (pp. 9-12). IEEE.

PROBLEM 3
Trust-Based Adaptive Safety

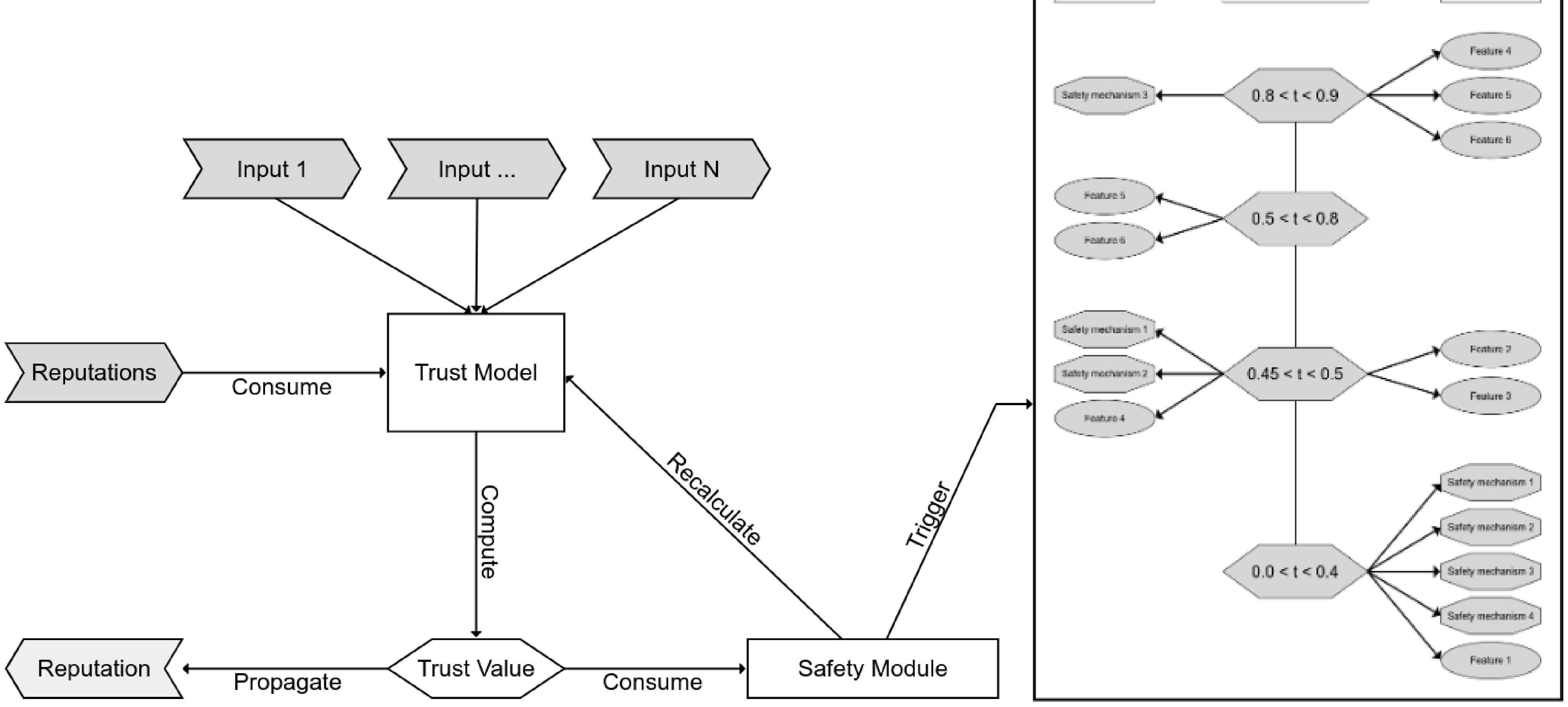
Safety Assurance in Face of Untrusted Agents

- **Run-time adaptive safety:** Evolution of safety mechanisms is needed to support dynamic and self-adaptive architectures of autonomous ecosystems.
- **Adaptation to the level of trust:** Responding to the level of trust among autonomous agents.
- **Safety supervision and control:** An agent that is reported as untrusted might fall under safeguarding of its trustworthy operation, with enabling/disabling its (safety) features.
- **False positives and negatives:** Need to accommodate for trust misjudgment with gradual safety mechanisms.
- **Technical feasibility:** The safeguarding mechanisms can be checked/downloaded on entry to the ecosystem (e.g., the city, highway).

Sandboxing within our Conceptual Framework



Trust-Driven Adaptive Safety



PROBLEM 4
Trust Management and Governance

Trust Governance Mechanisms

- **Trust score** calculation, propagation, update
- **Incentives**, i.e., rewards and punishment mechanisms
- **Reparation** and redemption mechanisms
- **Evidence** collection
  - Pre-incident to predict somebody is attempting misbehaviour
  - Post-incident to either identify the source of misbehaviour,
    or to understand whether a corrective action needs to be taken

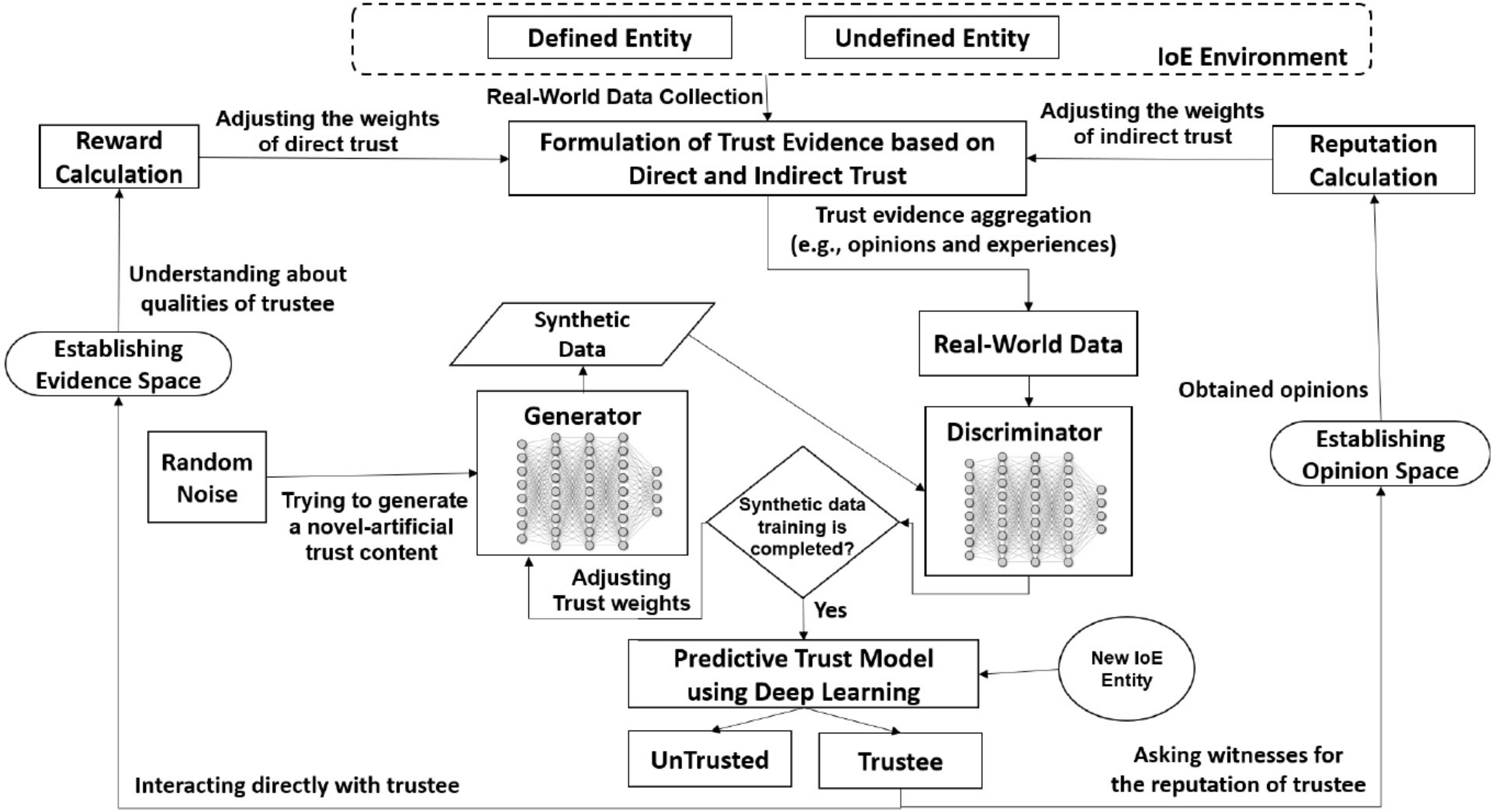
Trust Governance Mechanisms

- **Trust score** calculation, propagation, update
- **Incentives**, i.e., rewards and punishment mechanisms
- **Reparation** and redemption mechanisms
- **Evidence** collection
  - Pre-incident to predict somebody is attempting misbehaviour
  - Post-incident to either identify the source of misbehaviour, or to understand whether a corrective action needs to be taken

Further Mechanisms to Consider

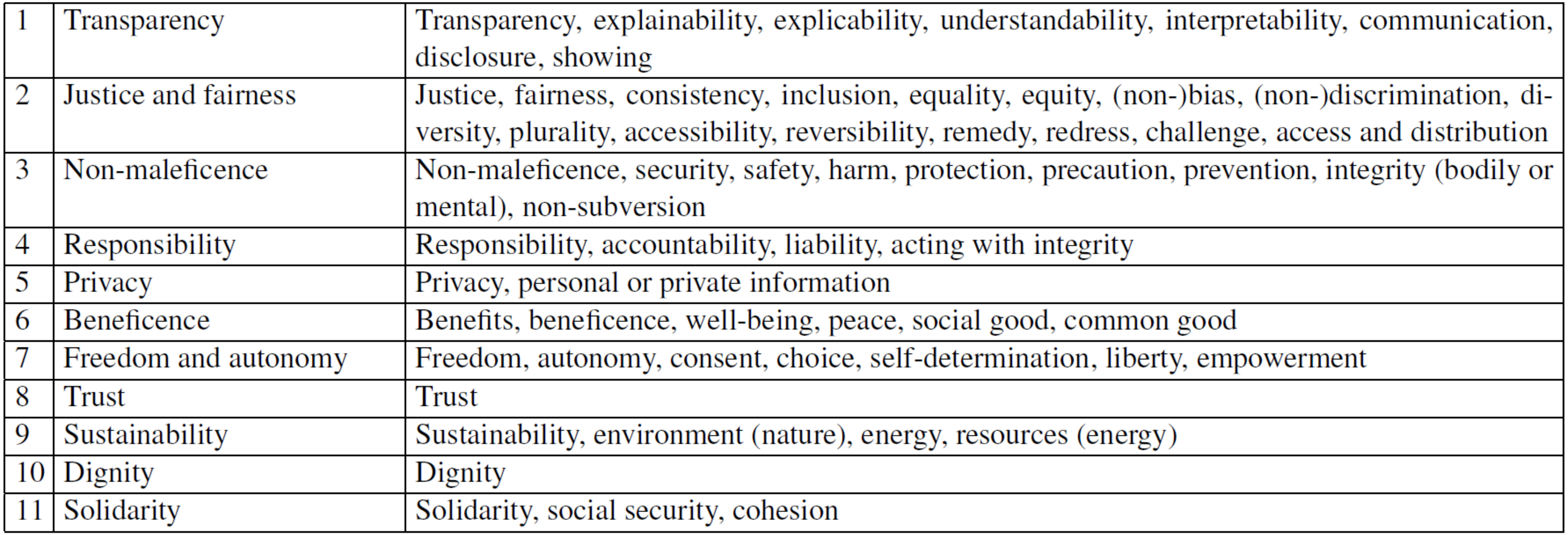
- **Default Trust Score of New Agents:** on which trust score shall a new agent start
- **Trust Erosion:** trust score is subject to decay in case of no or too few interactions
- **Building Trust in the Trustworthy:** employing explanation to give evidence of trustworthiness
- **Black Swan Blindness** and its other sources

Deep-Learning based Trust Management


Defined Entity
Undefined Entity
IoE Environment
Real-World Data Collection
Adjusting the weights of direct trust
Reward Calculation
Formulation of Trust Evidence based on Direct and Indirect Trust
Adjusting the weights of indirect trust
Reputation Calculation
Trust evidence aggregation (e.g., opinions and experiences)
Understanding about qualities of trustee
Synthetic Data
Real-World Data
Establishing Evidence Space
Generator
Discriminator
Obtained opinions
Establishing Opinion Space
Random Noise
Trying to generate a novel-artificial trust content
Synthetic data training is completed?
Adjusting Trust weights
Yes
Predictive Trust Model using Deep Learning
New IoE Entity
UnTrusted
Trustee
Asking witnesses for the reputation of trustee
Interacting directly with trustee

PROBLEM 5
Management and Governance for Other Values

Social Values and Ethical Principles



1	Transparency	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
2	Justice and fairness	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
3	Non-maleficence	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
4	Responsibility	Responsibility, accountability, liability, acting with integrity
5	Privacy	Privacy, personal or private information
6	Beneficence	Benefits, beneficence, well-being, peace, social good, common good
7	Freedom and autonomy	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
8	Trust	Trust
9	Sustainability	Sustainability, environment (nature), energy, resources (energy)
10	Dignity	Dignity
11	Solidarity	Solidarity, social security, cohesion

Ethical Digital Identities (EDIs)

- For methodological transition towards **fully ethical agents**
  - that involves the contribution of multiple stakeholders in **making and justifying the moral judgements**, achieving a societal-driven justification of morality.
- Can be used by providers of digital assets to **safeguard the ethical quality** of, e.g.,
  - an AI component and receive a positive social reputation.
- Through **EDI**s assets can become living identities with **traceable evidence of moral implications** evaluated from:
  - the perspective of ethical and social concerns of **experts**
  - as well as **non-experts**, possibly engaging large population in return of suitable incentives.

Ethical Digital Identities (EDIs)

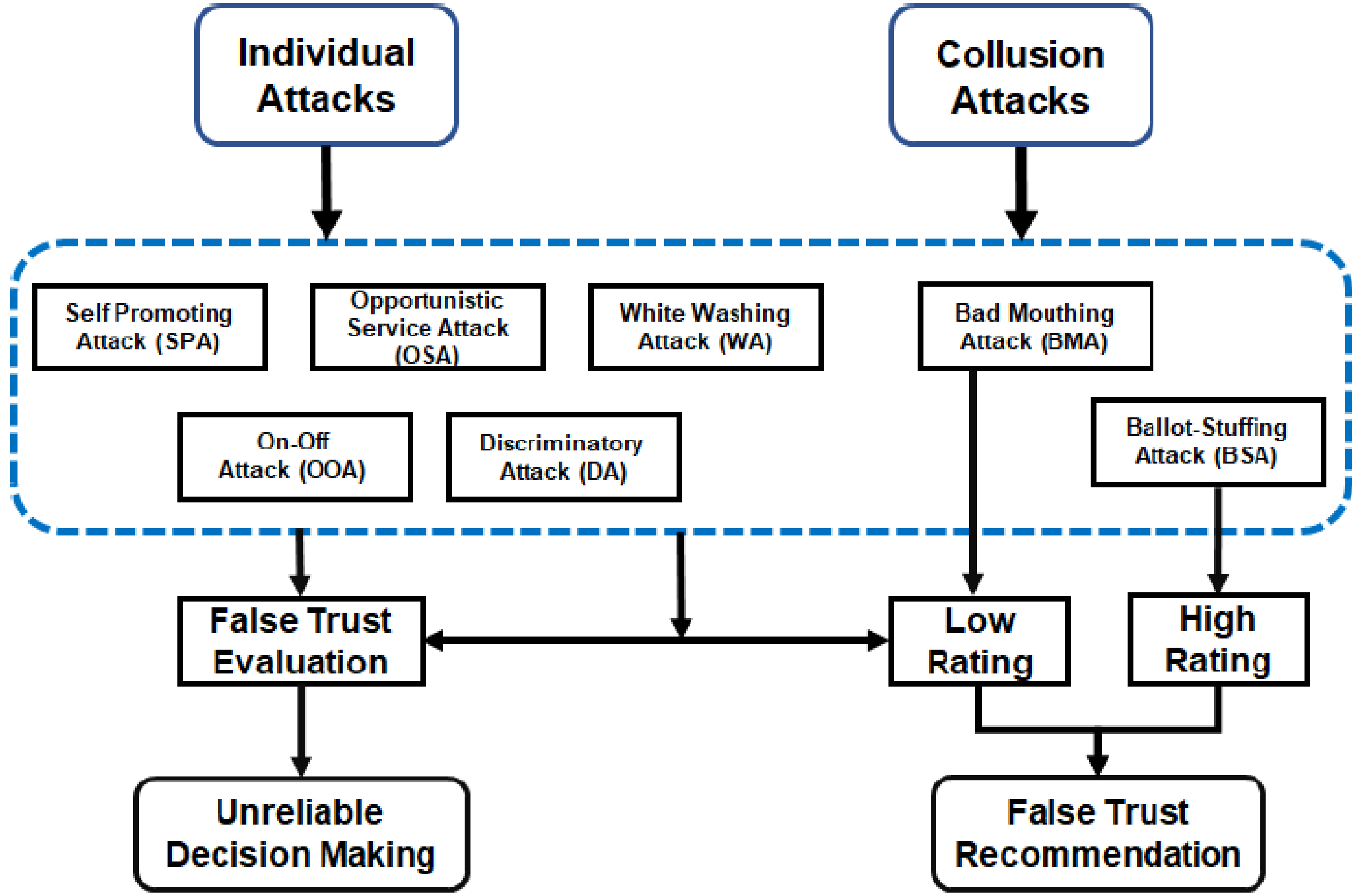
- The social gain is evaluated by multiple **ethical observers** which form a **virtual commission** that provides certified **accreditation**.
- The **EDI** can be customized to the digital asset which is under ethical evaluation, and it is **continuously updated** during the lifetime of its corresponding digital asset.
- The ethical identity has its **digital signature** and is updated according to **evidence** that supports the creation of a holistic perspective of its moral dimensions.
- The balanced perspective integrates both **supporting evidence** and **counter evidence**.
  - **Supporting evidence –** generated from successful evaluation scenarios
  - **Counter evidence –** generated from cases in which a particular component cannot be trusted

WHAT ARE THE CHALLENGES OF TRUST-MANAGEMENT SYSTEMS?

Challenges of Trust Management in IoE

- **Situational Scope of Trust:** high dependence of trust building on the context of trustor
- **Subjectivity of Trust:** influence by the factors inherent to the trustor (e.g. in taking risks)
- **Default Trust Score of New Agents:** on which trust score shall a new agent start
- **Trust Erosion:** trust score is subject to decay in case of no or too few interactions
- **Detection of Hidden Malicious Intentions:** hard to detect, likely to make mistakes in detection
- **Safety Assurance in Face of Untrusted Agents:** an ingredient of immune-response capability
- **Building Trust in the Trustworthy:** employing explanation to give evidence of trustworthiness
- **High Degree of Dynamism and Uncertainty in IoE:** possibly with missing information that is needed to make a decision, leading to misjudgment and bias

Trust Attacks


Individual Attacks
Collusion Attacks
Self Promoting Attack (SPA)
Opportunistic Service Attack (OSA)
White Washing Attack (WA)
Bad Mouthing Attack (BMA)
On-Off Attack (OOA)
Discriminatory Attack (DA)
Ballot-Stuffing Attack (BSA)
False Trust Evaluation
Low Rating
High Rating
Unreliable Decision Making
False Trust Recommendation

Trust Attacks – Individual Attacks

Individual Attacks refer to the attacks launched by an individual agent, which can take form of:

- **Self-Promoting Attacks:** an agent promotes its significance by providing good recommendation for itself so as to be selected as a service provider, and then acts maliciously.
- **Whitewashing Attacks:** an agent exits and re-joins the ecosystem to recover its reputation and to wash-away its own bad reputation.
- **Discriminatory Attacks:** an agent explicitly attacks other agents that do not have common friends with it, i.e. it performs well for a particular service/agent and badly for some other services/agents.
- **Opportunistic Service Attacks:** an agent might offer a great service to improve its reputation when its reputation falls too low.
- **On-Off Attacks:** an agent provides good and bad services on and off (randomly) to avoid being labeled as a low-reputed agent.

Trust Related Attacks – Collusion-based Attacks

Collusion-based Attacks represent the attacks launched by a group of agents to either provide a high rating or low rating to a particular agent, such as:

- **Bad-Mouthing Attacks:** In this type of attack, a group of agents diminishes the reputation of a trustworthy agent within the ecosystem by providing bad recommendations about it.
- **Ballot-Stuffing Attacks:** In this type of attack, a group of agents boosts the reputation of bad agents within the ecosystem by providing good recommendations for them.

THANK YOU

Thank you for your attention

Czech CyberCrime Centre of Excellence C4e

- A multidisciplinary center that brings together expert academic departments to address complex cyberspace problems

MUNI

MUNI ICS

MUNI FI

MUNI LAW

NÚKIB

CONCORDIA

Cyber Security for Europe

National Cybersecurity R&D Laboratory

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MŠMT
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

FACULTAS ARTIS INFORMATICÆ • UNIVERSITAS MASARYKIANA •

Barbora Buhnova, FI MU Brno
buhnova@fi.muni.cz
www.fi.muni.cz/~buhnova