| galois |

A Commitment Logic for Reasoning about Trust in Complex Systems

David Burke

HCSS 2017

0. Overview

The demand for certainty is one which is natural to man, but is nevertheless an intellectual vice.

Bertrand Russell

Overview

Synopsis: When reasoning about trust in complex system, design for a clean separation between behavioral claims by one entity/agent, and the trust granted by other entity/agent as a result of this claim.

Another way of thinking about this approach: obligations versus offers/promises/commitments/ claims.

Table of Contents

1. Obligations

I have always thought music as a way out of the mundane obligations of life.

Martha Reeve

Deontic Logic

Deontic logic is a member of the class of modal logics.

Obligation (O) and *Permission* (Ƥ) are the key concepts in deontic logic.

O And Ƥ can be defined (elegantly!) in terms of each other:

$$O(x) = \sim Ƥ(\sim x)$$

$$Ƥ(x) = \sim O(\sim x)$$

Obligation Dynamic

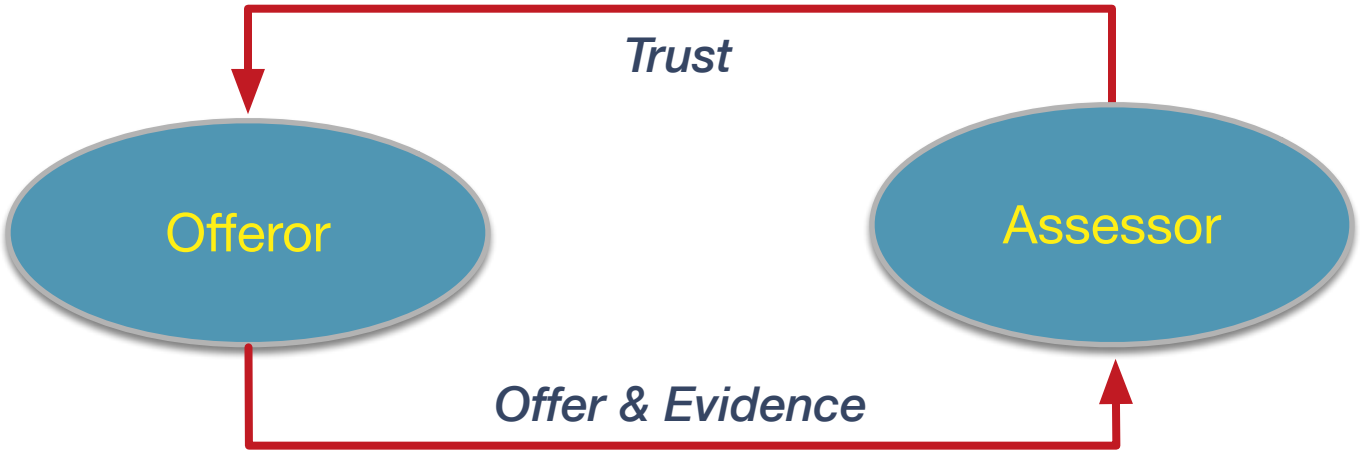
- This dynamic appropriate if you have enforcement: either the sender, or a third party.
- Not necessarily a good fit for distributed systems.

2. Offers

Promises are the uniquely human way of ordering the future, making it predictable and reliable to the extent that this is humanly possible.

Hannah Arendt

Fundamental Dynamic



Offeror: Responsible for describing offer and supporting evidence.

Assessor: Responsible for the decision to grant trust; takes action in accordance with that assessment. The assessor can bring additional evidence to bear in this assessment (not limited to offeror's exhibits).

Brief Historical Interlude

- These ideas have been around…
- For example, Contract Theory is a branch of moral theory that attempts to make sense of morality by reducing it to agreements between parties.
  - "Thou shalt not steal" becomes instantiated by "I'll pay you back the money I owe you".
  - Common meta-rule: *you promise to keep all your promises*.

3. Evidence

A wise machine proportions its belief to the evidence.

(with apologies to David Hume)

Evidence Classes

1. ***Direct Evidence*** – credible, direct testimony coming from a shared domain and ontology.
2. ***Indirect*** – evidence that is not directly derived from or associated with the current offer, but is deemed relevant for the assessor.

Direct Evidence

Testimony from the offeror that is directly related to the offer.

"The confidentiality of the transmissions will be protected by end-to-end AES-256 encryption."

"The hotel elevator has been built and tested to a load of 3000kg or 13 persons in accordance with the ASME A17.1-2016/CSA B44-16 standards."

"The engine is capable of providing a minimum of 365 horsepower over the required 3-hour period of performance."

This class of evidence requires a shared ontology between offeror and assessor.

Indirect Evidence

Indirect evidence is an observational pointer to existing evidence, which can be direct or indirect.

1. Reciprocity
2. Social Proof
3. Consistency & Commitment
4. Authority
5. Liking (Similarity)
6. Scarcity (Abundance)

4. Assessment

Never trust anything that can think for itself if you can't see where it keeps its brain.

J.K. Rowling

Trust

Voluminous literature on the subject. And nobody seems to agree on an exact definition for the term.

- The notion of delegation?
- A dependency or vulnerability on another entity?
- A cognitive shortcut?

Distinction between competence and willingness.

What about transitivity?

We're going to be pragmatic and operational: trust is a decision made by an entity to take action.

How Assessments Work

Each node/entity/agent is responsible for determining:

1. Classes of evidence it will consider as inputs
2. Aggregation methods
3. Actions it is willing to take as a result of the assessment

Evidence Classes for Assessments

- Direct Evidence
- Indirect Evidence
- Contextual Evidence

Assessing Indirect Evidence

Assessing Contextual Evidence

Agents can choose to explicitly weigh the context surrounding the decision to grant trust:

- Pragmatism: it serves my current interests to trust/not-trust.
- Risk: The risk in trust/not-trusting is too high.
- Consequences: a potential outcome is too important/horrible, and should dominate the trust-granting process.

Aggregating Evidence - Dimensions

- Dimensions of a Single Piece of Evidence
  - Probative Force
  - Relevance
  - Confidence

Aggregation Strategies

We're running experiments with two aggregation strategies:

- Social Choice Theory (SCT)
- Heuristics

Social Choice Theory (SCT)

SCT refers to a means by which individual opinions, preferences, or judgments are aggregated into a collective decision.

The reasons that SCT is relevant to the problem at hand is twofold:

1. Many assessments depend on explicitly fusing evidence from other agents; SCT gives guidance on how to do this fusion/ aggregation.
2. More generally, every piece of evidence is weighing in on the question “Do I grant trust?” or “What action do I take given the level of trust I’m willing to grant?” We can treat each piece of evidence as weighing in with a preference or judgment.

SCT Judgment Paradox

	P	P → Q	Q
Judge 1	True	True	True
Judge 2	True	False	False
Judge 3	False	True	False
Majority	True	True	False

Each judge uses their judgments of the premises P, P→Q and propositional logic to draw conclusions about Q.

However, the majority vote has both P and P→ Q both being true, and Q as false.

We're looking at extensions to traditional SCT to avoid these paradoxes.

Heuristics

- Conventional Wisdom: More information is always better, full information is best. More computation is always better, optimization is best.
- A heuristic is a strategy that ignores available information; it focuses on a few key cues.
- In order to make good decisions under uncertainty, humans (and machines) should ignore part of the available information.
- "Satisficing", not optimizing – "fast and frugal" heuristics.
- *Humans very often use the heuristic "Make the decision that can most easily be justified/explained to others".*

5. Implementation

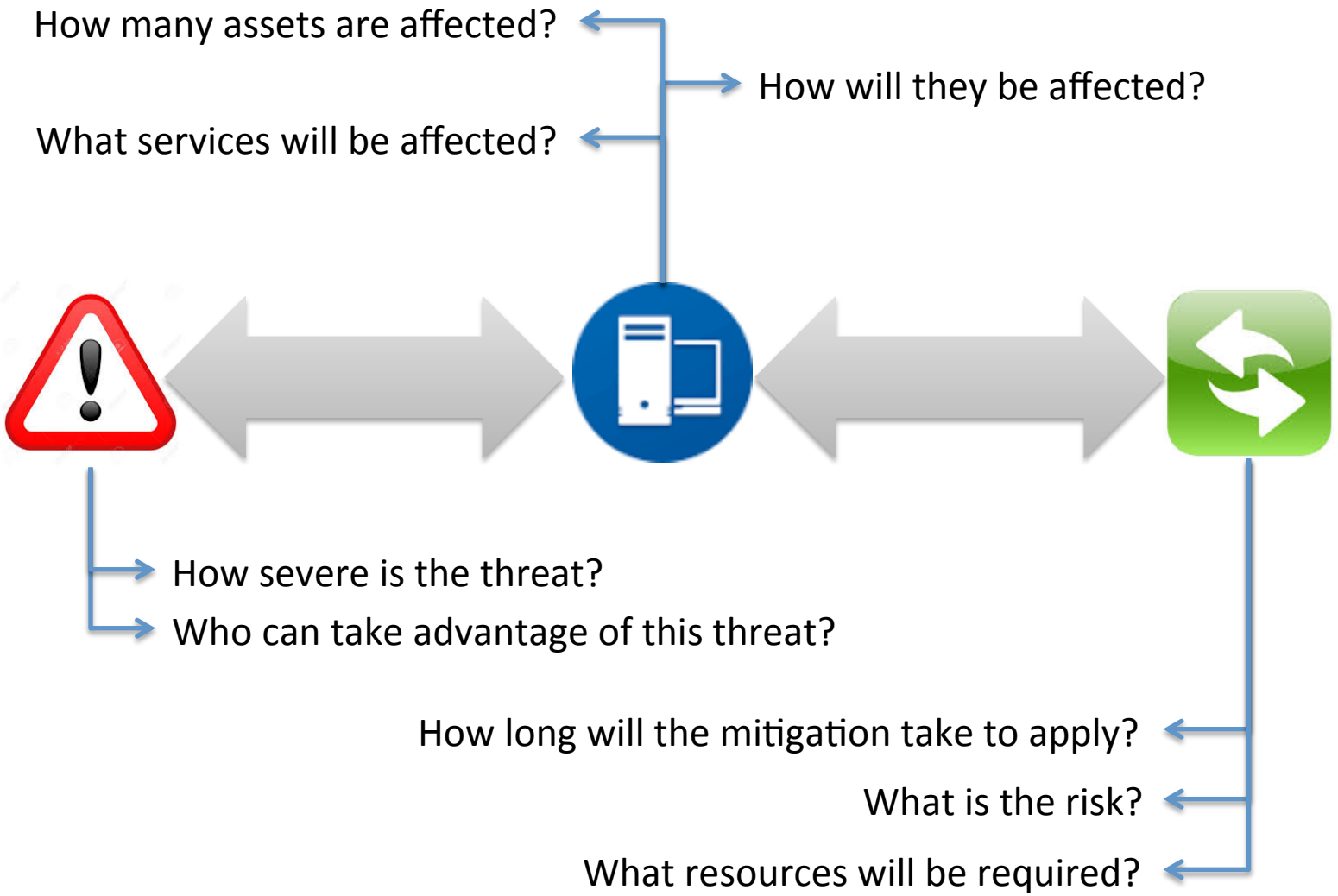
The way to build a complex system that works is to build it from very simple systems that work.

Kevin Kelly

Relevant Domain – Cyber Analysts

- Cyber analysts make decisions under conditions of:
  - Many, diverse inputs
  - Time pressure
  - Dynamic environment
  - Consquences matter
- Analyst attention is the most limited resource; how do analysts allocate their time?
- Which information sources (human and machine) can be trusted?

TFER as Validation Testbed



Message Structure

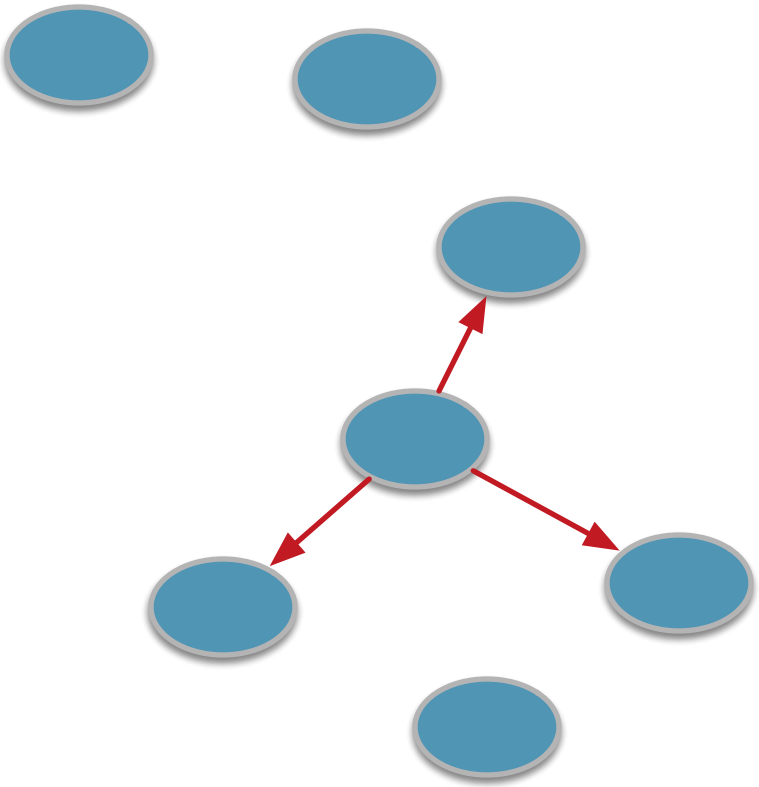
Offeror:

{entity_id, claim_id, claim, [evidence, dim, conf]}

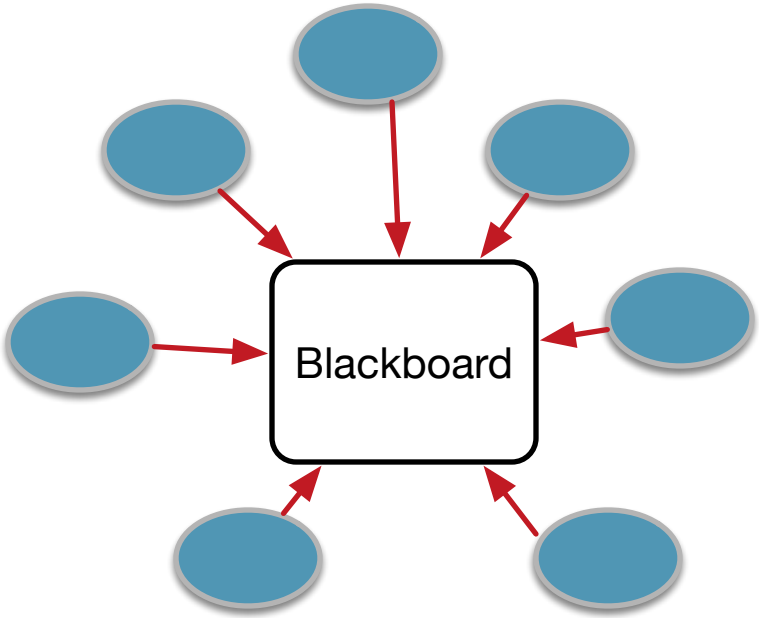
Assessor:

{entity_id, claim_id, trust_level, [justifications]}

Candidate Architectures



Localized Communication



Centralized Communication

6. Conclusions

As you deal with more and more complex systems, it becomes harder and harder to find deep and interesting properties.

Noam Chomsky

Emerging Design Principles

- Agnostic to whether the entities are humans, machines, or human-machine ensembles
- Agnostic to level of abstraction in a machine – no single level is privileged.
- Pertinent to modern distributed systems with increasing amounts of autonomy.
- Expressiveness – diverse classes of evidence
- Transparency & Justification

Questions

- Adversarial actions – how can adversaries game the system?
- How to incorporate time – staleness of evidence & assessments?
- Dealing with features of complex systems
  1. Emergent properties
  2. Lack of ergodicity
  3. Radical uncertainty (i.e. “black swans”)
  4. Computational irreducibility

| galois |

David Burke
davidb@galois.com
503-330-9512

Backups &
Scrap Material

Majority Grade Example

Wine	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
St. Amour	Very Good	Passable	Good	Very Good	Good
Bourgueil	Good	Very Good	Good	Mediocre	Excellent
Cahors	Mediocre	Excellent	Excellent	Passable	Good

Put each wine's grades into lexicographic order

1st - St. Amour	2nd - Bourgueil	3rd - Cahors
Very Good	Excellent	Excellent
Very Good	Very Good	Excellent
Good	**Good**	**Good**
Good	Good	**Passable ← 3rd place**
Passable	**Mediocre ← 2nd place**	Mediocre

Using the taxonomy:

Ex2: bomb-disposal robot

- High on competence
- High on consequence
- High on likability
- Medium on loyalty
- Medium on reciprocity
- Medium on authority

etc.

Ex1: Airplane Autopilot

- High on competence
- High on consequence
- High on social proof
- Medium on authority
- Low on likability
- Low on reciprocity

etc.

Dimensions of Human-Machine Trust-Granting

Individual Dimension

Predisposition **- the individual's general tendency (perhaps partially genetic) to grant trust to others.**

Commitment/Consistency **- an assessment of the person's commitment to interacting with ASes; consistency with prior engagements**

Relationship Dimension

Reciprocity **- an assessment of what benefits the human has received from the AS in the past; what they believe they owe in return.**

Likability **- the emotional valence of the relationship between the human and the AS.**

Loyalty **- an assessment of how strongly the human feels that the AS is "part of the team" and therefore deserves support.**

Organizational Dimension

Social Proof **- examples of how successful ASes have been in similar situations; endorsements.**

Authority **- the amount of 'top-down' or policy support for the human/AS teaming by authority figures.**

Environmental Dimension

Competence **- the human's judgment of the competence of the AS for the task at hand.**

Consequence **- An assessment as to what is at risk in the current scenario.**

Willingness **- A belief that the AS is willing to assume the trust granted, and to take action.**

I have always thought music as a way out of the mundane obligations of life.

Martha Reeve

As you deal with more and more complex systems, it becomes harder and harder to find deep and interesting properties.

Noam Chomsky

It is wrong always, everywhere, and for everyone, to believe anything upon insufficient evidence.

William James

The way to build a complex system that works is to build it from very simple systems that work.

Kevin Kelly

Never trust anything that can think for itself if you can't see where it keeps its brain.

J.K. Rowling

A wise machine proportions its belief to the evidence.

(with apologies to David Hume)

Promises are the uniquely human way of ordering the future, making it predictable and reliable to the extent that this is humanly possible.

Hannah Arendt