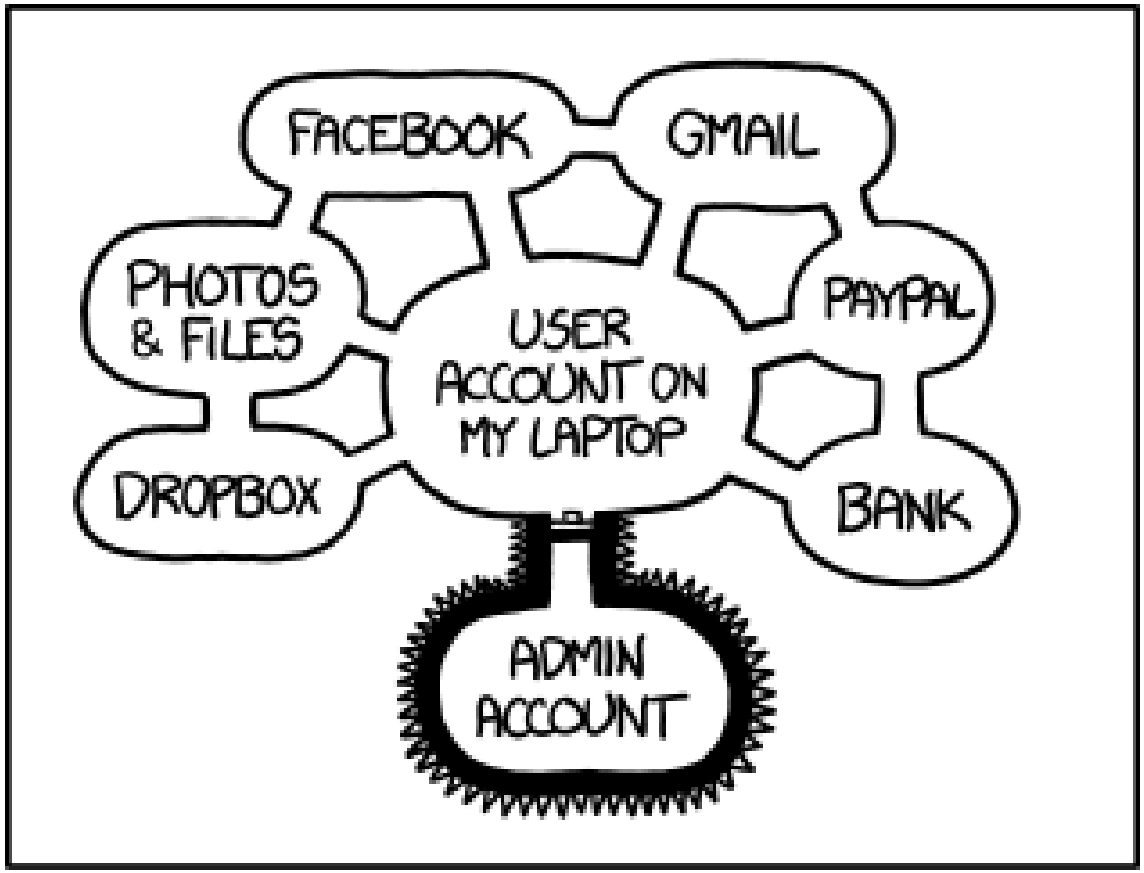
Toward New Foundations for Computational Trust

Munindar P. Singh
singh@ncsu.edu

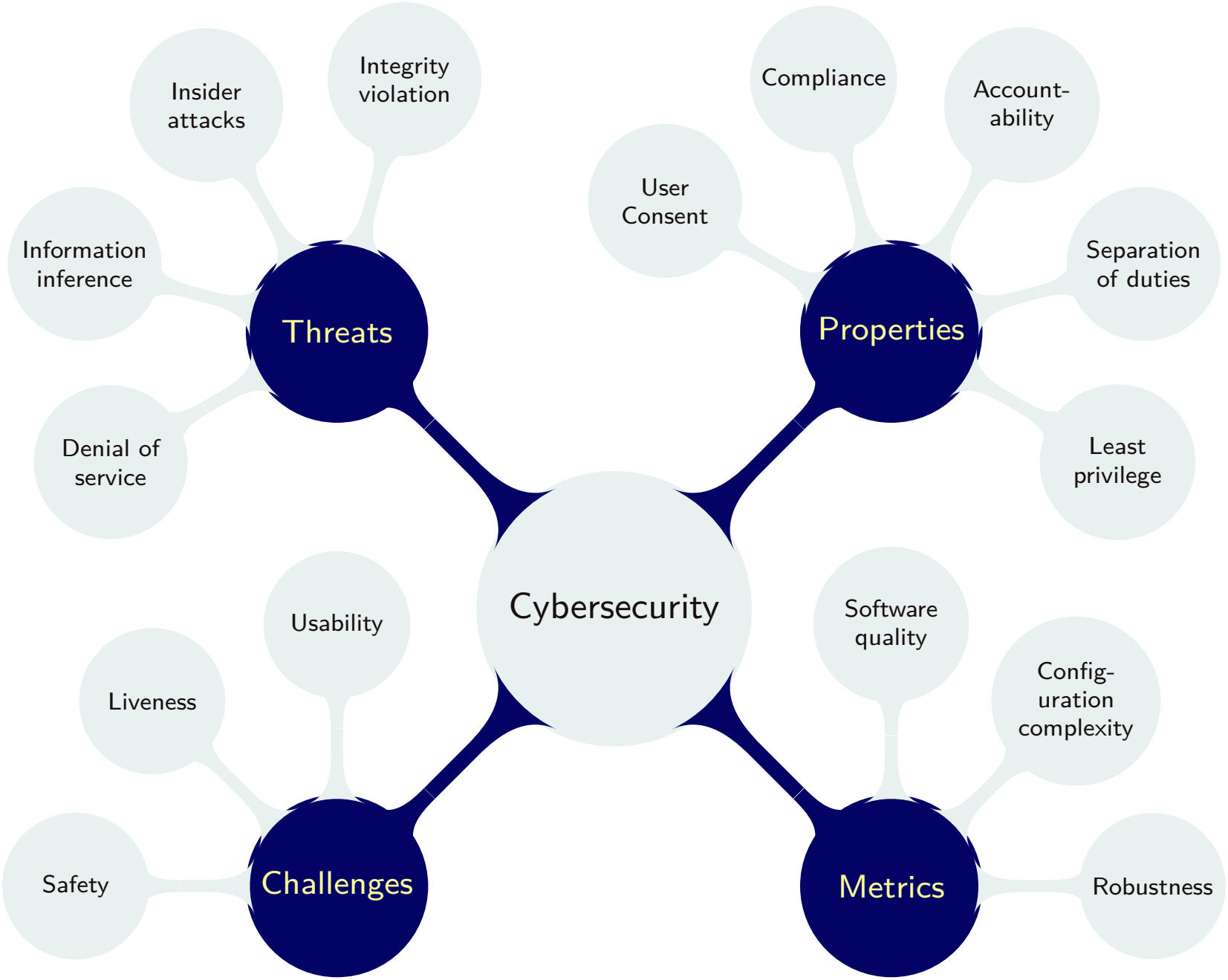
Department of Computer Science
North Carolina State University

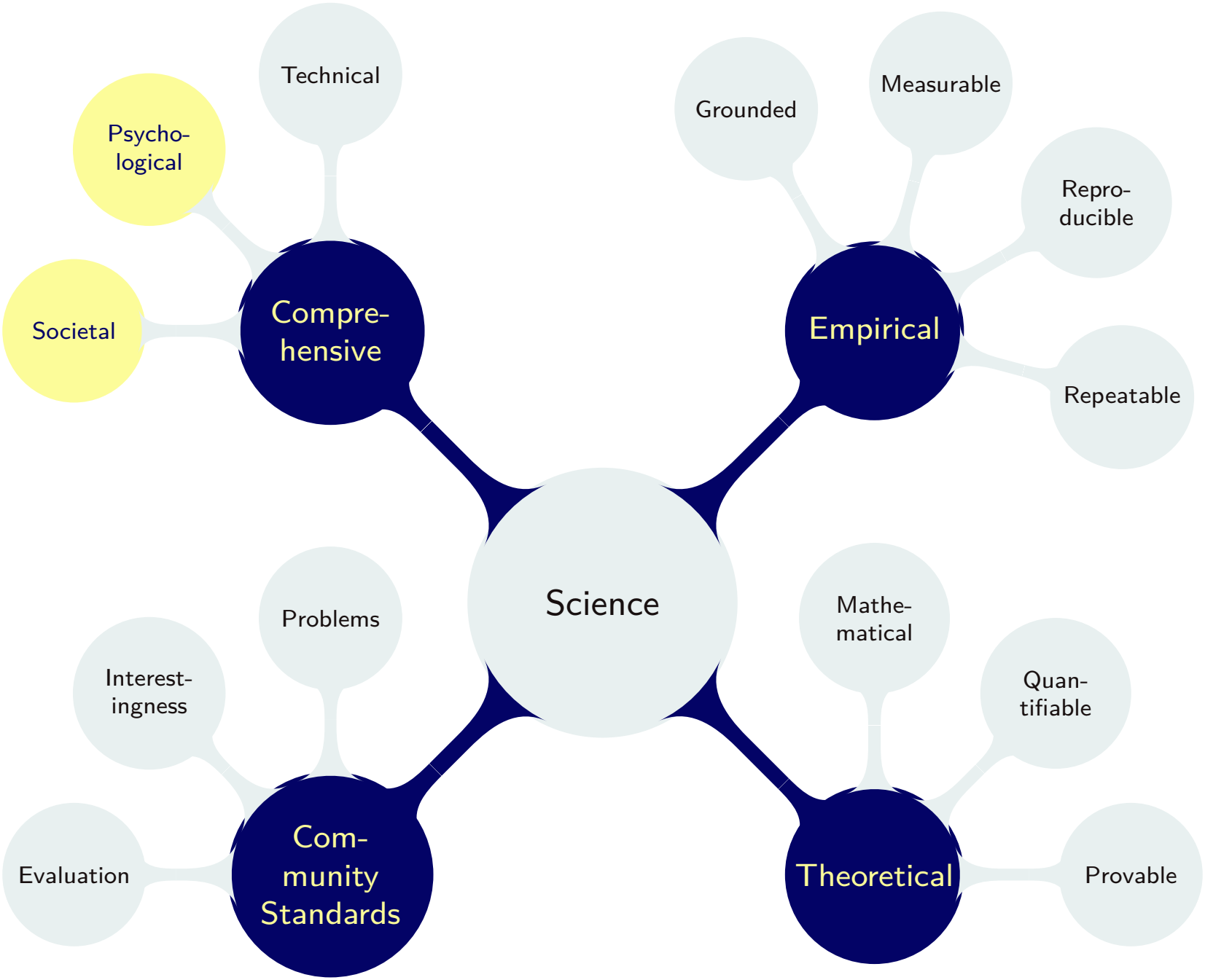
XKCD's Assessment of Security Today

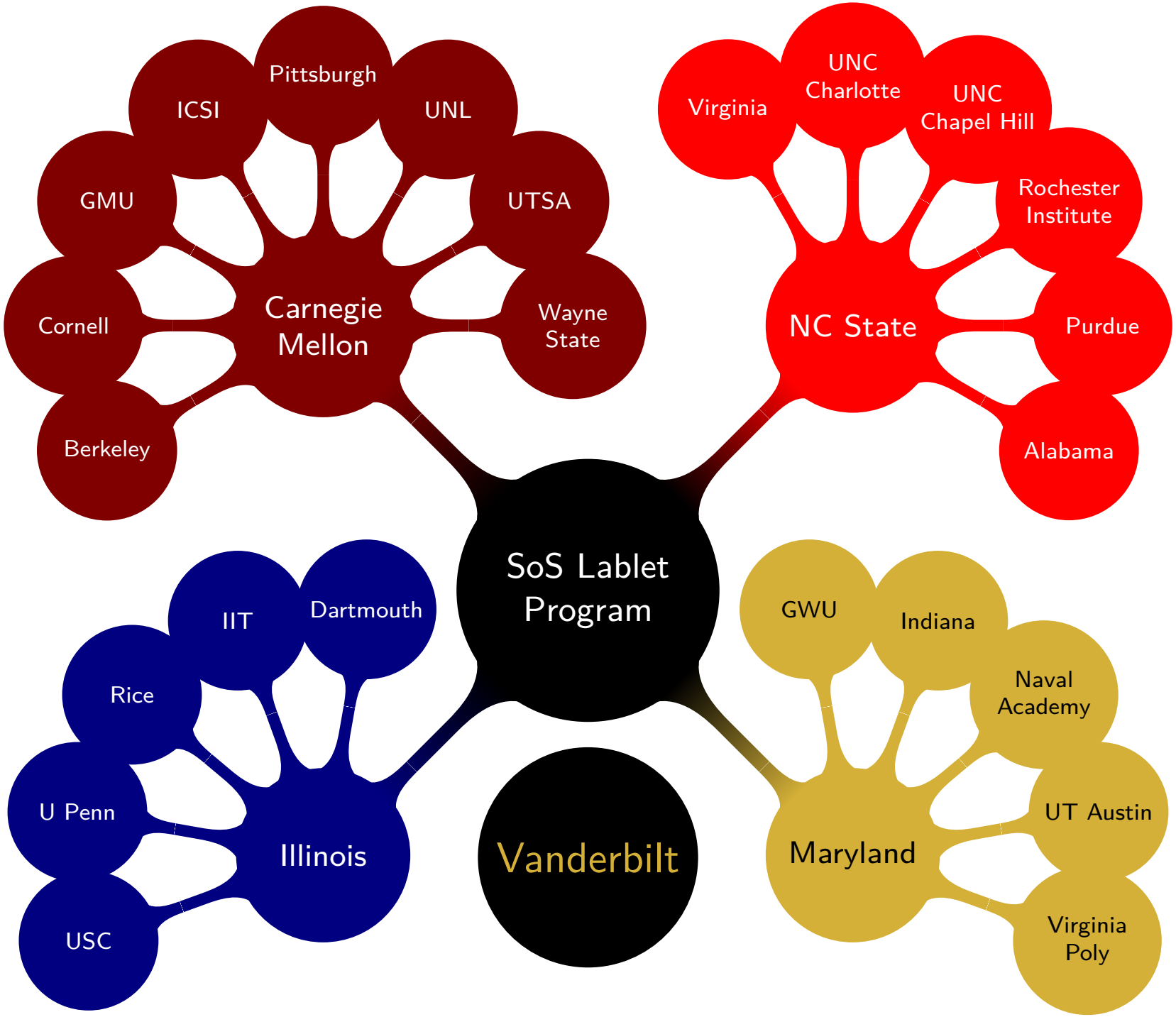


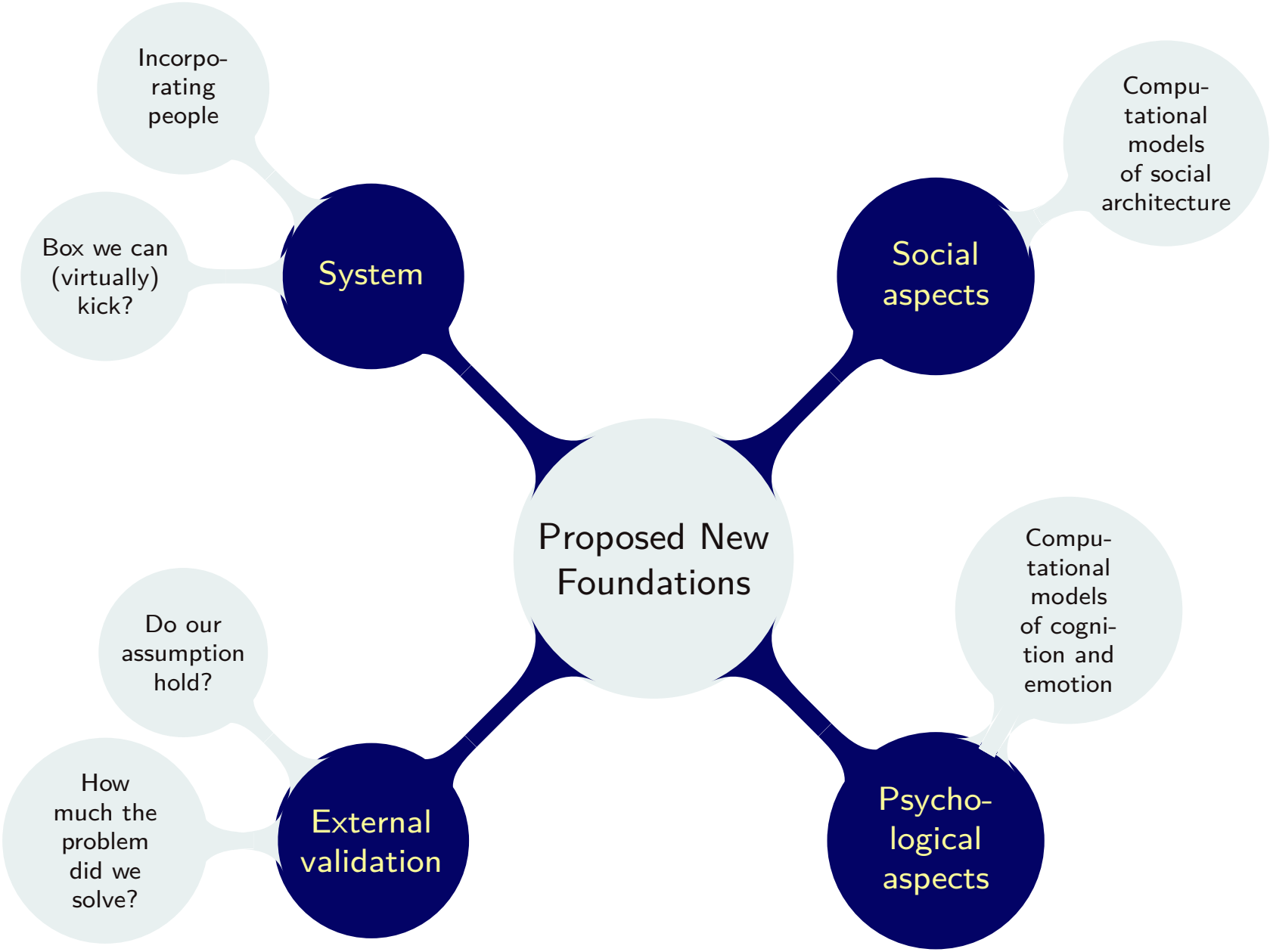
IF SOMEONE STEALS MY LAPTOP WHILE I'M LOGGED IN, THEY CAN READ MY EMAIL, TAKE MY MONEY, AND IMPERSONATE ME TO MY FRIENDS,

BUT AT LEAST THEY CAN'T INSTALL DRIVERS WITHOUT MY PERMISSION.


Cybersecurity
Threats
Insider attacks
Integrity violation
Information inference
Denial of service
Properties
Compliance
Account-ability
User Consent
Separation of duties
Least privilege
Challenges
Usability
Liveness
Safety
Metrics
Software quality
Config-uration complexity
Robustness

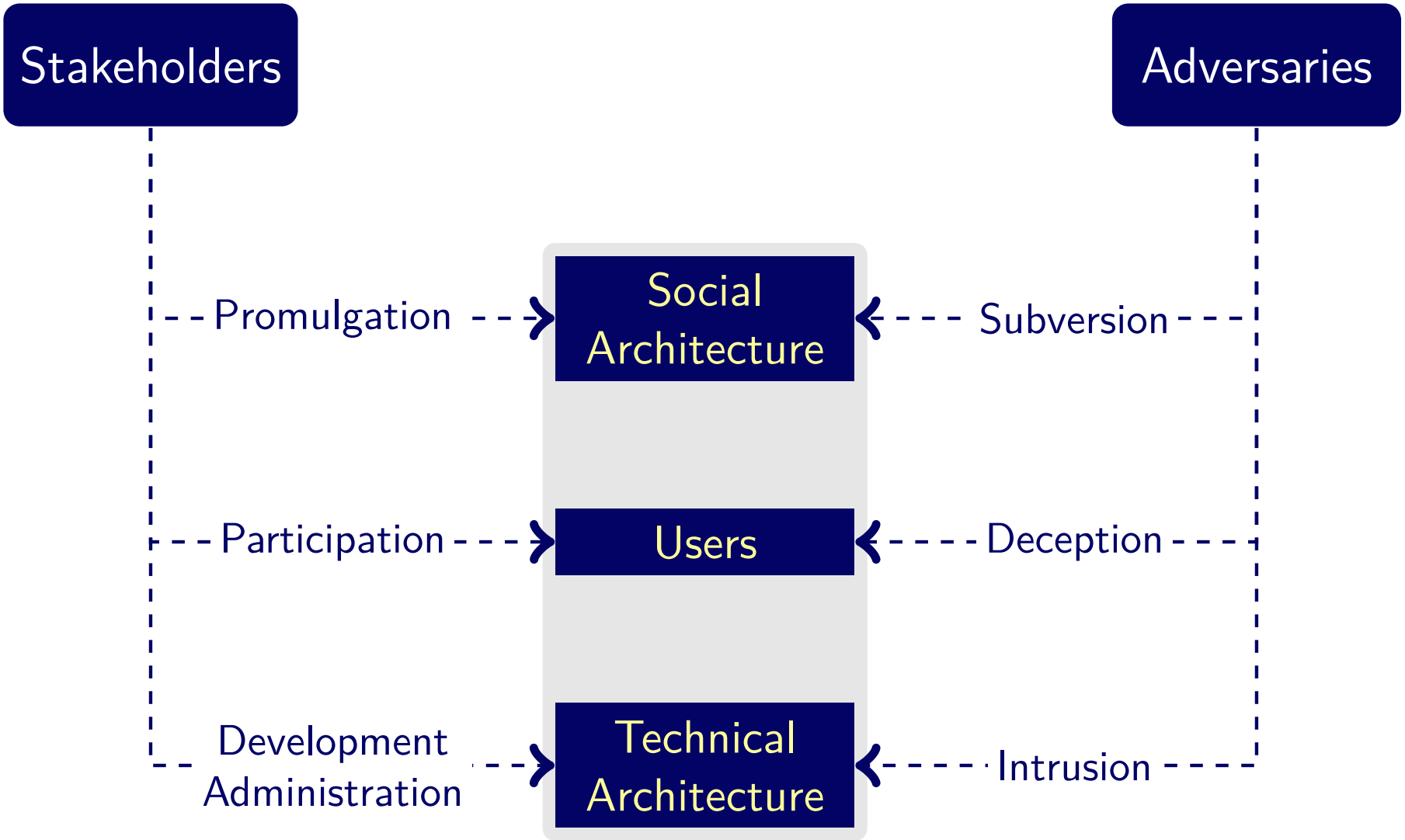

Science
Comprehensive
Technical
Psychological
Societal
Empirical
Grounded
Measurable
Reproducible
Repeatable
Community Standards
Problems
Interestingness
Evaluation
Theoretical
Mathematical
Quantifiable
Provable


SoS Lablet Program
Carnegie Mellon
Pittsburgh
ICSI
UNL
GMU
UTSA
Cornell
Wayne State
Berkeley
NC State
UNC Charlotte
Virginia
UNC Chapel Hill
Rochester Institute
Purdue
Alabama
Illinois
IIT
Dartmouth
Rice
U Penn
USC
Vanderbilt
Maryland
GWU
Indiana
Naval Academy
UT Austin
Virginia Poly


Proposed New Foundations
System
Incorporating people
Box we can (virtually) kick?
Social aspects
Computational models of social architecture
External validation
Do our assumption hold?
How much the problem did we solve?
Psychological aspects
Computational models of cognition and emotion

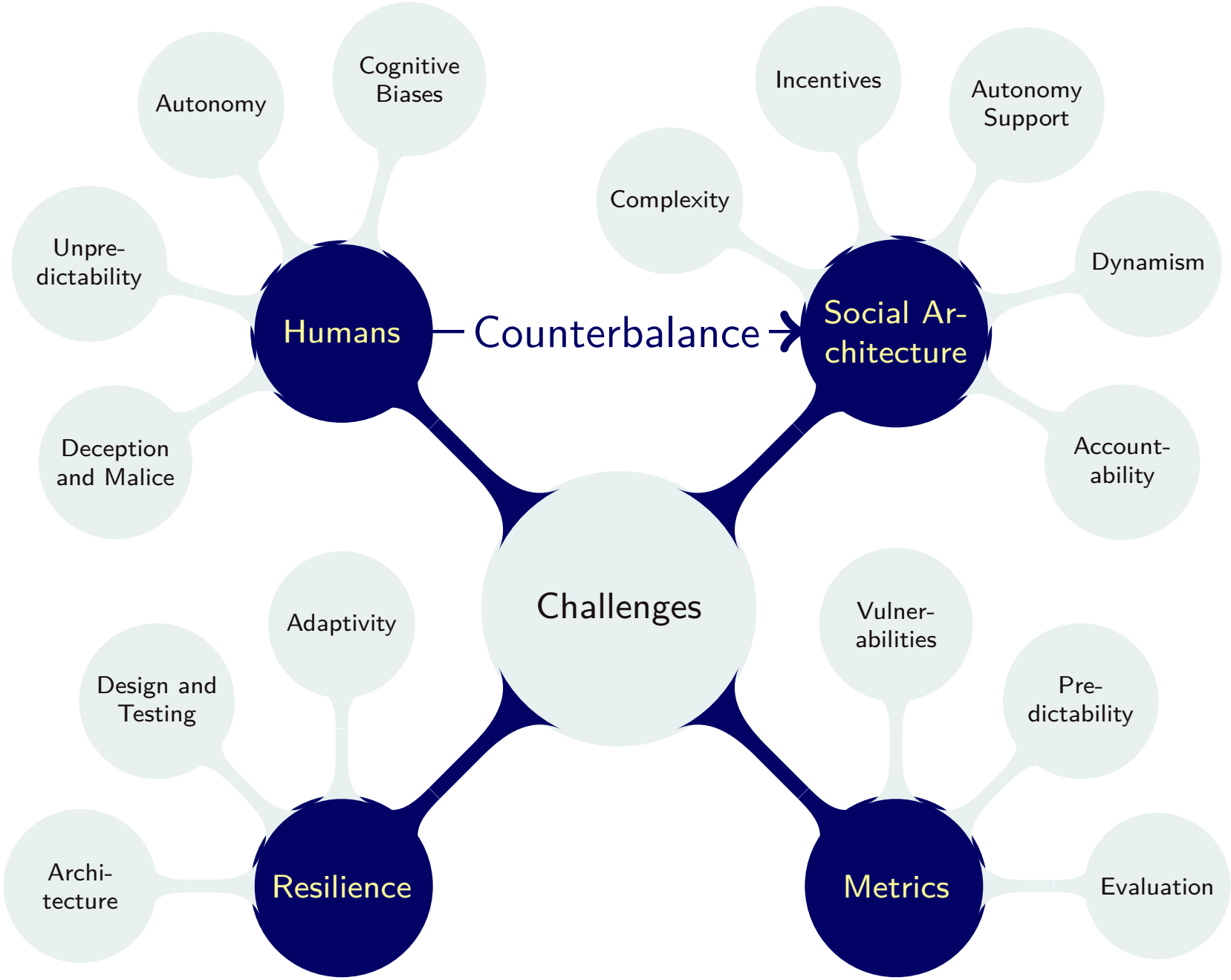
Participants and Artifacts in Security

Greatest challenges arise in the upper two; most past effort is on technical architecture



Usability and Strange User Behavior

Can we protect users from themselves? Can we validate our approaches in real environments?

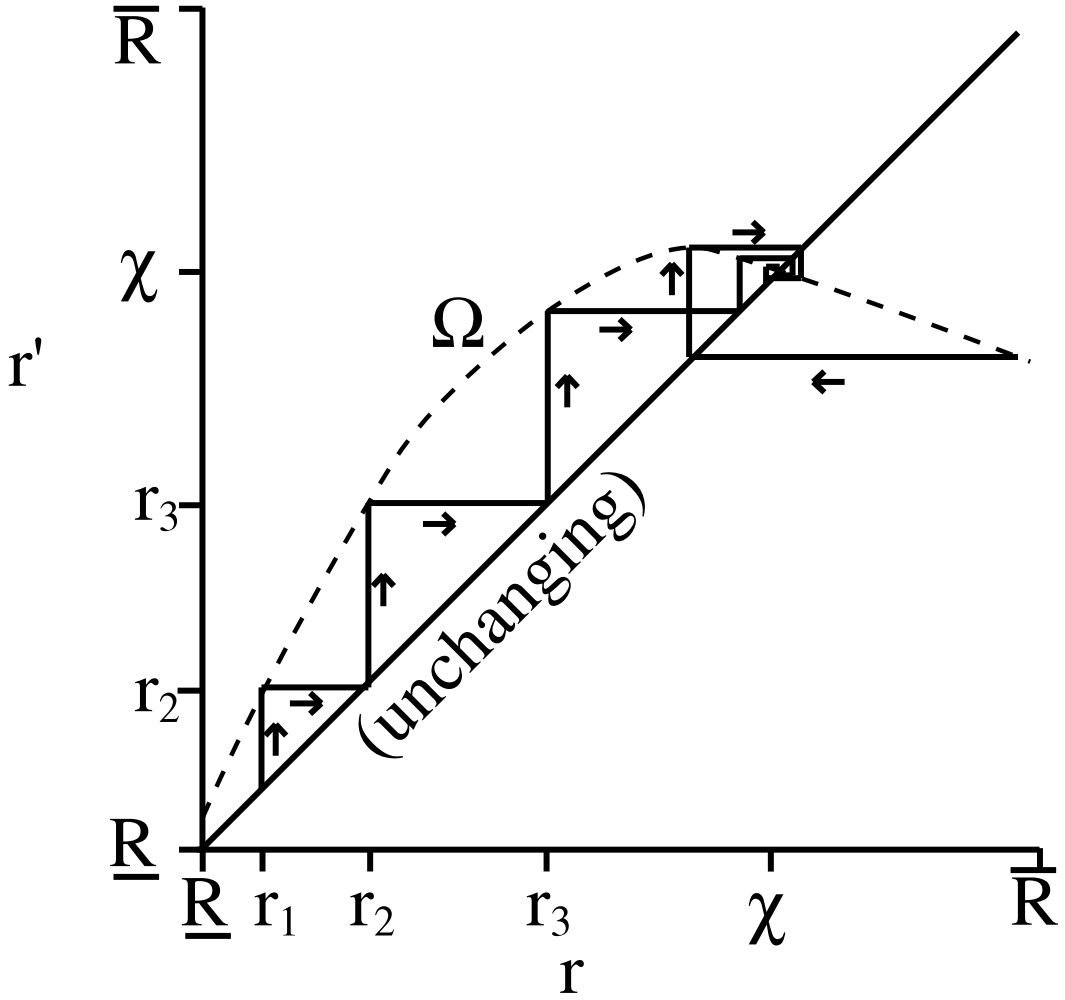

Autonomy
Cognitive Biases
Unpre-dictability
Humans
Deception and Malice
Counterbalance
Incentives
Autonomy Support
Complexity
Dynamism
Social Architecture
Account-ability
Challenges
Adaptivity
Design and Testing
Archi-tecture
Resilience
Vulner-abilities
Pre-dictability
Metrics
Evaluation

Understanding Economic Foundations of Trust

Focus on dynamical properties

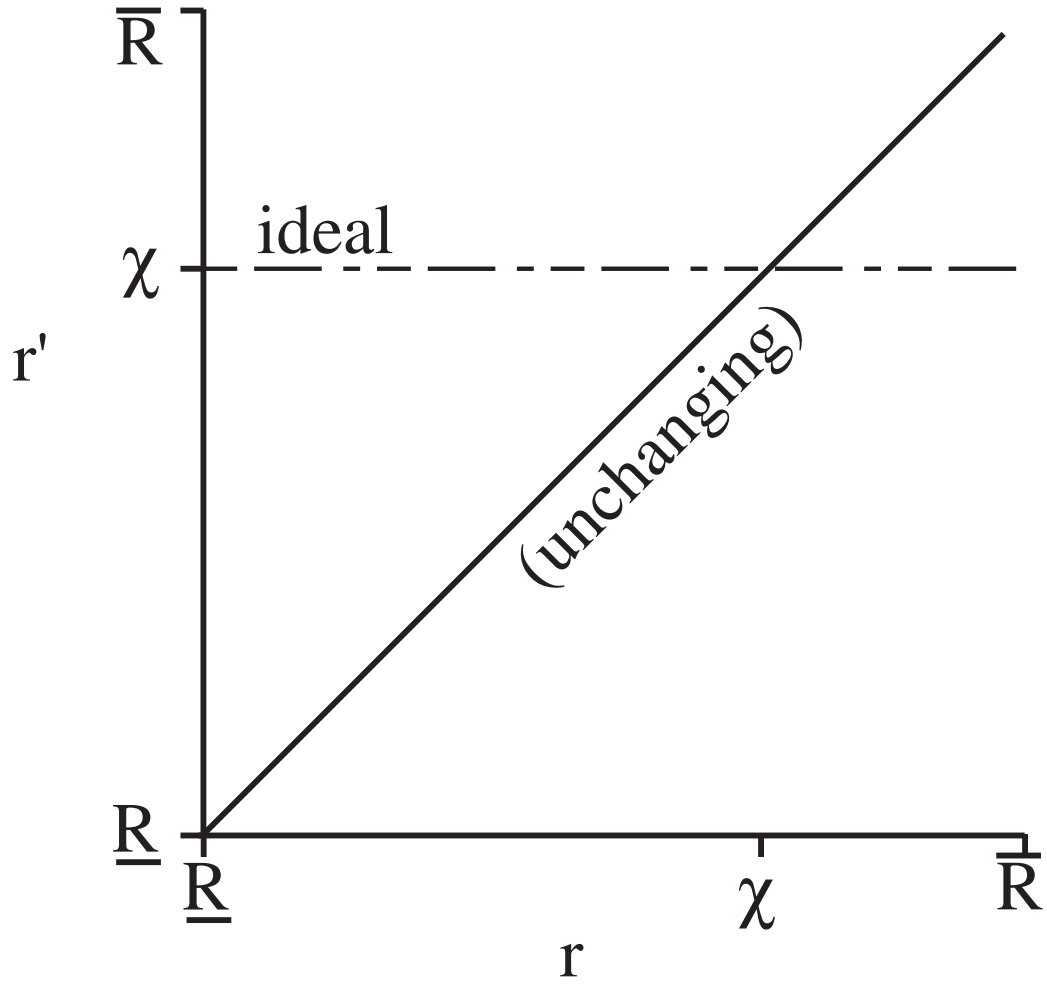
- Axioms of trustworthiness of trustee judged via truster's utility
  - Scalar, e.g., $\in [0, 1]$
  - Strength: Trustee willing to do more $\Rightarrow$ Willing to do less
  - Comparison: Trustee with greater effort $\Rightarrow$ More trustworthy
  - Stability: Preferences of all are stable if time shifted
- Understand reputation models as mechanisms, which
  - Govern agents' behaviors
  - Can be analyzed as dynamical systems
- Technical properties of dynamical systems
  - Criteria for comparing systems
  - Contrast with traditional, anecdotal evaluations

Dynamic Reputation Graphs

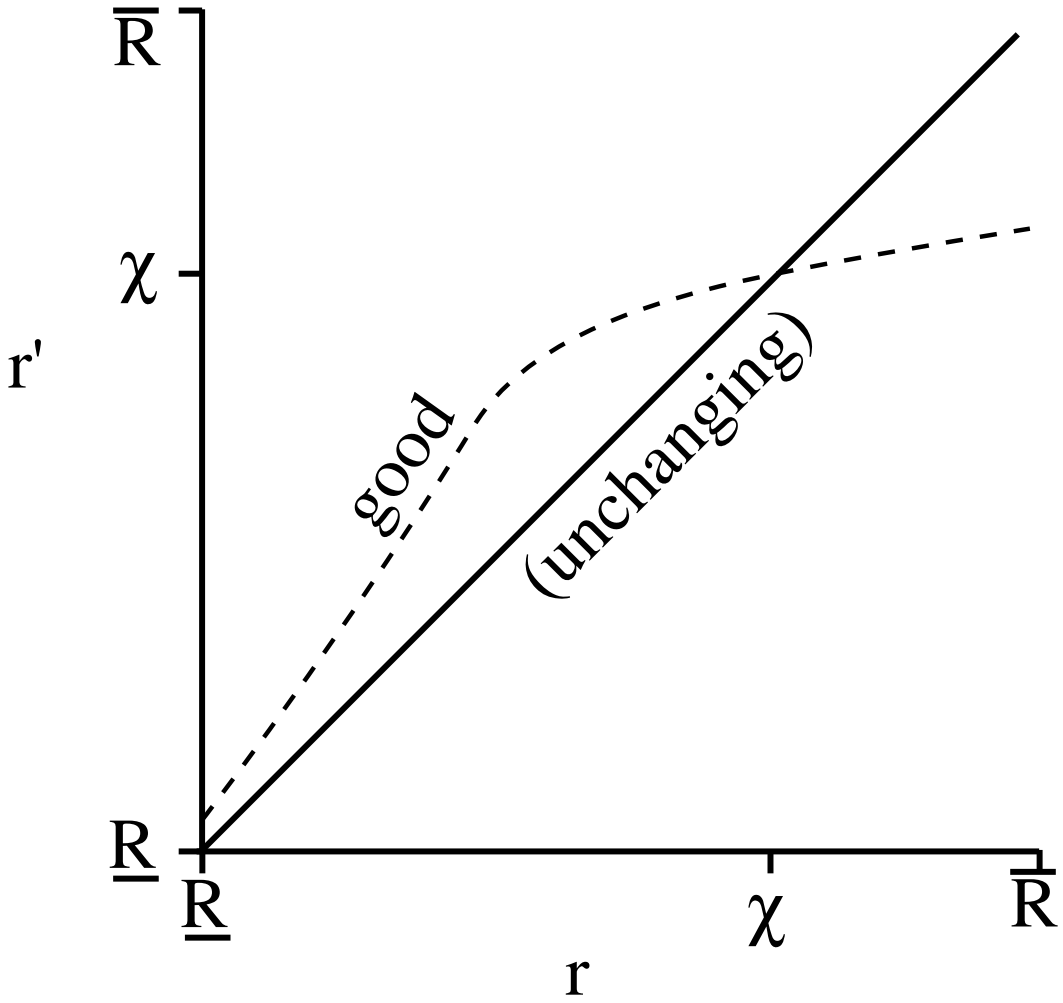


- *Update function:* next reputation after action
- *Payoff function:* reward for action given reputation
- Agent utility: function of reputation
- Agent type: $\theta \in \Theta$
- Current reputation (projection): $r \in R$
- Next reputation $r' = \Omega_\theta(r)$
- Fixed point: $\chi(\theta) = \lim_{n\to\infty} \Omega_\theta^n(r_{\text{initial}})$

Ideal Trust System



Good Trust System

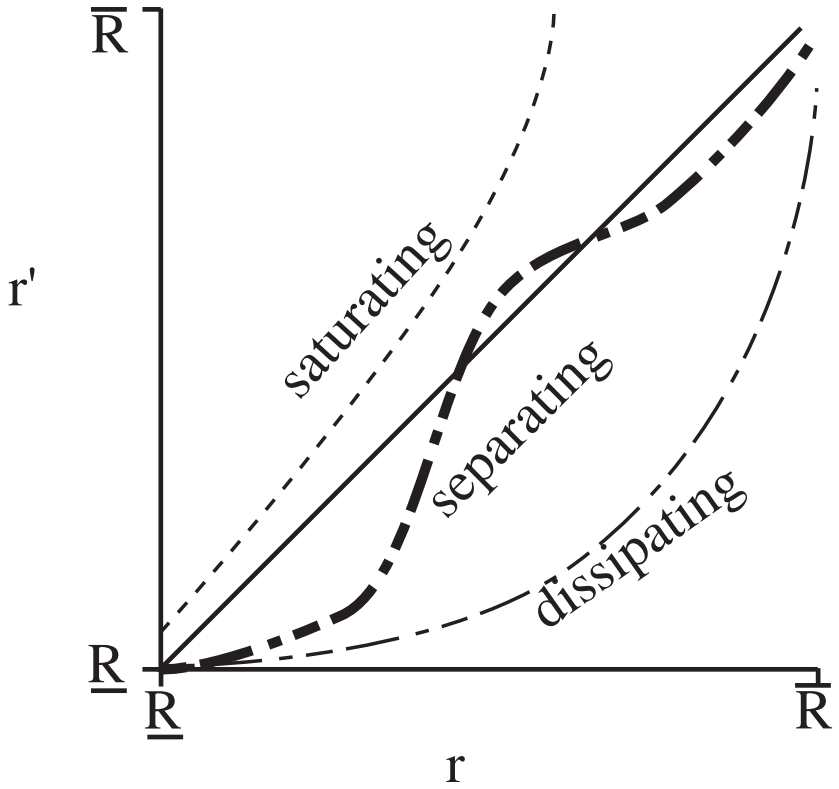


Trust System Metric: Monotonicity

- *Ideally Patient Strategic* (IPS) agent
  - Infinite horizon, maximize utility
- If θ_a is weakly preferable to θ_b to IPS agent $\Rightarrow$ asymptotically, a's reputation $\geq$ b's reputation

Trust System Metric: Unambiguity

Each agent type asymptotically maps to a single reputation value

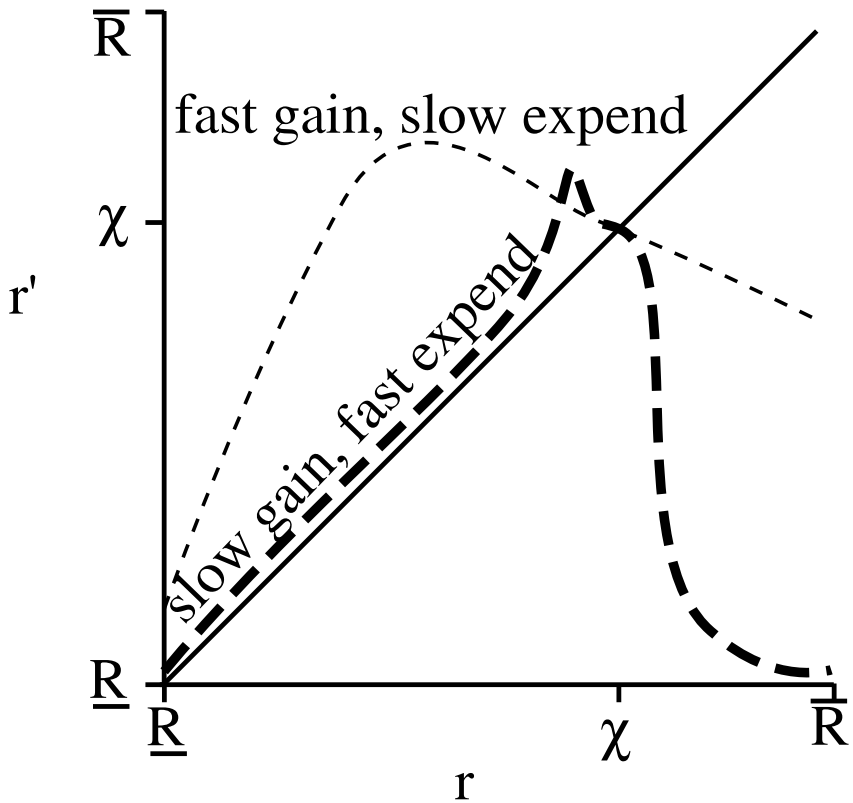


Ambiguous Trust Systems

Trust System Metric: Accuracy

Minimize average error over distribution of types

- Error: absolute distance from ideal reputation
- Reputation system performance when beliefs far from fixed point

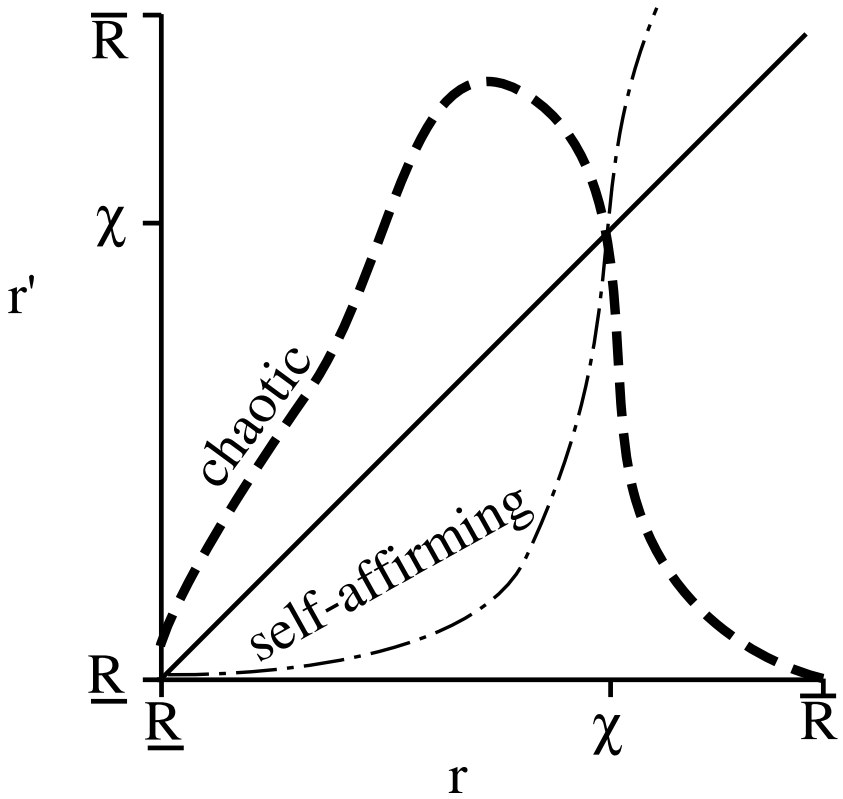


Differing Accuracy

Trust System Metric: Convergence

Reputation should converge quickly near the fixed point

- Max component of gradient: $||\nabla\Omega(r)||_\infty < 1$ and minimized
- Lyapunov stability may be acceptable



Divergent Trust Systems

Simulation Study Design: Real Approaches; Streamlined Interventions

Exchange of favors between a trustee and an infinitely patient truster

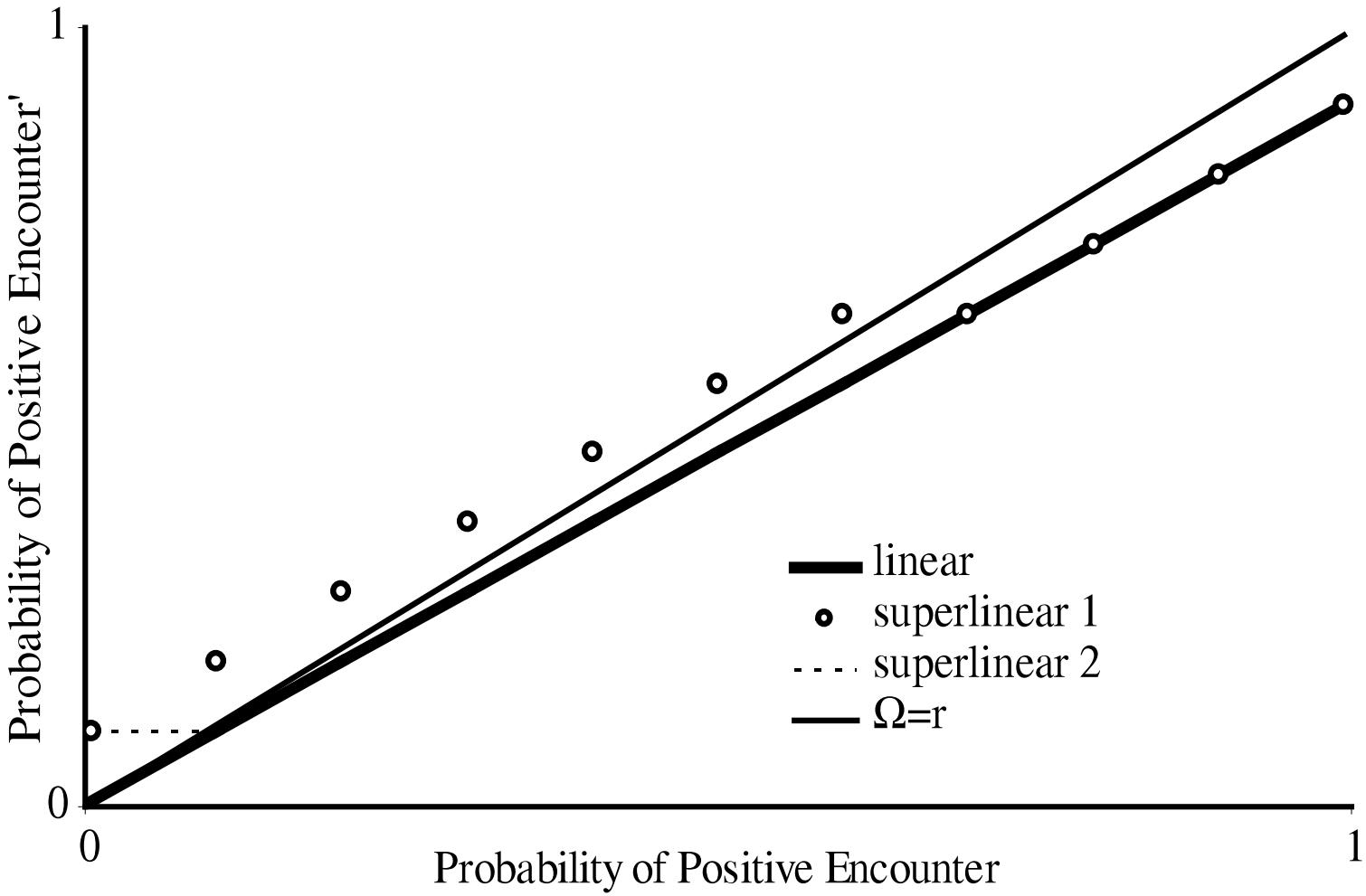
- Trustee offers a favor to truster
  - Small cost [1, 12] to offerer
  - High benefit [10, 30] to offerer
- Reverse roles
- Target's strategy: finite series of binary decisions
  - Bounded to 95% of total utility over infinite horizon
- Infinite patience of truster: hence, perfect trustee won't defect
- Use expected payoffs to determine best response strategies to compute
  - Updated reputation

Methodology Example: Beta Model

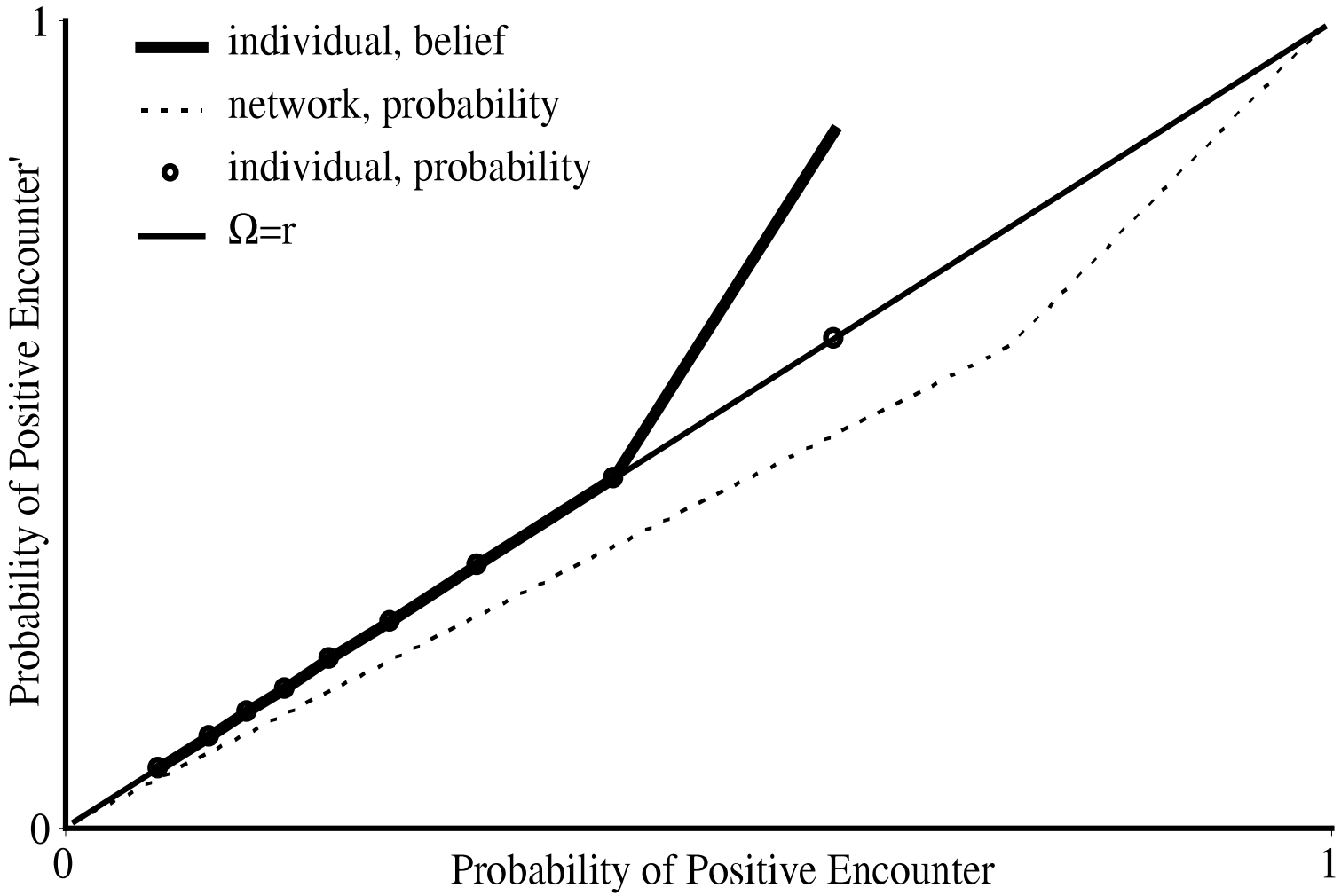
Quantize interactions into positive and negative experiences

- Determine the update function
  - Total number of positive and total number of negative experiences
  - Update according to favor being positive, else negative
- Determine the payoff function
  - Linear with reputation
- Consider each possible reputation of 10 total observations with a range of parameter settings
- Evaluate the metrics

Beta Model Variants

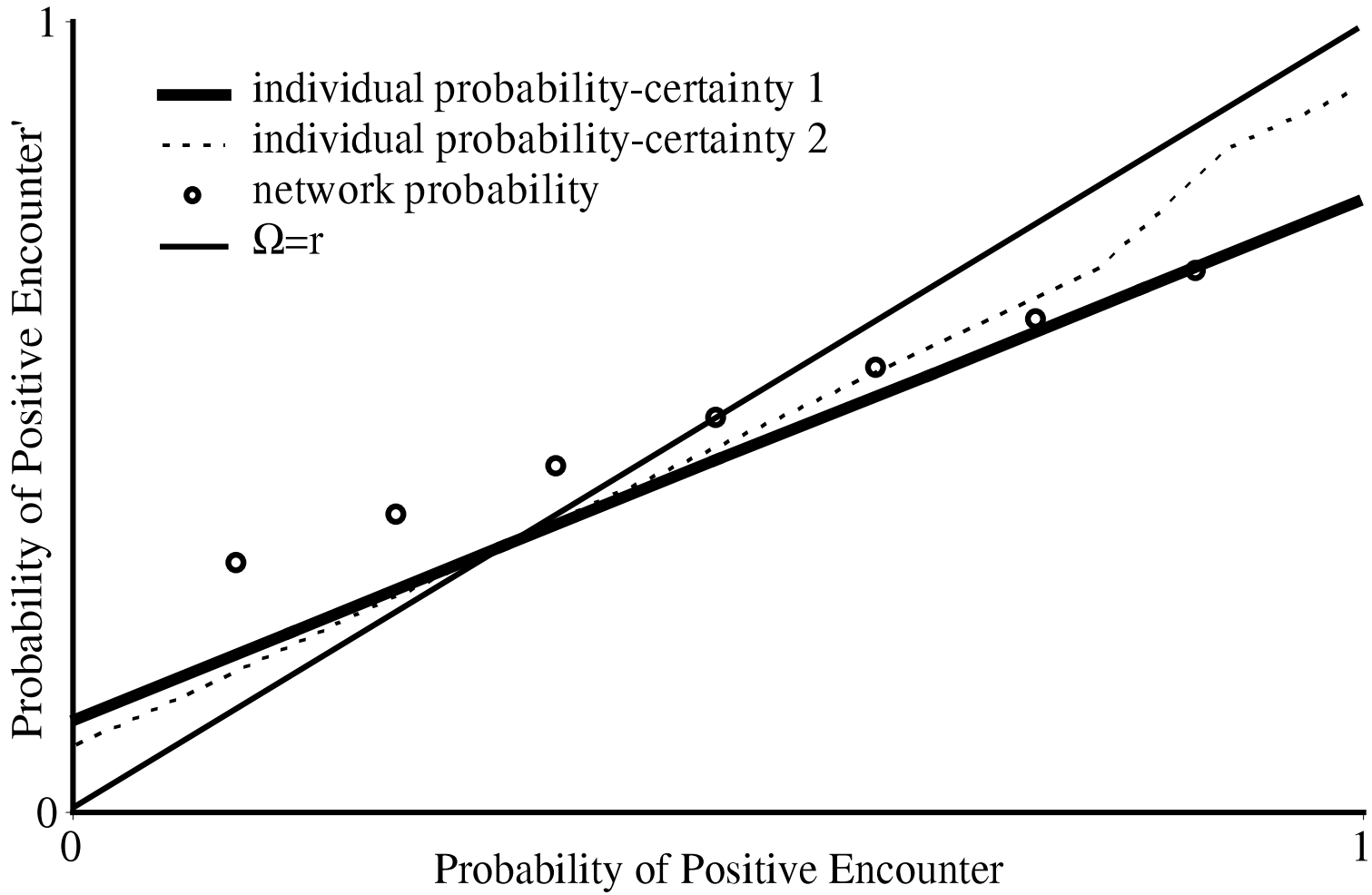


The Certainty Model



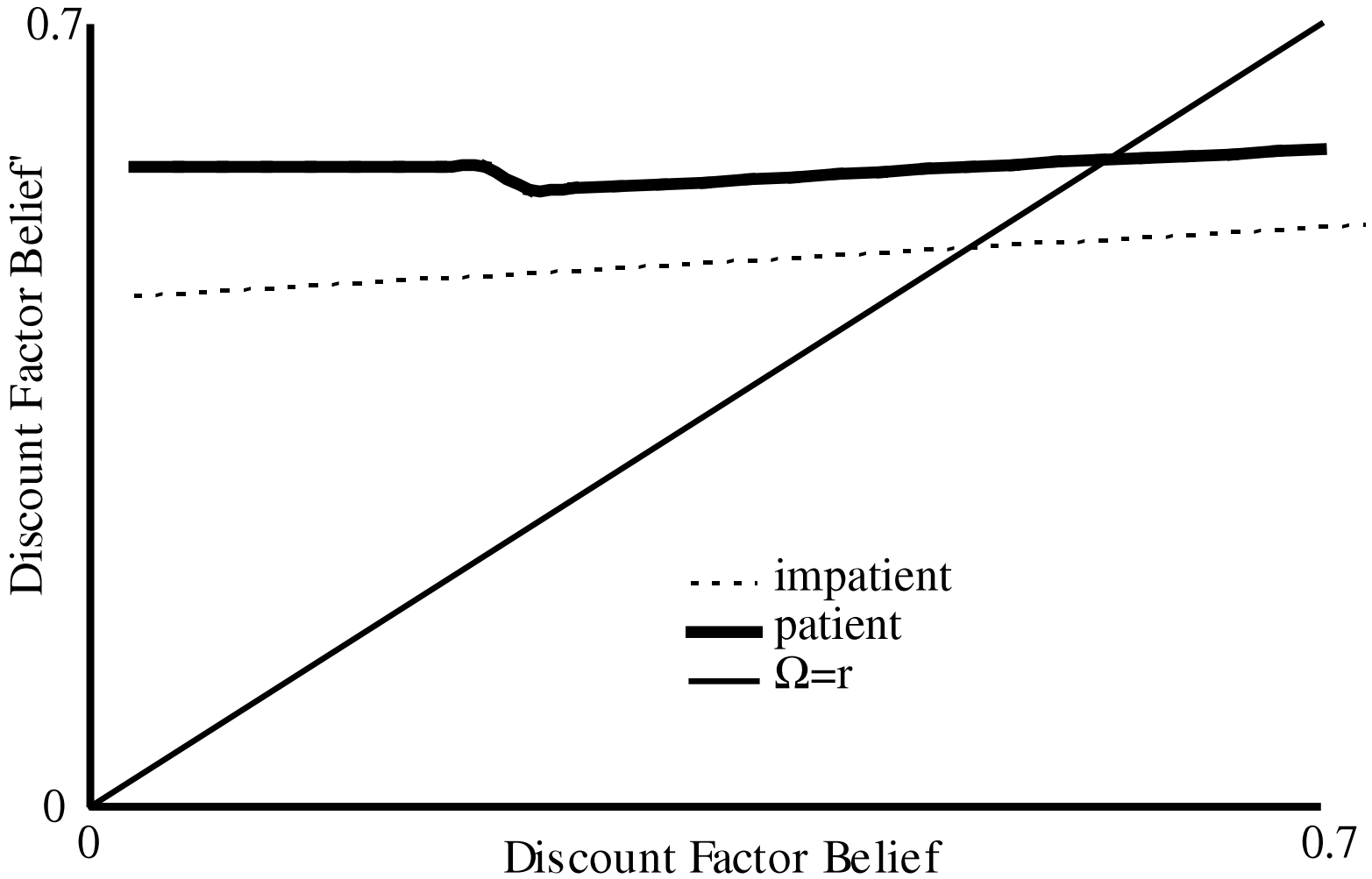
Belief $\sim$ certainty $\times$ expected value

The Travos Model



Discount Factor Graph

Trustworthiness $\sim$ patience [Hazard & Singh 2010; Smith & desJardins 2009]

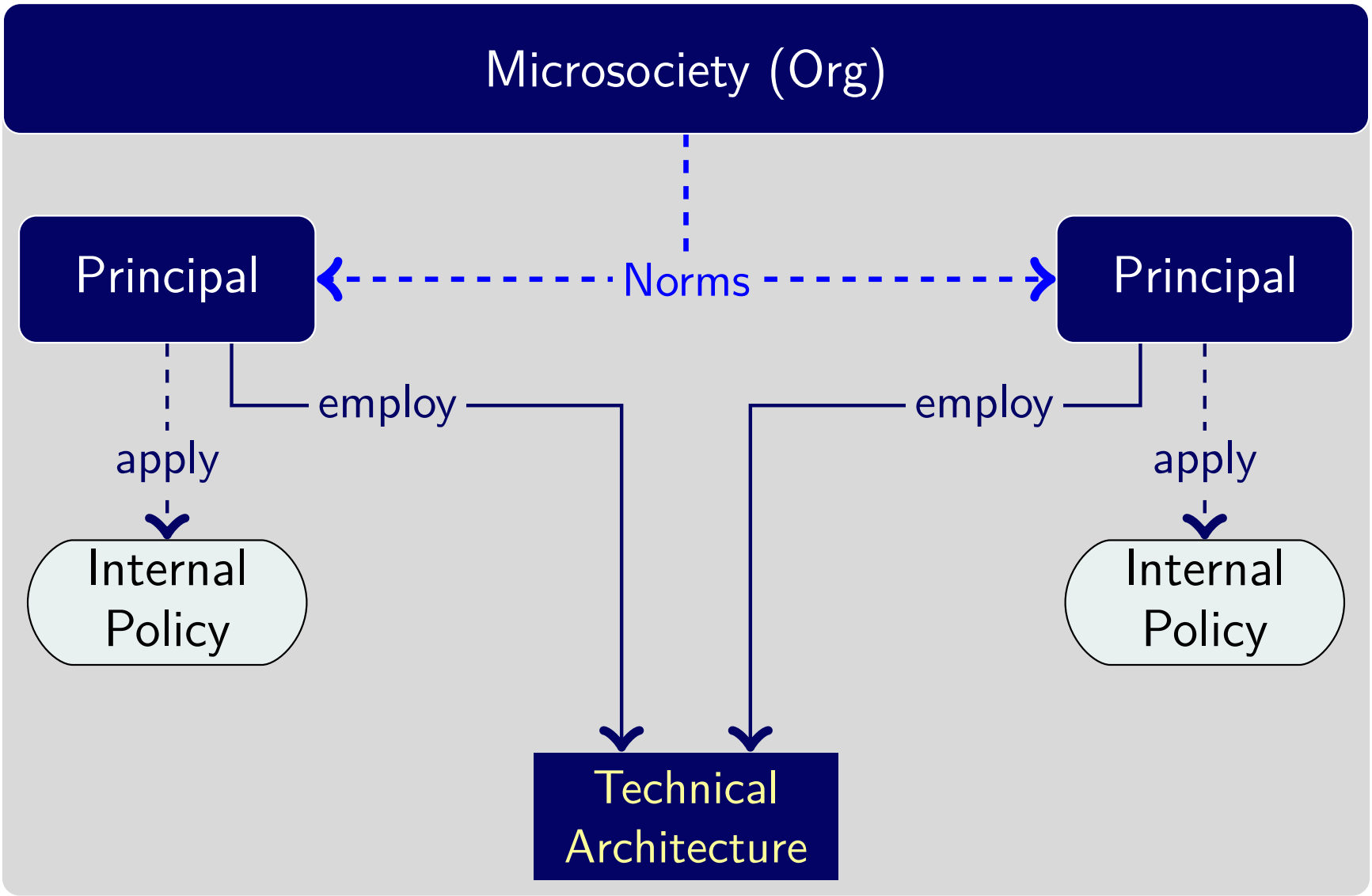


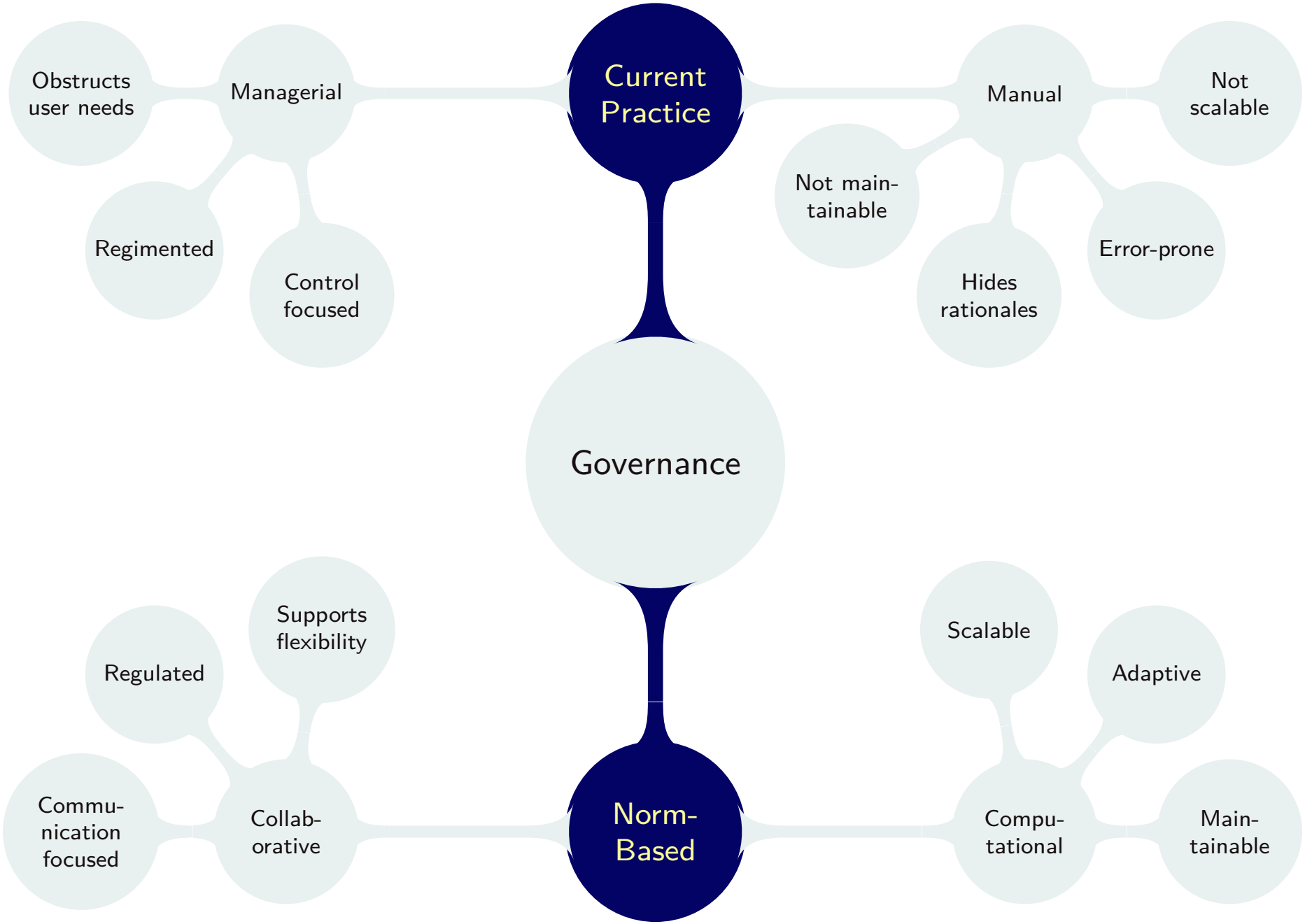
Results

Reputation System	Unambiguity	Monotonicity	Convergence (defined so lower is better)	Accuracy
Beta superlinear	yes	monotonic	0 and 0.9	0.4
Beta (sub)linear	yes	nondiscriminatory	0.9	0.45
Certainty	no	–	1	–
Discount Factor	yes	monotonic	< 0.1	0.02
Prob. Reciprocity	no	monotonic	no	0.2
Travos	yes	monotonic	0.8	0.2

A System is a Microsociety

Traditional view: A system is an artifact

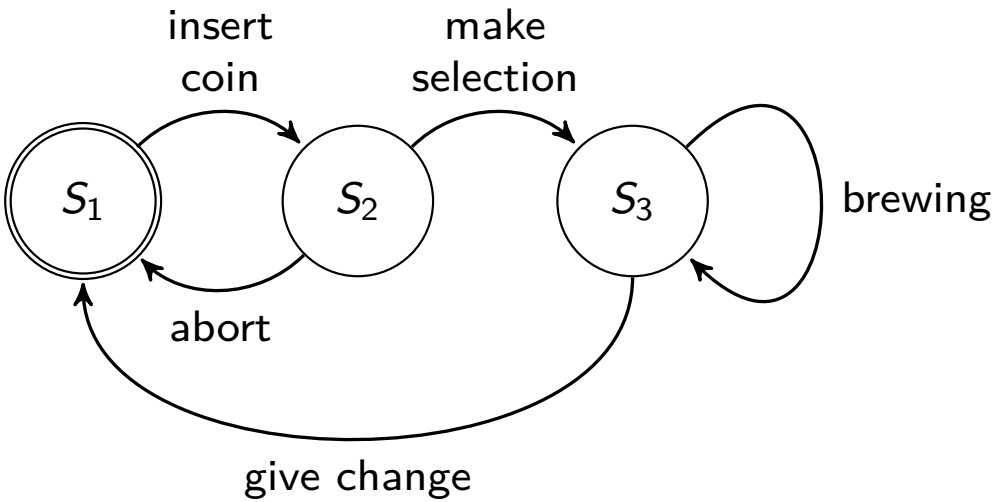


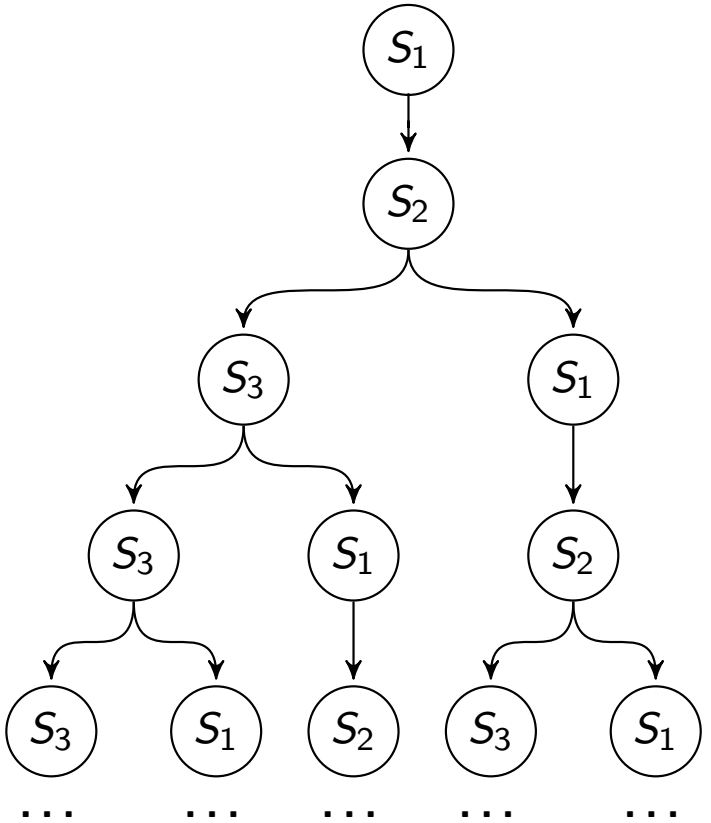

Current Practice
Obstructs user needs
Managerial
Regimented
Control focused
Manual
Not scalable
Not maintainable
Error-prone
Hides rationales
Governance
Norm-Based
Regulated
Supports flexibility
Communication focused
Collaborative
Scalable
Adaptive
Computational
Maintainable


Accountability
Traditional
Traceability
Fails when traceabiity fails
Insufficient
Unnec-essary
Deterrence
Penalty voids account-ability?
Treats account-ability as sanctioning
Norm-Based
Intuitive
Indepen-dent of mechanism
Captures domain re-quirements
Opera-tionalizable
Sanctioning
Logging
Monitoring

Vending Machine in Vienna

Conventional formal methods assume regimentation, i.e., a technical service





AF[Brew]: On every path, coffee is eventually brewed

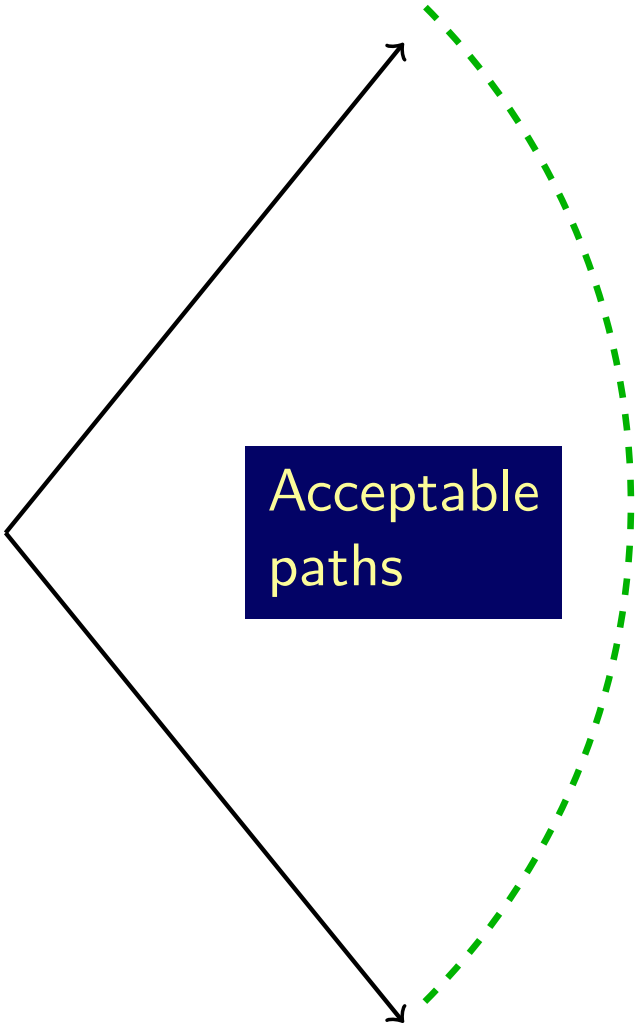
A[¬Brew U Coin]: On every path, no coffee is brewed prior to payment

http://embsys.technikum-wien.at/projects/decs/verification/formalmethods.php

Regimentation: Violations are Impossible

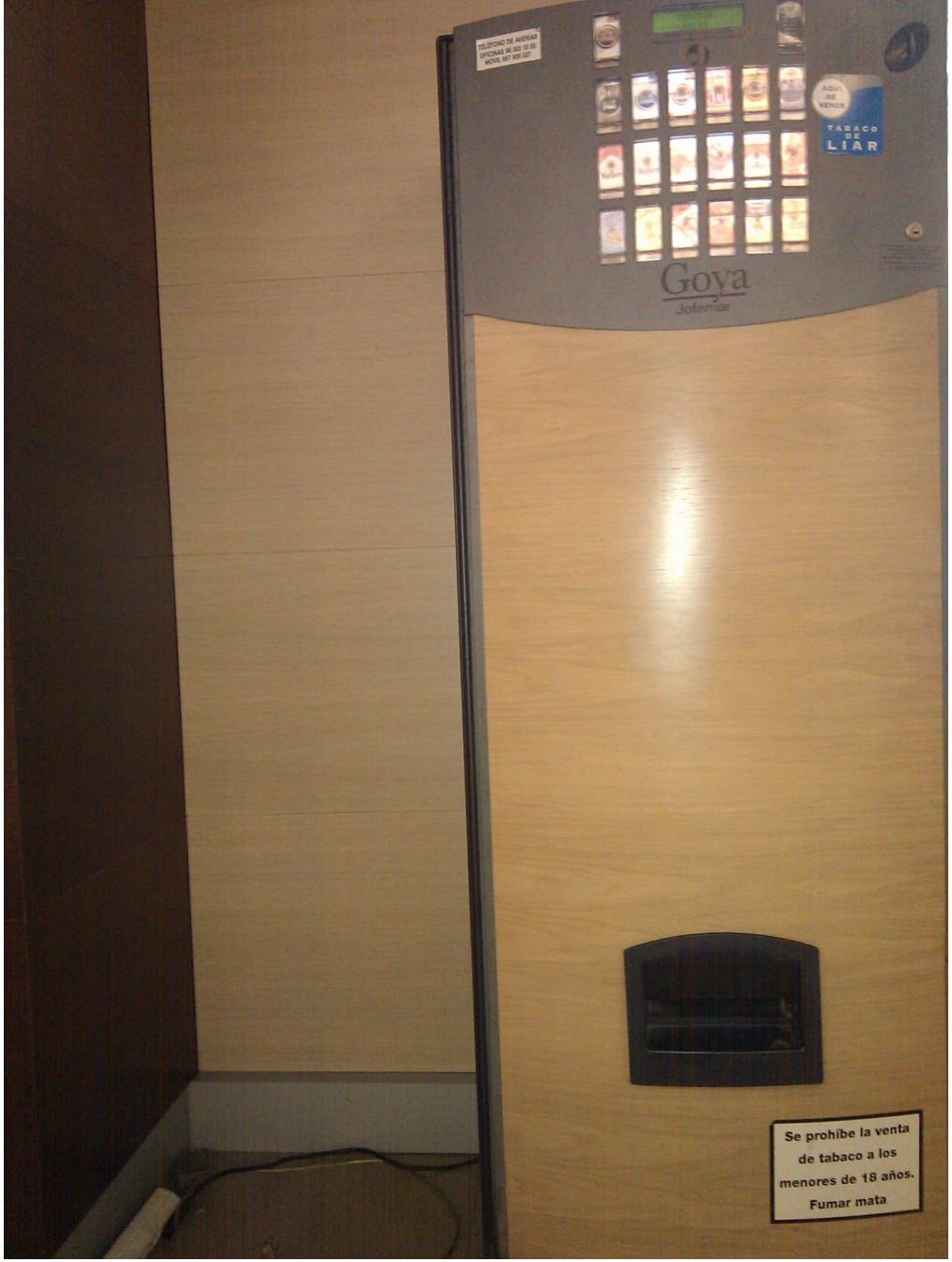
Viable assumption in a closed system

All paths the machine can generate in its environment



Vending Machine in Valencia

A business service



- Tall structure
- Hard to reach for short people
- Is that a bug or a feature?

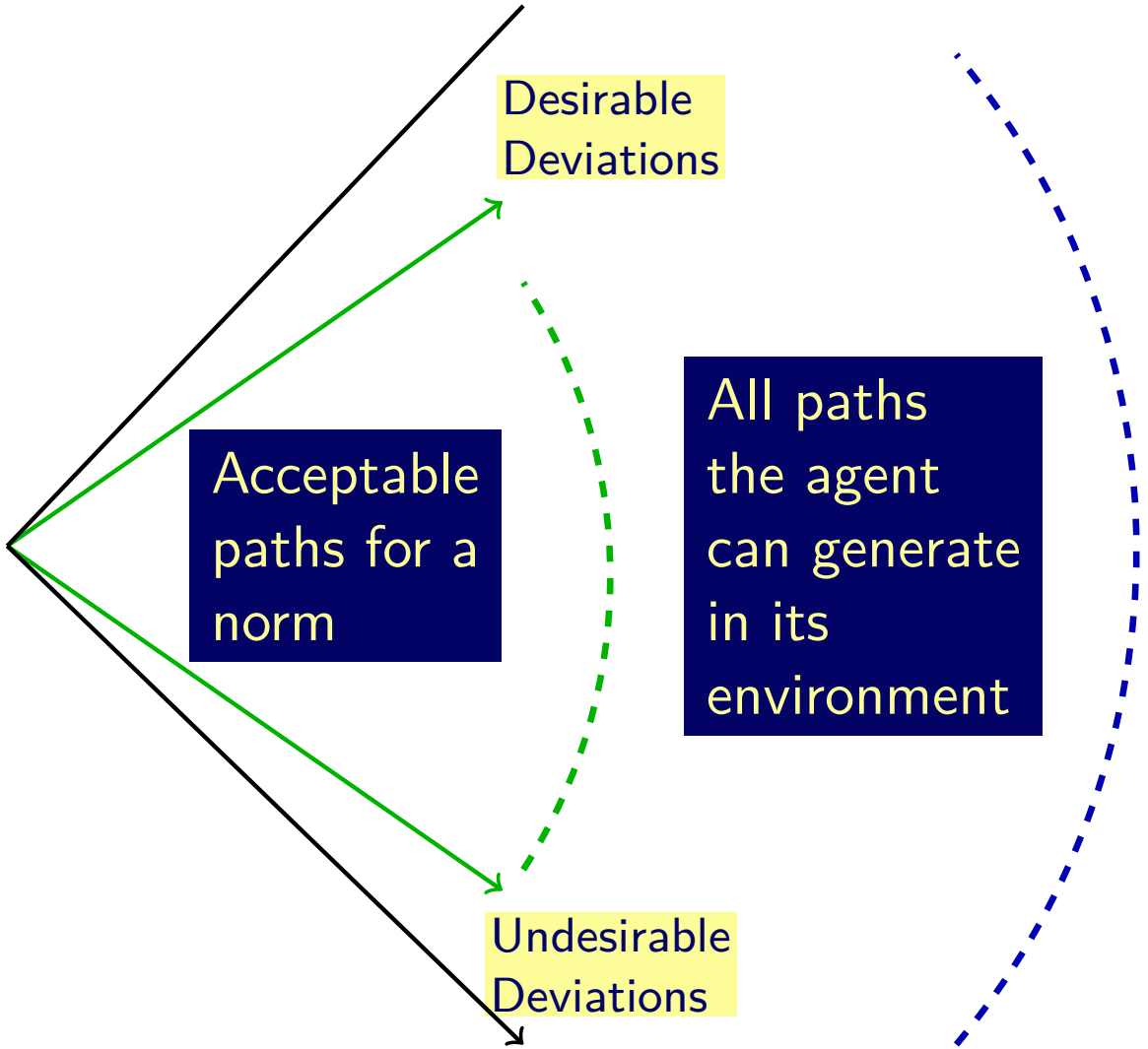
Vending Machine Close Up: Cigarettes!

Regulation

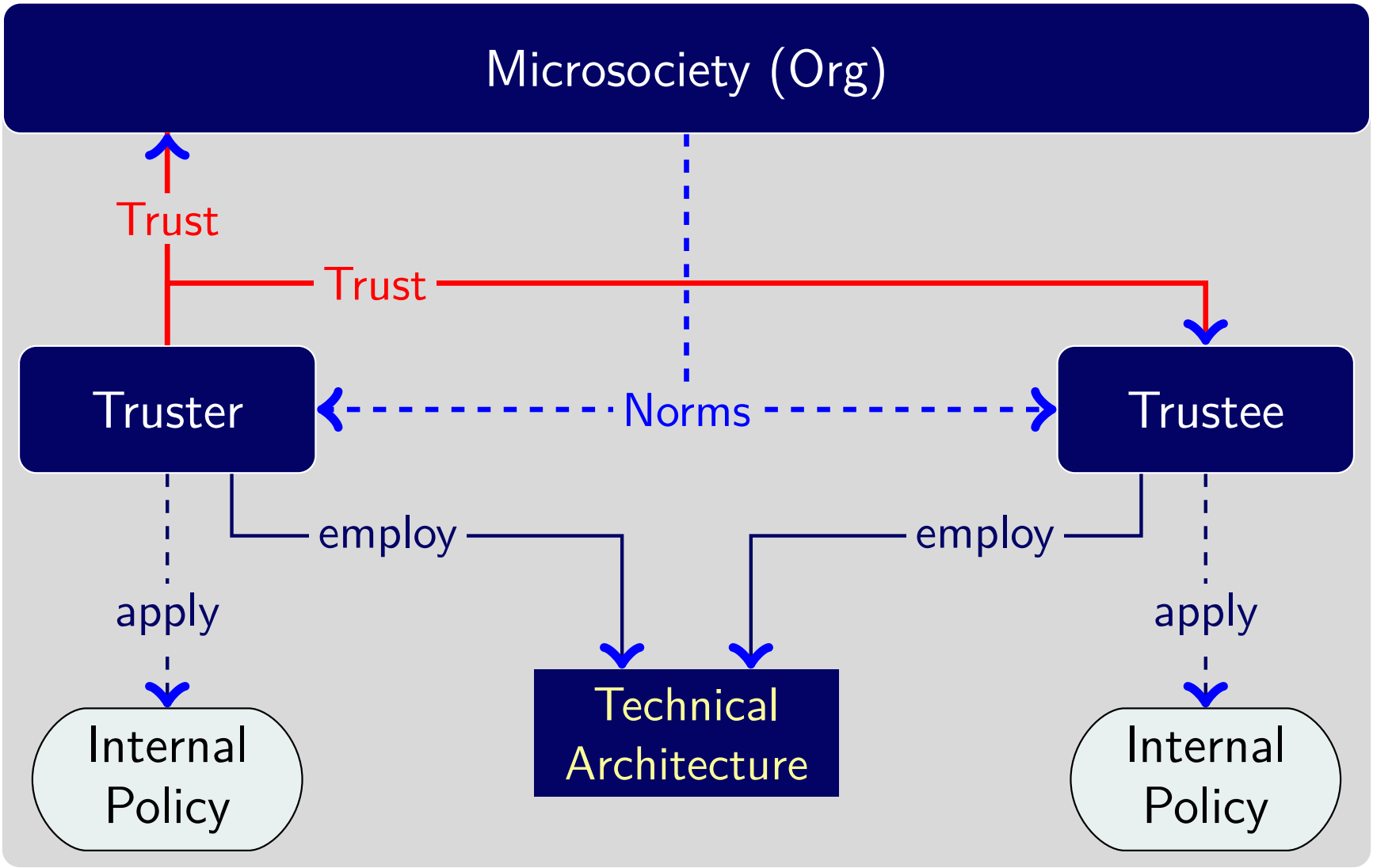

Se prohíbe la venta
de tabaco a los
menores de 18 años.
Fumar mata

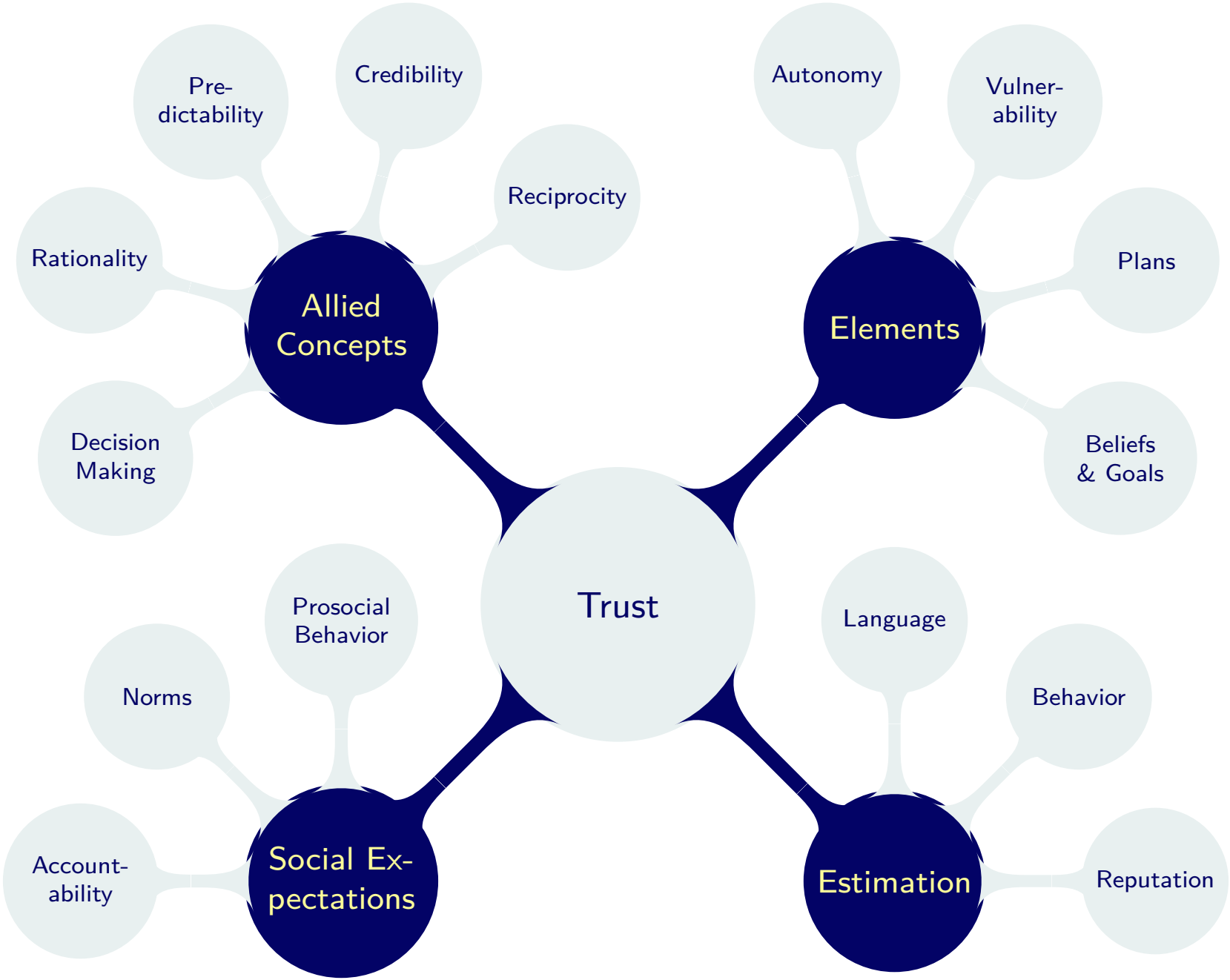
Regulation: Violations are Possible

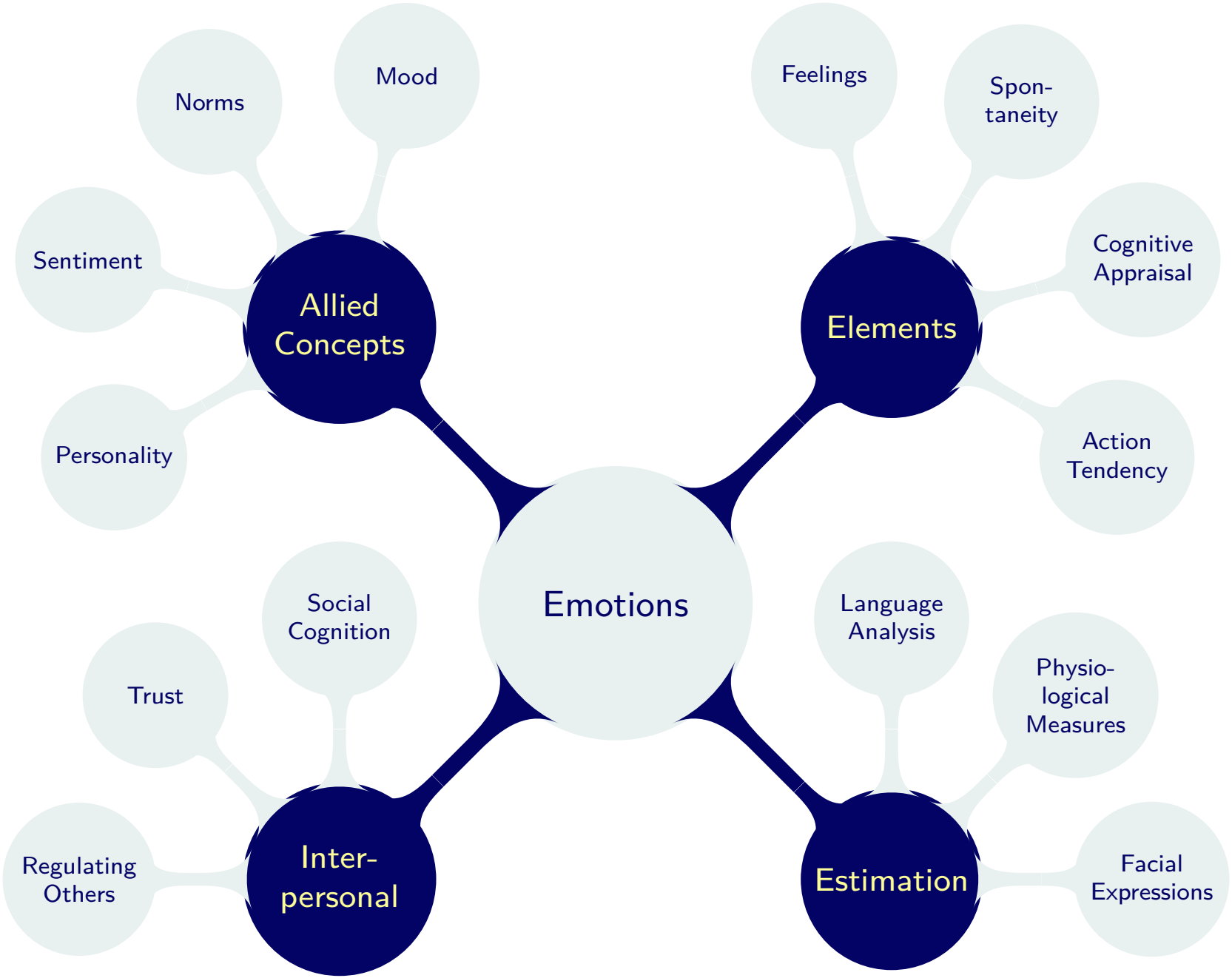
Appropriate assumption when dealing with autonomous parties



Trust via Norms: Trust as Presumed Accountability



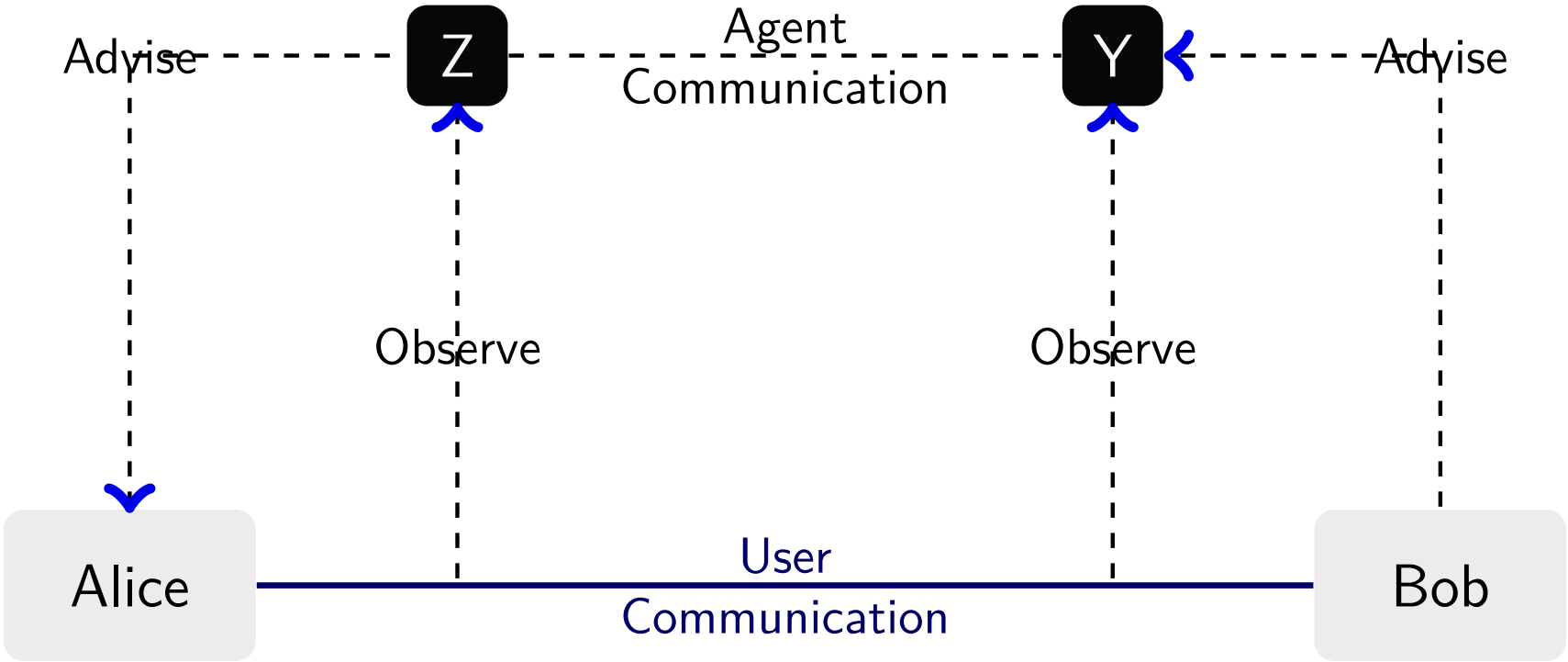

Trust
Allied Concepts
Pre-dictability
Credibility
Reciprocity
Rationality
Decision Making
Elements
Autonomy
Vulner-ability
Plans
Beliefs & Goals
Social Ex-pectations
Prosocial Behavior
Norms
Account-ability
Estimation
Language
Behavior
Reputation


Emotions
Allied Concepts
Norms
Mood
Sentiment
Personality
Elements
Feelings
Spon-taneity
Cognitive Appraisal
Action Tendency
Inter-personal
Social Cognition
Trust
Regulating Others
Estimation
Language Analysis
Physio-logical Measures
Facial Expressions

Possible Benefit of Understanding Trust and Emotions

Decision quality often depends upon the user's emotions

Could an agent support a user by conveying or inferring trust and emotions?



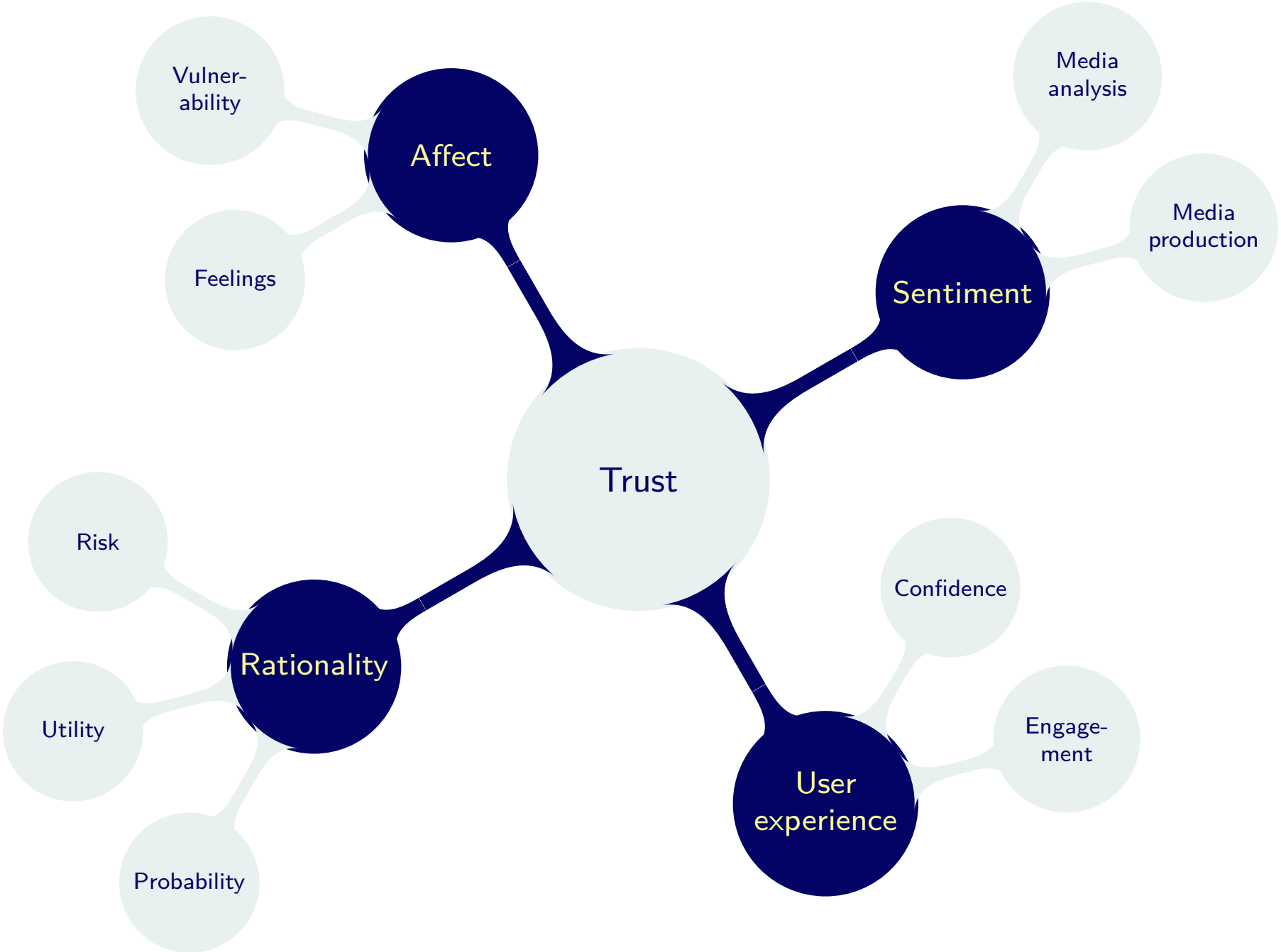
Paradox of Propensity

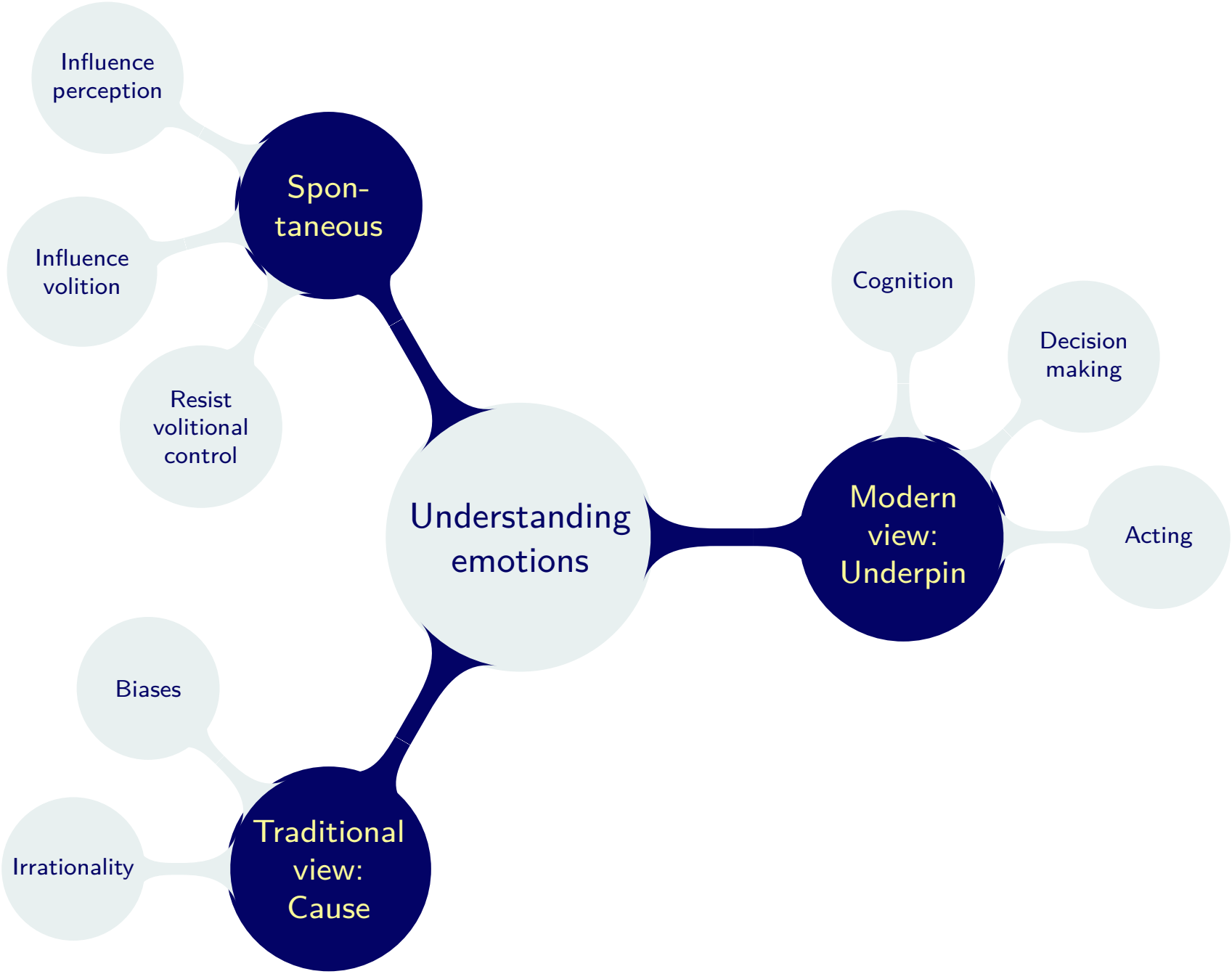
Humans have varying propensities to trust—independent of, or prior to, any experience with the trustee

- Propensity to trust
  - Makes intuitive sense, subjectively
  - Descriptive of people
- Trust applies in decision-making
  - Two or more choices, potentially involving differing vulnerabilities
  - Wouldn't a skeptical agent trust each choice less?
  - Wouldn't a credulous agent trust each choice more?
  - Thus, all else being equal, the decision is unaffected whether the agent is skeptical or credulous
- How can we reconcile
  - Subjective experience of trust and psychological realism
  - Elementary level of rationality

Authentic Trust

- Trust as understood today is not authentic
  - Consider probabilistic or utilitarian aspects
  - Determine trustworthiness of others
  - Establish incentives that promote trustworthiness
    - No need for trust in the psychological sense
- Authentic trust
  - Captures human view of trust
    - Arising from social interactions
    - Influencing expectations
    - Influenced by expectations
  - Affective process
  - Proposal: meta-affective process


Trust
Affect
Vulner-
ability
Feelings
Sentiment
Media
analysis
Media
production
Rationality
Risk
Utility
Probability
User
experience
Confidence
Engage-
ment


Influence perception
Spon-taneous
Influence volition
Resist volitional control
Understanding emotions
Cognition
Decision making
Modern view: Underpin
Acting
Biases
Traditional view: Cause
Irrationality

Psychoevolutionary Theory of Emotions

Darwin, Plutchik, ...

- Emotions are a communication and survival mechanism based on evolutionary adaptation
  - Anger protects against exploitation
  - Anxiety leads to goal fulfillment
  - Guilt leads to discharging commitments
- (Attempt to) influence and regulate interpersonal relations
- Events interrelated with feedback loops
  - Stabilize social relationships
- Expressed emotions
  - Human sneer derived from lower primates' snarl

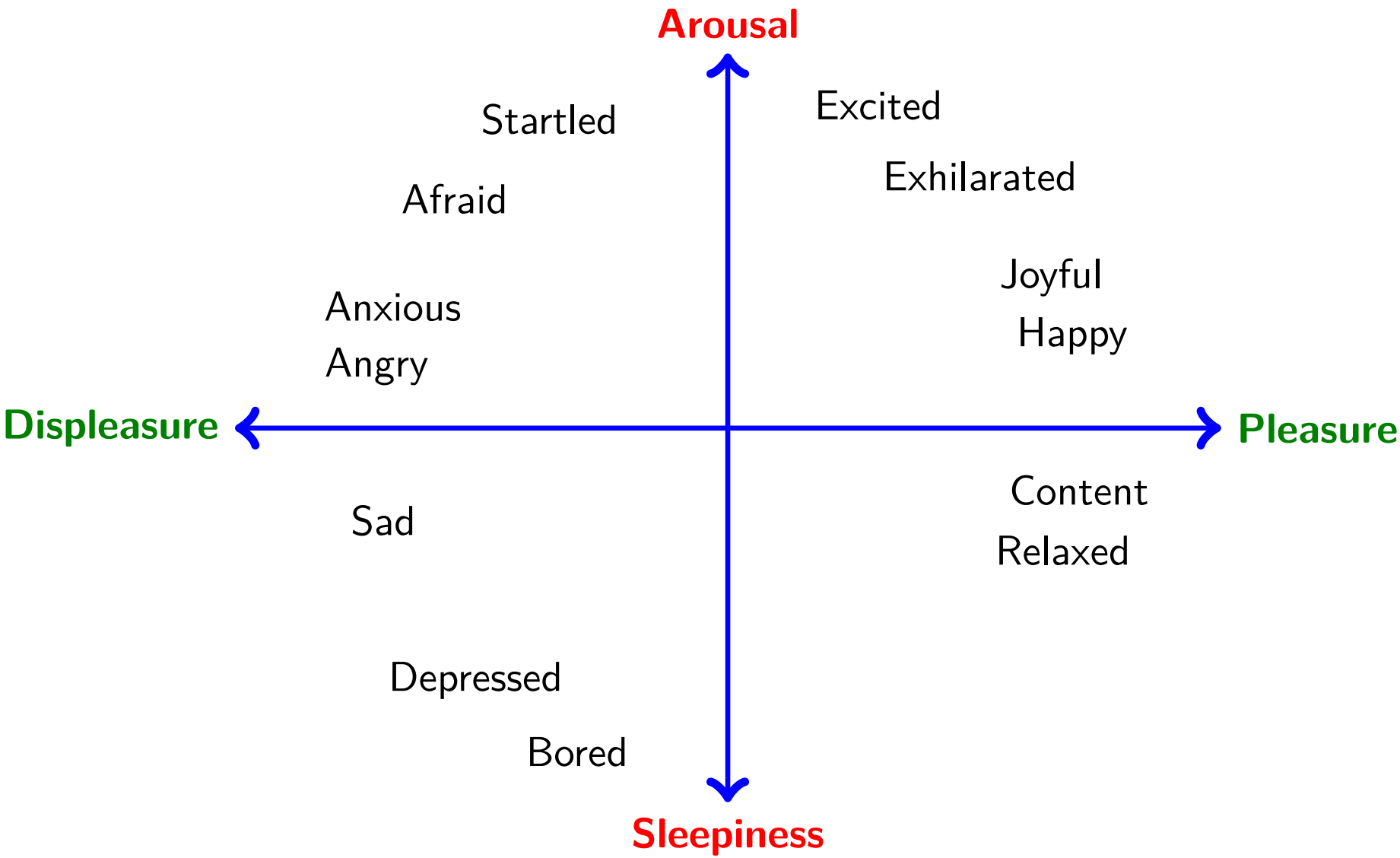
How to Describe Emotions

Russell

- Real emotions cut across lexical categories, such as angry and anxious
  - Nonbinary degree of membership in each category
  - Our language with specific lexemes may be limiting
- We never experience the same emotion (instance) twice
- Each category is a *script*
  - Prototypical causal and temporal chain
  - Explains how emotions in that category come about

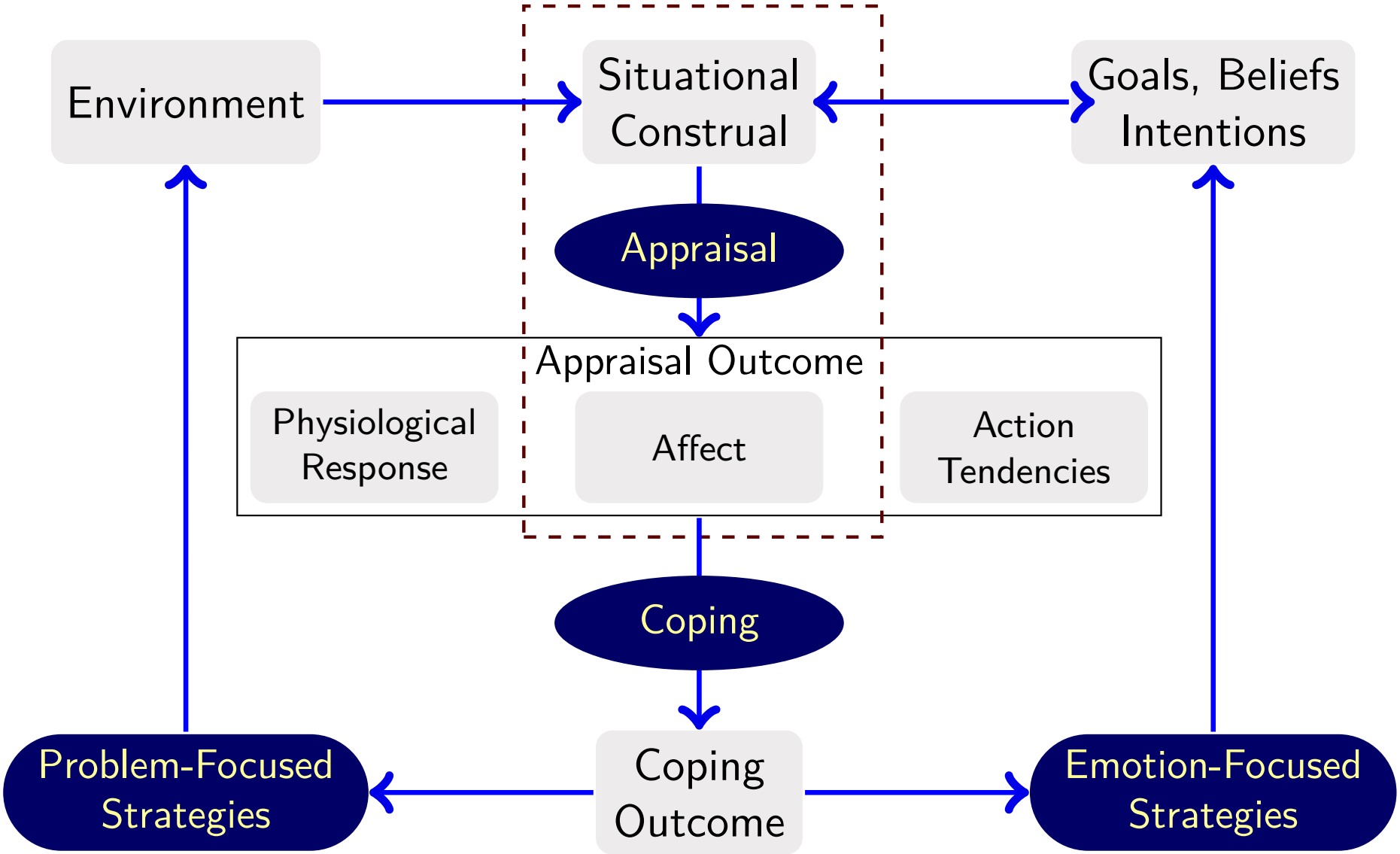
Circumplex Model of Emotions

Russell



Appraisal Theory

Lazarus and Smith



Self-Awareness and Reflection on Emotions

Lazarus and Kemper

- Emotion is a complex chain of *covert* responses
- Directly knowing one's feelings
  - I am angry
  - I am depressed
- Inferring one's emotional state by observing oneself
  - I have sweaty palms, so I must be anxious
- Additional components
  - Secondary cognitive appraisal
    - Evaluating one's coping resources
  - Reappraisals
    - Feedback from the environment to the individual's reactions

Emotions versus Personality

States versus Traits

- Emotions are *states*, i.e., in the moment
- Personality is a a *trait*, i.e., stable notion
  - Repeated measurements produce correlated results
  - Indicated by similar folk language, e.g.,
    - Gloomy
    - Cheerful
- Sometimes a matter of timescale and contextual information, e.g.,
  - Fear (specific stimulus) versus
  - Nervousness (personality) versus
  - Shyness (interpersonal relationship)
- Feelings
- Mood
  - An aggregation of feelings over time
  - Sometimes based on self-directed appraisals
- Sentiment: view of an agent
- Affect: any "unconscious" process, including the above

Some Relevant Brain Systems

- Emotions appear in multiple brain systems
  - Amygdala, an almond-shaped structure in the brain
    - Once considered the unique place for emotions
    - No longer so
  - Orbitofrontal cortex
    - Sometimes with its middle and lateral parts considered separately
  - Anterior insula
    - Latest addition
- Hippocampus—"sea horse" because of its shape
  - Important for related functions such as memory
  - Not emotions specifically

Neurological Observations

Studies often based on victims of accidents or illness

- Amygdala
  - Fear: lesions lead to psychic blindness
  - Face recognition, e.g., as evinced in abnormal eye movements
  - Feeling sadness
    - Controversy as to whether its left or right deals with sadness
- Orbitofrontal cortex
  - Anger
- Anterior insula
  - Feel disgust
  - Recognize disgust

Mirror Neurons

- Activation measured in terms of "lighting up" in fMRI (functional MRI) in the appropriate conditions
- The same parts of the brain are activated whether one
  - Perceives another agent performing an action
  - Performs the same action oneself
- Studies of monkeys seeing a researcher reaching for a peanut
  - Only when they know the researcher is reaching a peanut
  - The peanut can be invisible, though
- Explanation for mimicry in infants
  - Too young to have been trained

Social Cognition

Bridges emotions and trust

- Understanding one's social relationships
- Understanding others' social relationships
- Relies on the orbitofrontal cortex, lesions on which lead to awkward social interactions
  - Sitting too close to others
  - Talking too much to strangers

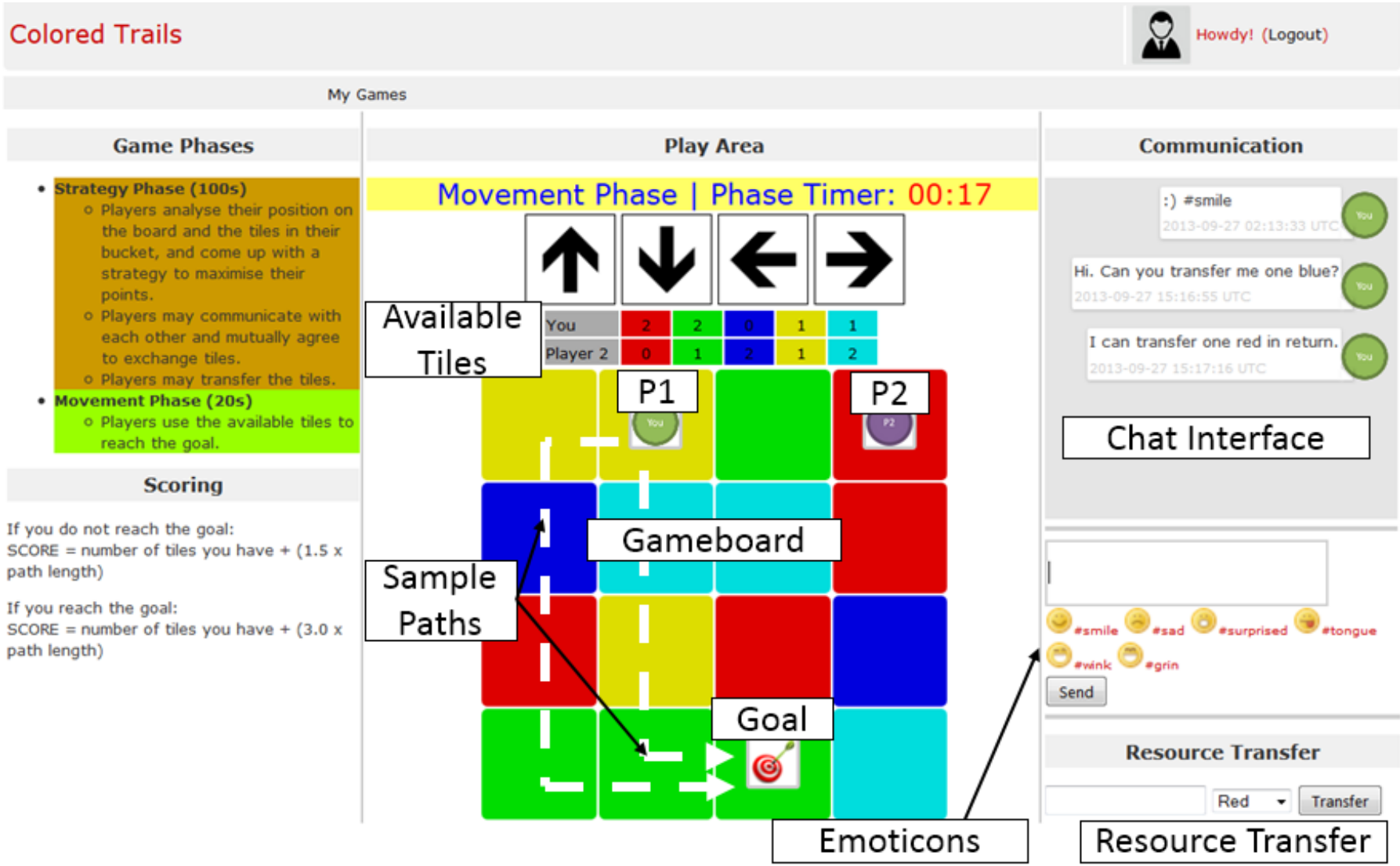
Empirical Studies Relating Emotions and Trust

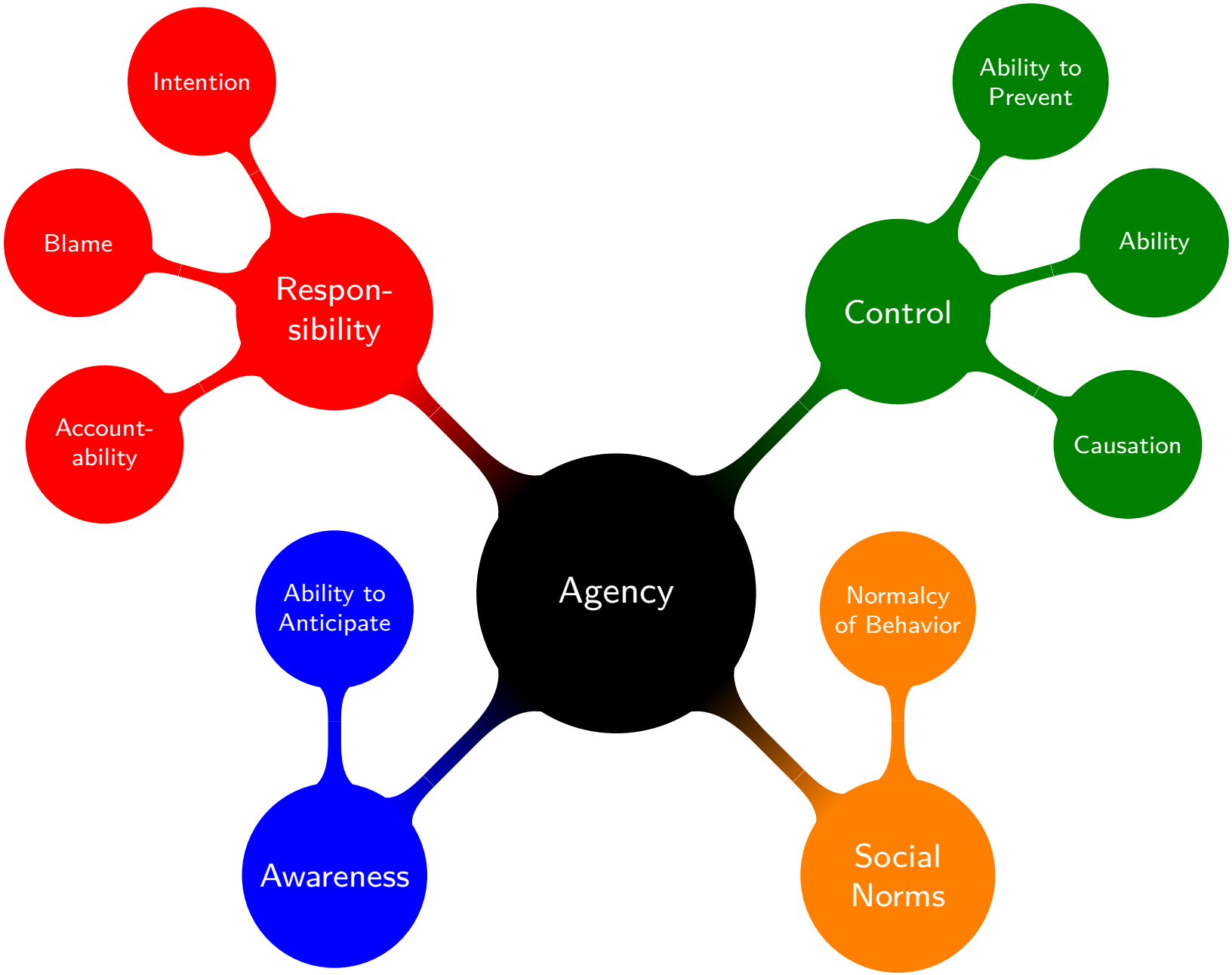
Fewer than one would expect

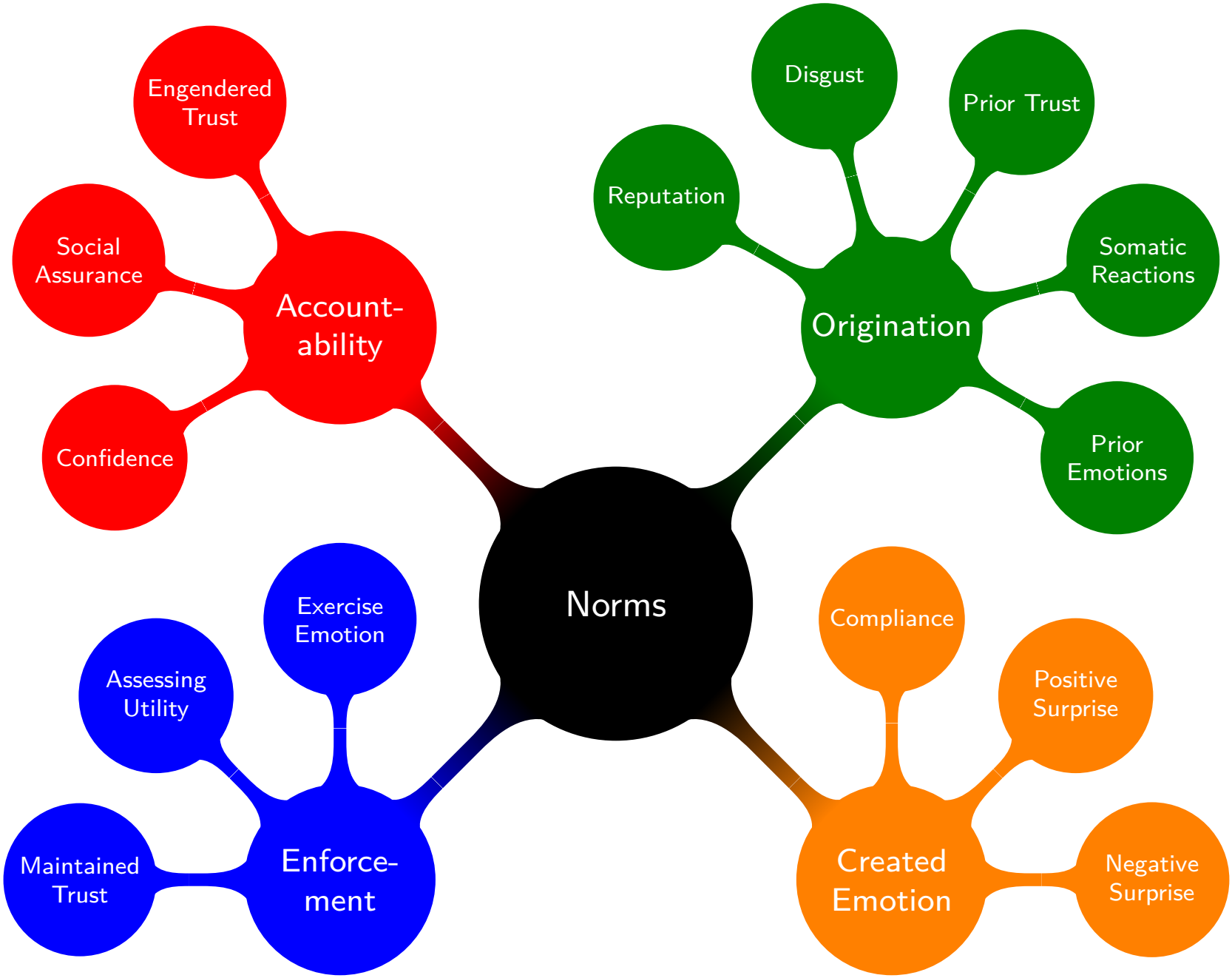
- Positive emotions increase trust whereas negative emotions decrease trust [Dunn and Schweitzer, 2005]
- Gratitude (an interpersonal emotion) triggers a positive mood [Sheldon and Lyubomirsky, 2007]
- de Melo et al. [2012]: People base expected cooperation based on emotions displayed facially or verbally; social appraisal matters
- Kalia et al. [2014]: Emotions, mood, goals, commitments relate to trust
  - Current trust depends upon trust at previous moment
  - Current mood depends upon success with previous goals
  - Current expectations depend upon own behavior (guilt or righteousness)
  - Own behavior depends upon trust in others

Empirical Study Design

- 30 subjects, all computer science students
- Data collected
  - 450 rows of data (30 subjects $\times$ 3 games $\times$ 5 rounds per game)
  - Chat messages between subjects
  - Surveys
- Threats to validity
  - Subjects unlike typical end users
  - Unrealistic situation
  - Interruptions and surveys can influence emotions

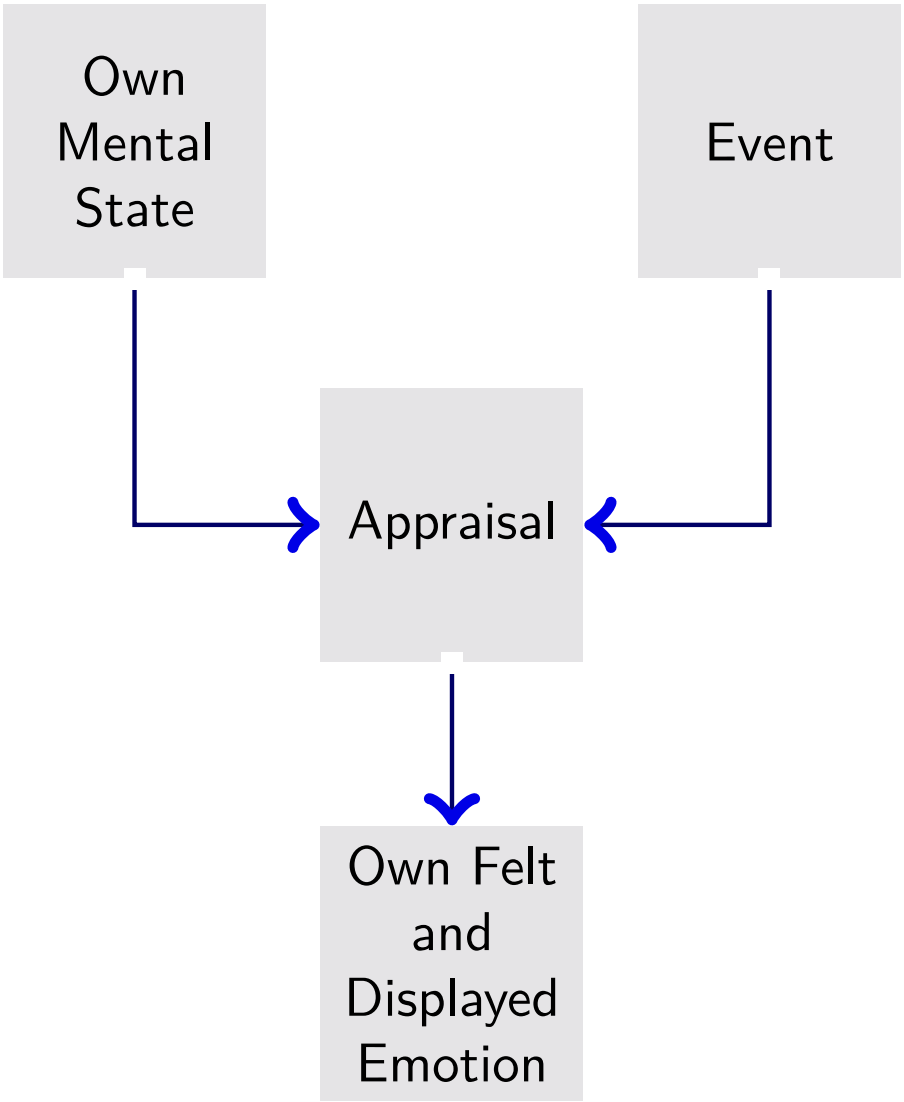

Colored Trails
Howdy! (Logout)
My Games
Game Phases
Strategy Phase (100s)
Players analyse their position on the board and the tiles in their bucket, and come up with a strategy to maximise their points.
Players may communicate with each other and mutually agree to exchange tiles.
Players may transfer the tiles.
Movement Phase (20s)
Players use the available tiles to reach the goal.
Scoring
If you do not reach the goal:
SCORE = number of tiles you have + (1.5 x path length)
If you reach the goal:
SCORE = number of tiles you have + (3.0 x path length)
Play Area
Movement Phase | Phase Timer: 00:17
Available Tiles
You 2 2 0 1 1
Player 2 0 1 2 1 2
P1
P2
Gameboard
Sample Paths
Goal
Emoticons
Communication
:) #smile
2013-09-27 02:13:33 UTC
Hi. Can you transfer me one blue?
2013-09-27 15:16:55 UTC
I can transfer one red in return.
2013-09-27 15:17:16 UTC
Chat Interface
#smile #sad #surprised #tongue #wink #grin
Send
Resource Transfer
Red
Transfer
Resource Transfer

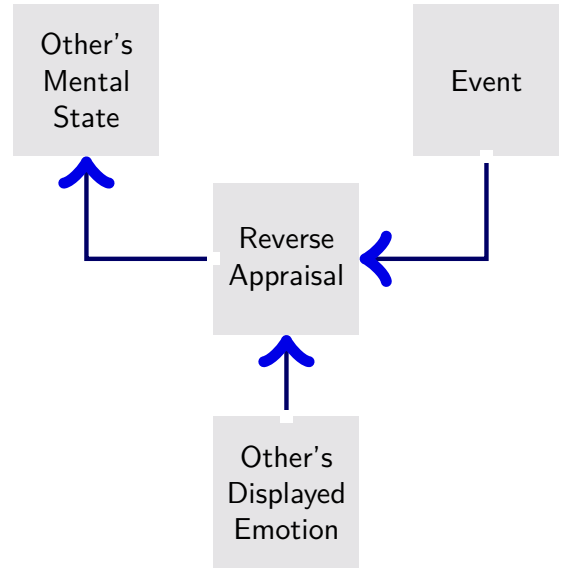

Intention
Blame
Respon-
sibility
Account-
ability
Ability to
Prevent
Ability
Control
Causation
Ability to
Anticipate
Agency
Normalcy
of Behavior
Awareness
Social
Norms

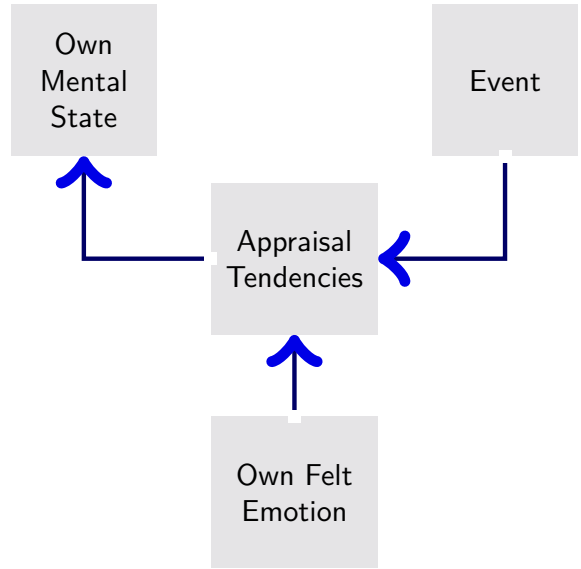

Norms
Account-ability
Engendered Trust
Social Assurance
Confidence
Origination
Disgust
Prior Trust
Reputation
Somatic Reactions
Prior Emotions
Enforce-ment
Exercise Emotion
Assessing Utility
Maintained Trust
Created Emotion
Compliance
Positive Surprise
Negative Surprise

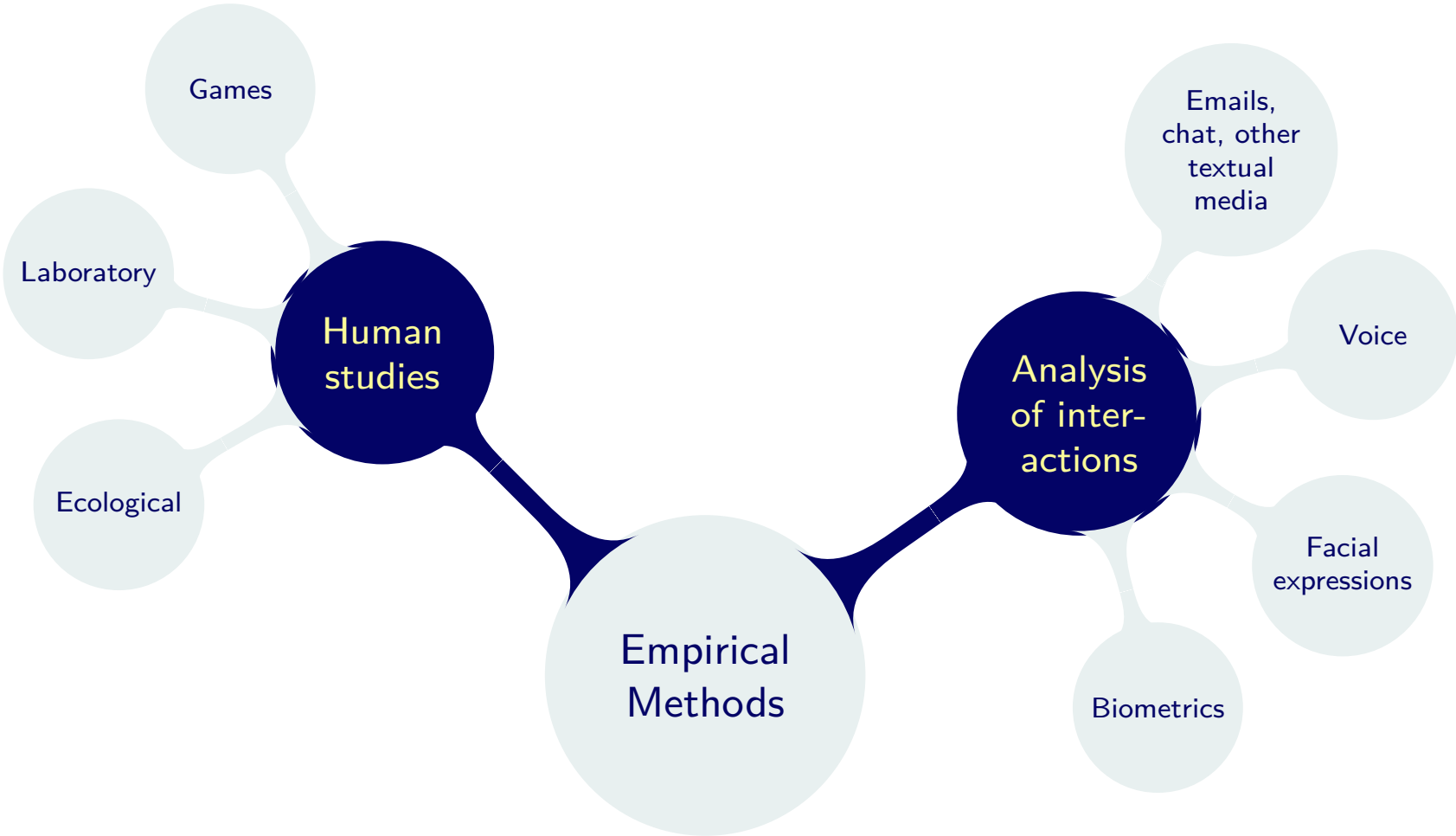
Social Appraisal Theory

Appraisals, reverse appraisals, and self-monitoring appraisal tendencies





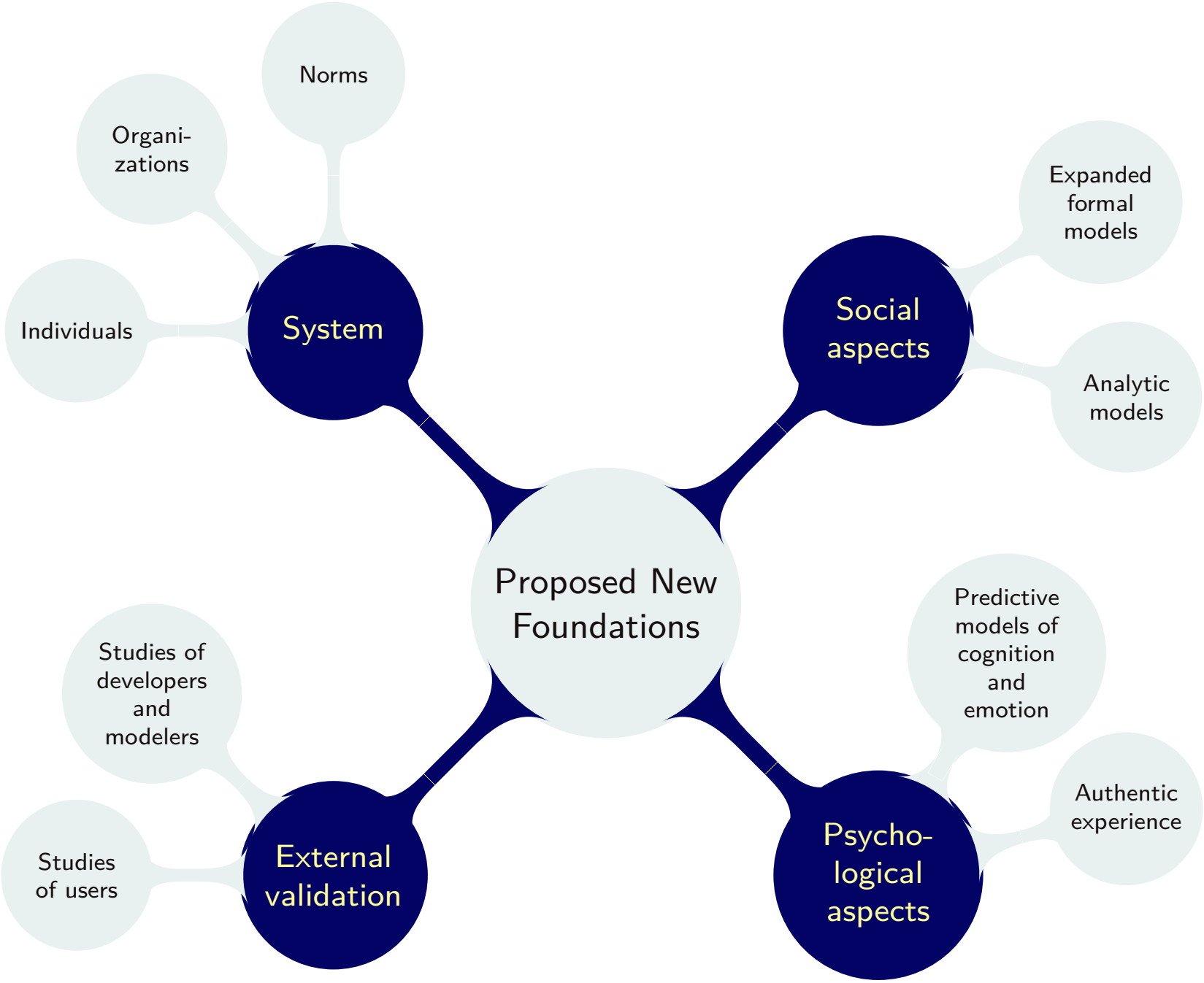



Games
Laboratory
Ecological
Human studies
Empirical Methods
Analysis of inter-actions
Emails, chat, other textual media
Voice
Facial expressions
Biometrics

Interesting Observations

Can we develop theories that explain the full range of trust phenomena?

- Many human norms associate with disgust (for violations)
  - Evolutionary basis for carrying out harmful behaviors
  - Trained through culture
- Trustworthy or prosocial behavior has interesting correlations with indicators of benevolence or malice
  - Implicit influence: Clean scents promote reciprocity and charity [Liljenquist et al., 2010]
  - Explicit influence: Black sports uniforms correlate with greater fouls [Frank and Gilovich, 1988]
- Emotional engagement required to punish norm violators
  - To help improve prosocial behavior
  - Even though the punisher usually has a negative utility


Norms
Organi-
zations
Individuals
System
Expanded
formal
models
Social
aspects
Analytic
models
Proposed New
Foundations
Predictive
models of
cognition
and
emotion
Studies of
developers
and
modelers
Authentic
experience
Studies
of users
External
validation
Psycho-
logical
aspects

Thanks and Plugs

- Acknowledgments
  - US Department of Defense
  - US National Science Foundation

- Consider submitting to
  - ACM Transactions on Internet Technology
  - IEEE Internet Computing

References I

Celso M. de Melo, Peter Carnevale, Stephen Read, Dimitrios Antos, and Jonathan Gratch. Bayesian model of the social effects of emotion in decision-making in multiagent systems. In *Proceedings of the 11th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 55–62, Valencia, Spain, June 2012.

Jennifer R. Dunn and Maurice E. Schweitzer. Feeling and believing: The influence of emotion on trust. *Journal of Personality and Social Psychology*, 88(5):736–748, May 2005.

Mark G. Frank and Thomas Gilovich. The dark side of self- and social perception: Black uniforms and aggression in professional sports. *Journal of Personality and Social Psychology*, 54(1):74–85, January 1988.

References II

Anup K. Kalia, Nirav Ajmeri, Kevin Chan, Jin-Hee Cho, Sibel Adalı, and Munindar P. Singh. A model of trust, moods, and emotions in multiagent systems, and its empirical evaluation. In *Proceedings of the 16th AAMAS Workshop on Trust in Agent Societies (Trust)*, Paris, May 2014.

Katie Liljenquist, Chen-Bo Zhong, and Adam D. Galinsky. The smell of virtue: Clean scents promote reciprocity and charity. *Psychological Science*, 21(3):381–383, March 2010.

Kennon M. Sheldon and Sonja Lyubomirsky. How to increase and sustain positive emotion: The effects of expressing gratitude and visualizing best possible selves. *The Journal of Positive Psychology*, 1(2):73–82, April 2007.