Autonomous Agents and Ethical Multiagent Systems

Challenges and Hints at a Research Program

Munindar P. Singh
singh@ncsu.edu
https://www.csc.ncsu.edu/faculty/mpsingh/

Department of Computer Science
North Carolina State University

Outline

Standing on the Shoulders of Giants

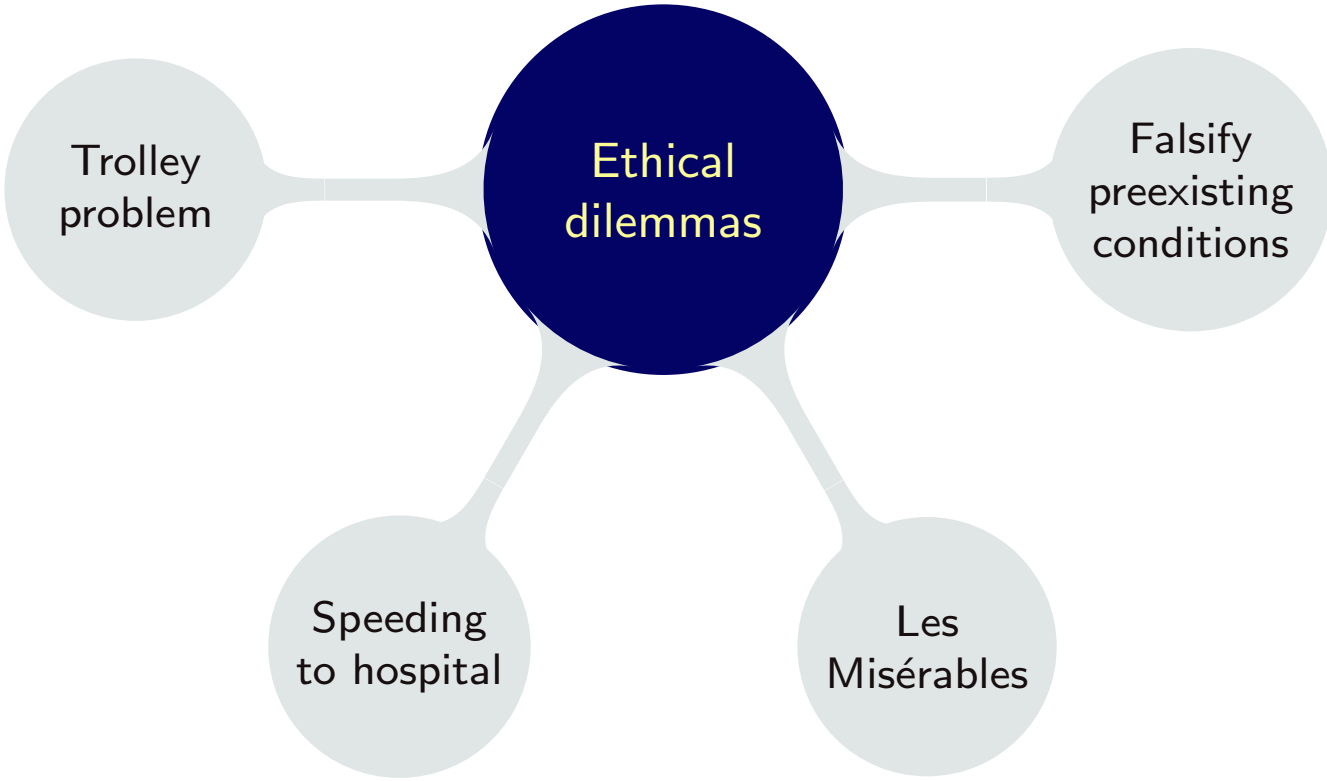
But looking in another direction?

US Library of Congress, manuscript c. 1410; from Wikipedia

Outline

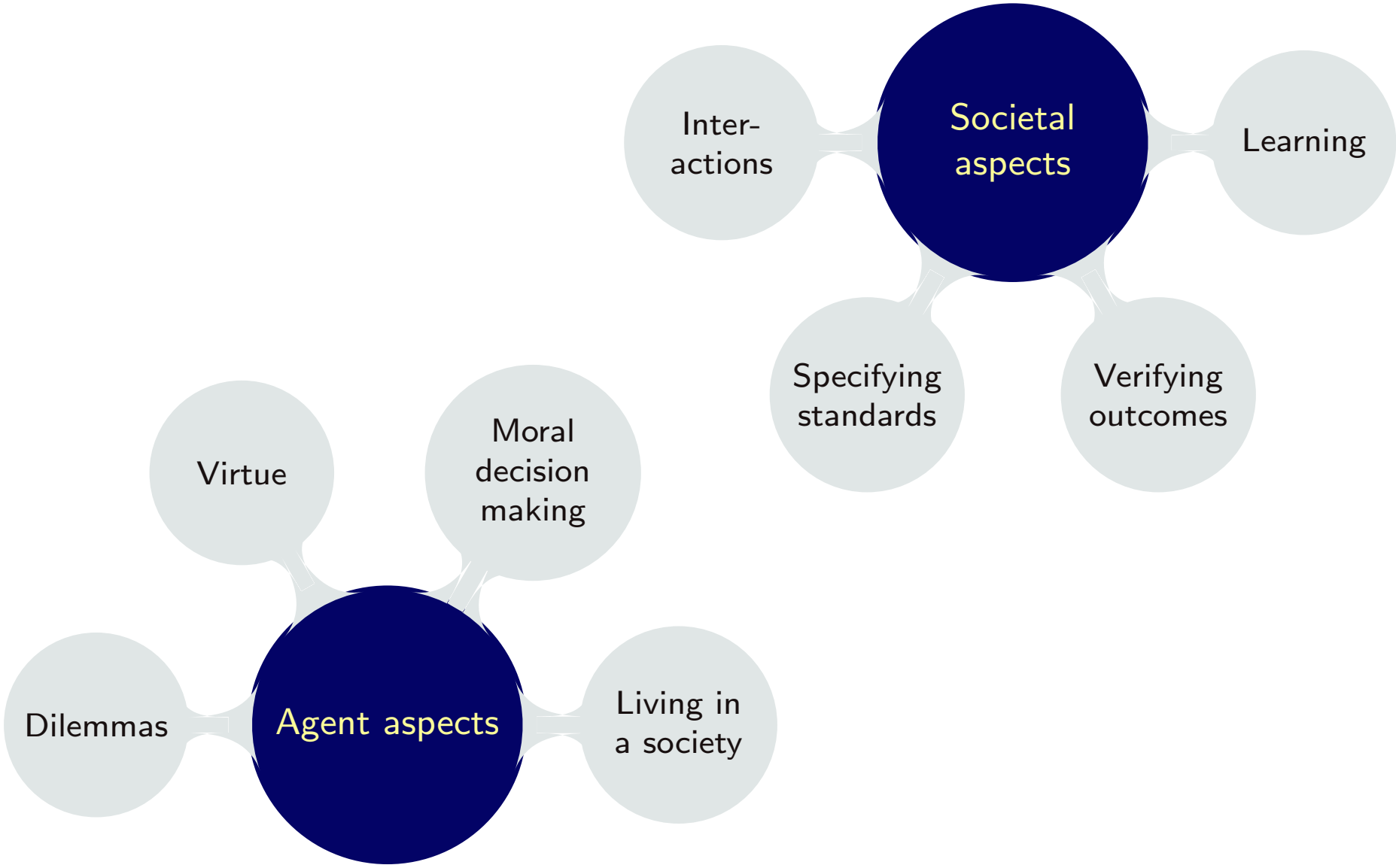
Ethical Dilemmas: No Good Choices

Contrast the following examples



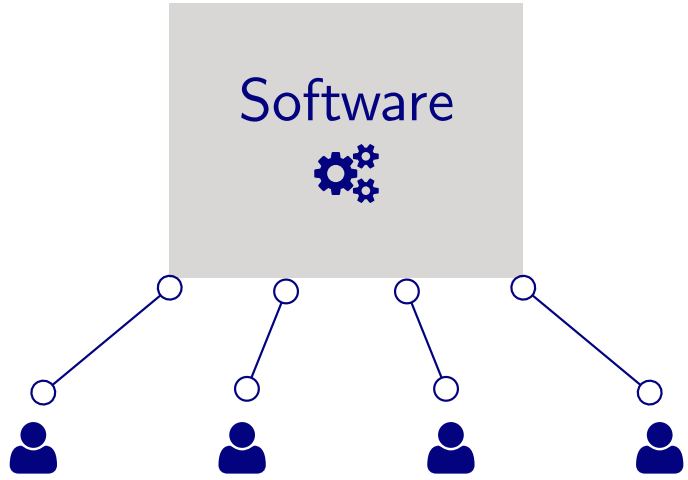
Ethics in Multiagent Systems

Ethics is an inherently multiagent concern, yet current approaches focus on single agents



“Ethics” of a Central Technical Entity

Today’s view of AI ethics involves how a single agent deals with people
Such as a prediction algorithm or an autonomous vehicle

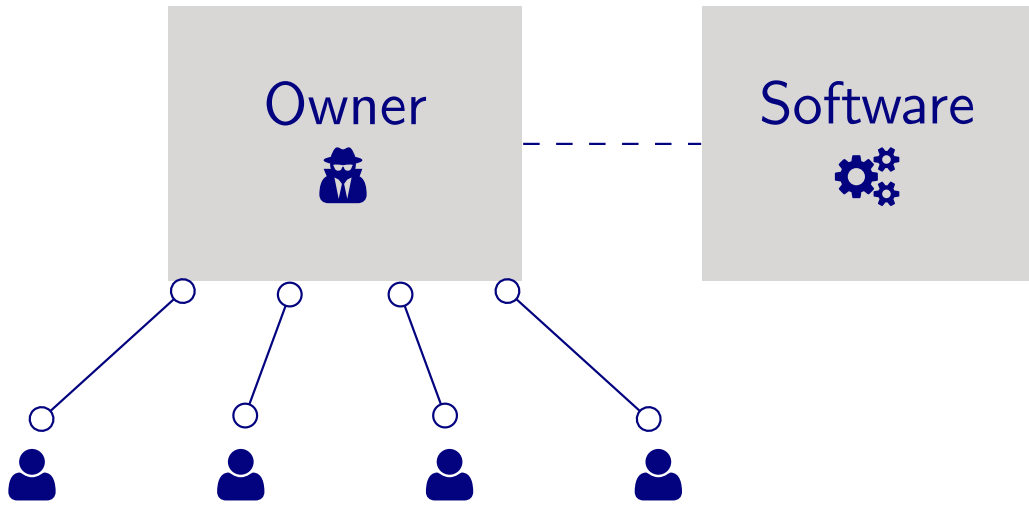


- Autonomy is defined as automation: complexity and intelligence
- Dilemmas à la trolley problems approached in an atomistic manner

Ethics of a Social Entity with a Software Decision Aid

A social entity, assisted by software, wields power over people

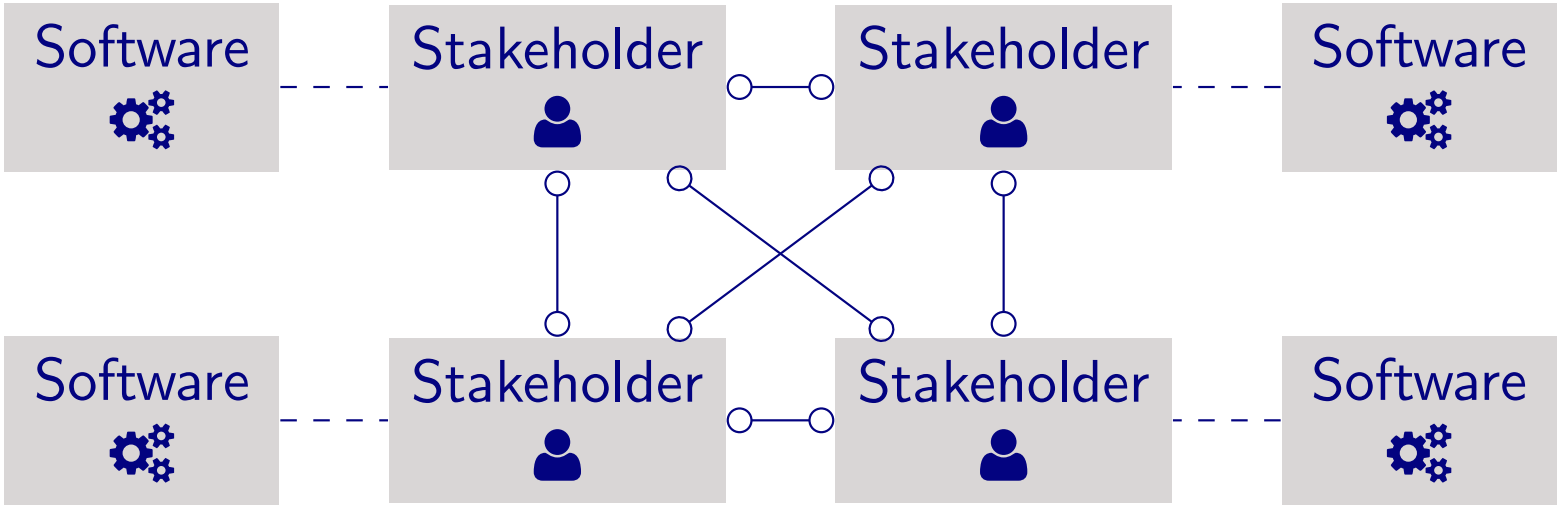
Ethical concerns focused on social entity

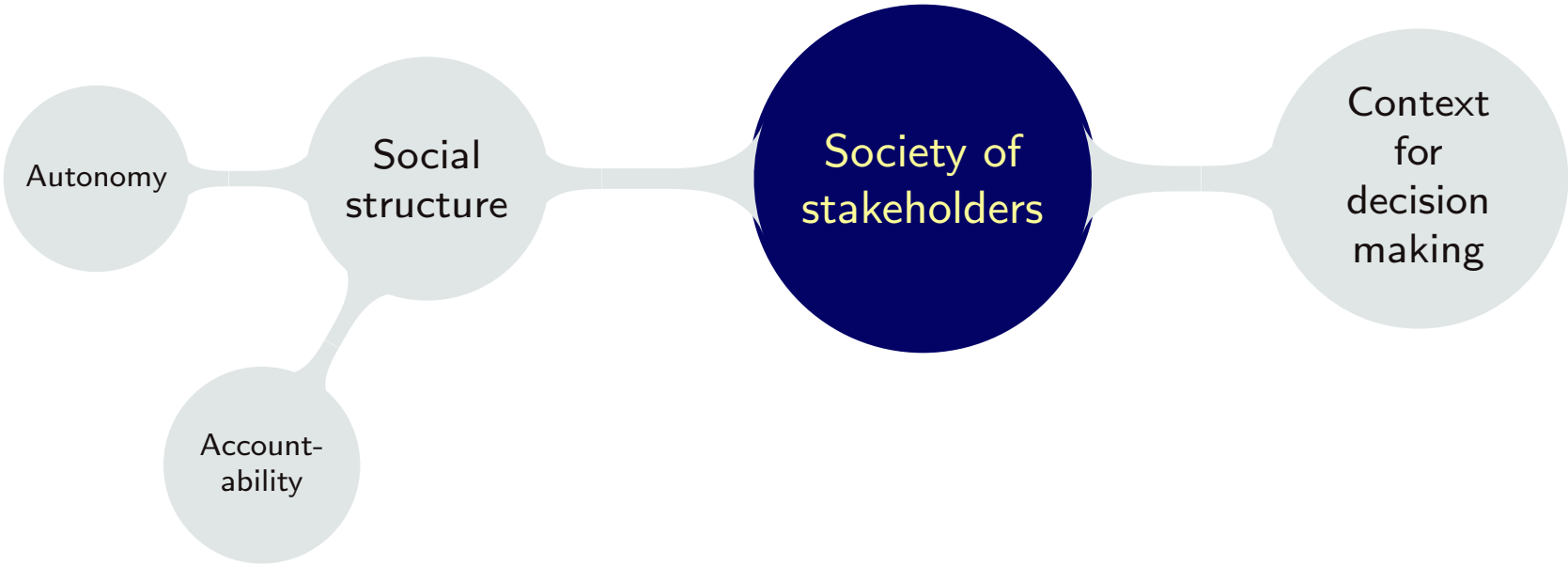


- Autonomy as a social construct; mirror of accountability
- Accountability rests with the social entity
- Powers and how they are exercised

Ethics in Society

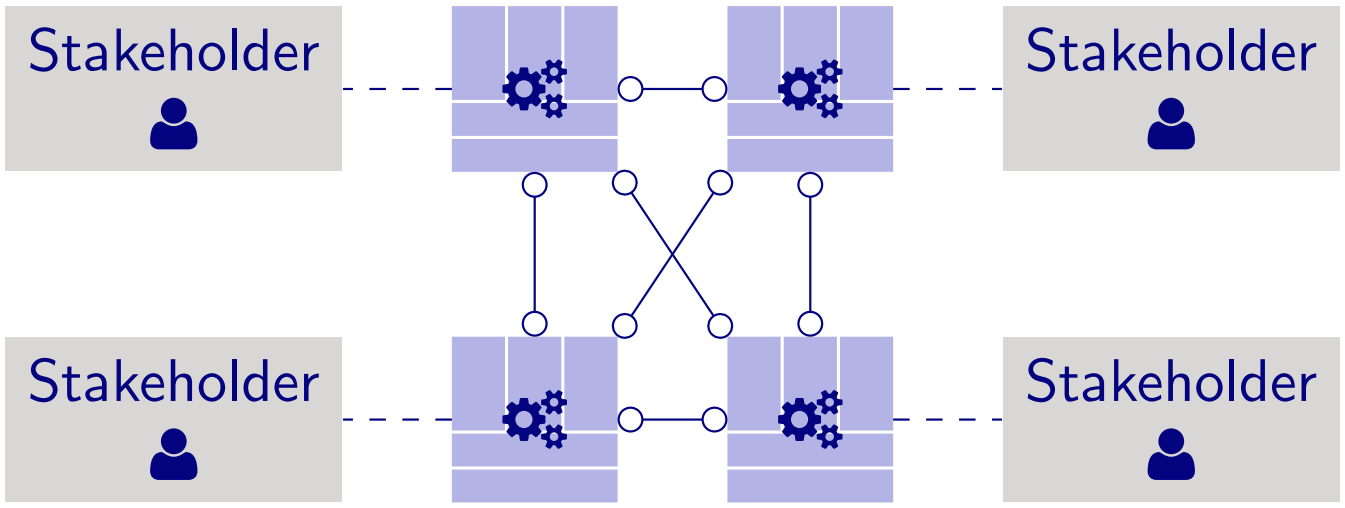
Ethical considerations and accountability arise in how social entities interact



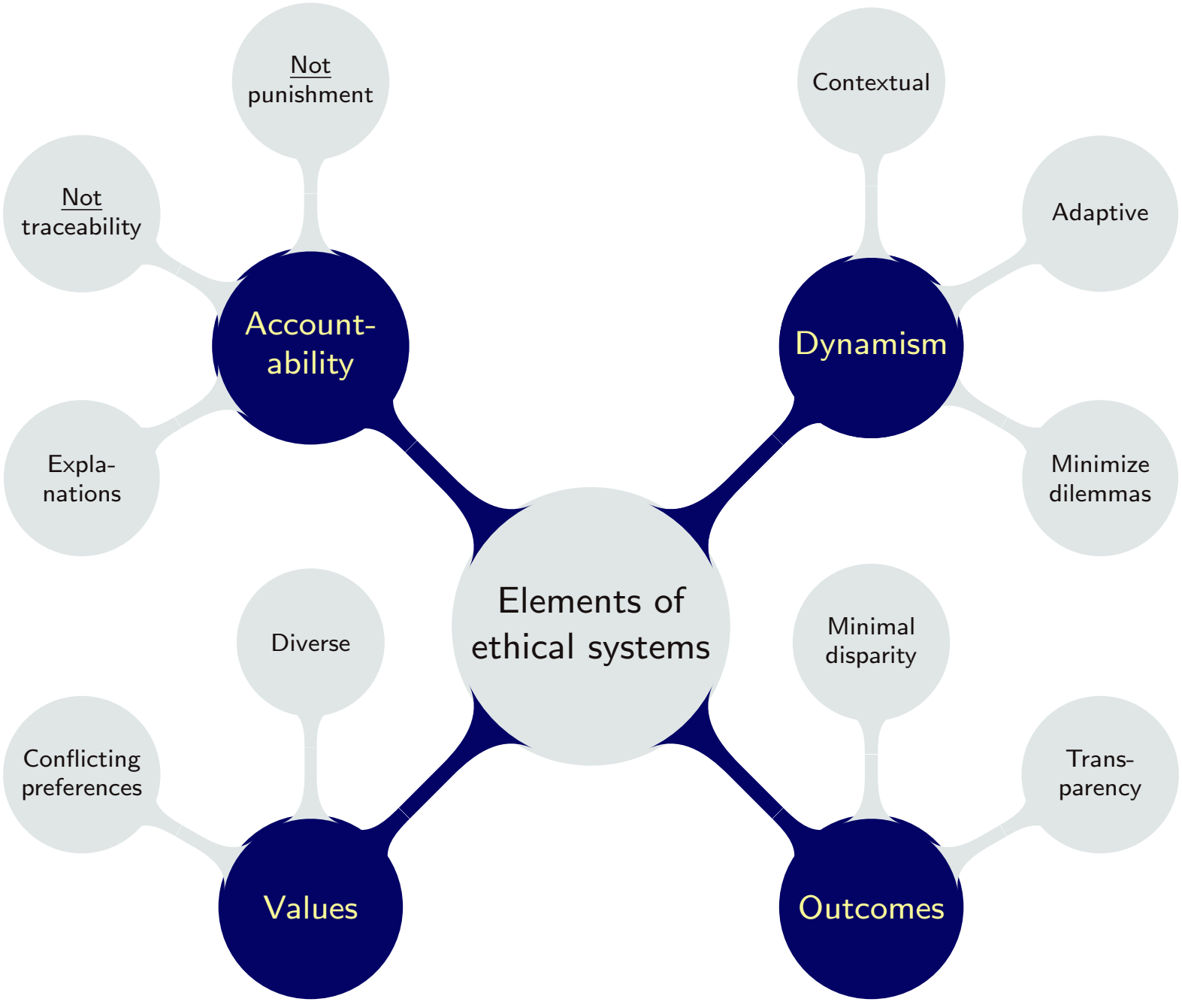


Ethics in Society with Agents Helping Stakeholders

Inherently decentralized



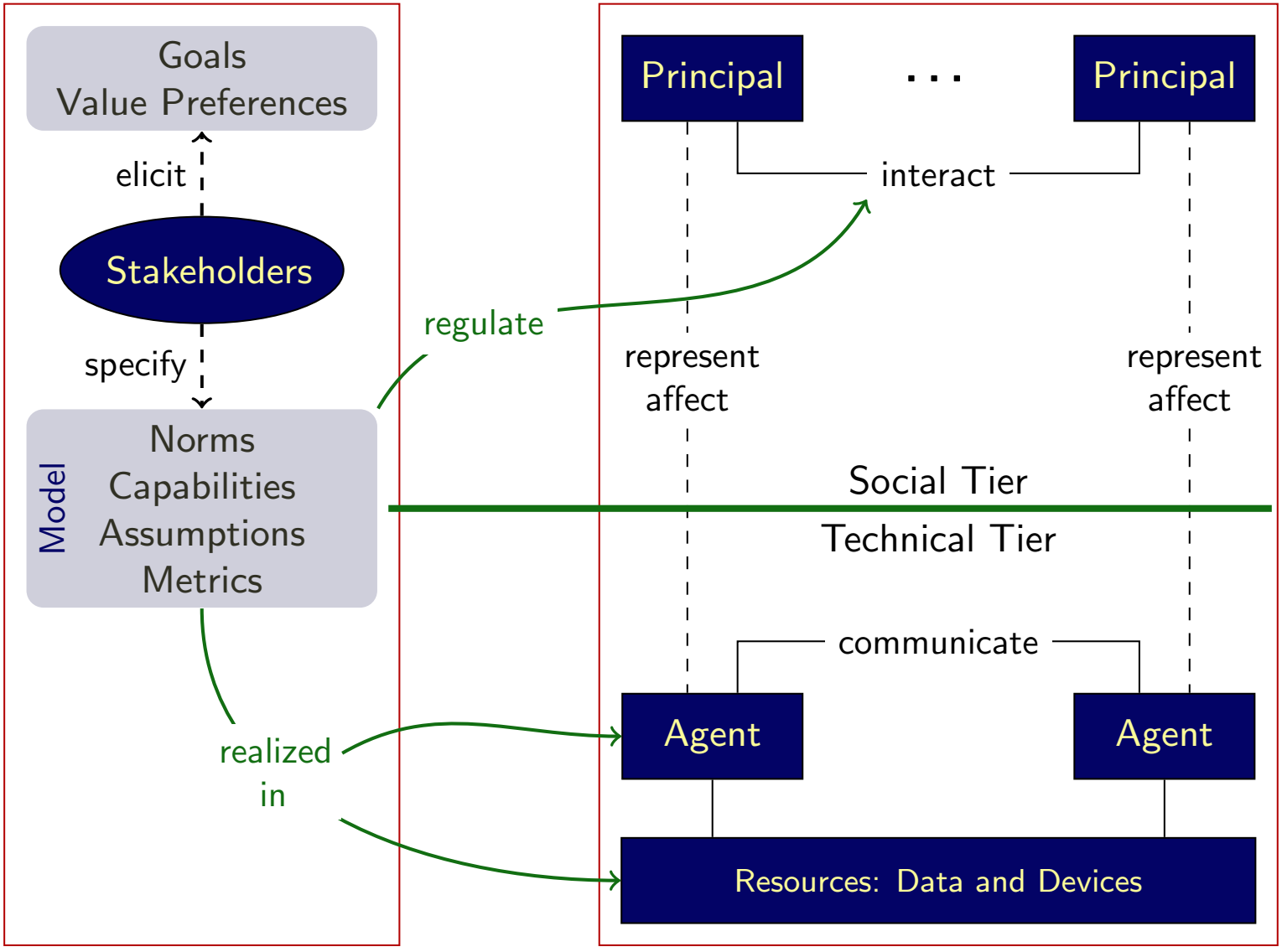
- Each agent reflects the autonomy of its (primary) stakeholder
- How can we realize a multiagent system based on the value preferences of its stakeholders?


Elements of ethical systems
Account-ability
Not punishment
Not traceability
Expla-nations
Dynamism
Contextual
Adaptive
Minimize dilemmas
Values
Diverse
Conflicting preferences
Outcomes
Minimal disparity
Trans-parency

Outline

Sociotechnical Systems

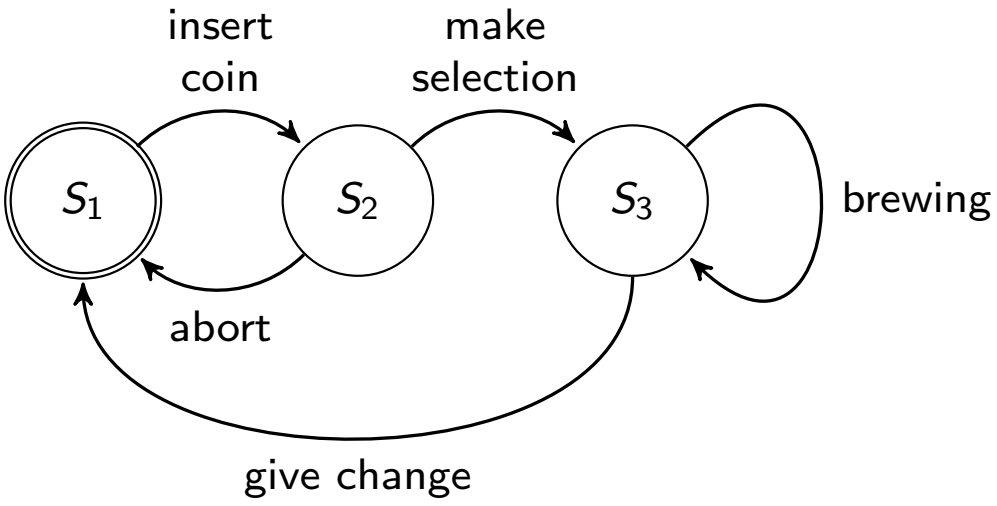
Current AI research: atomistic, single-agent decision-making focused on ethical dilemmas
Current social sciences research: Not computational in outlook

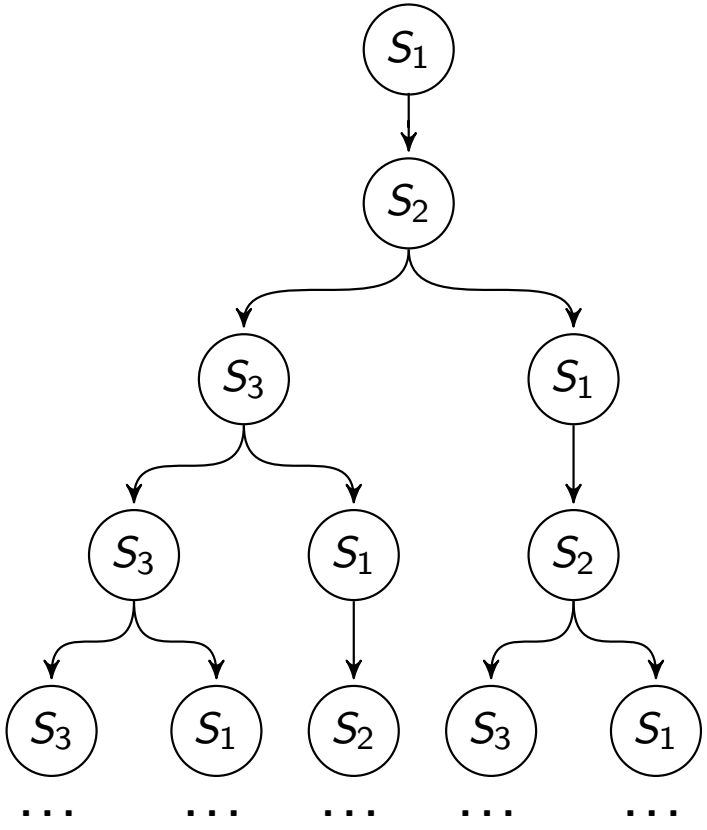


Vending Machine in Vienna

Conventional formal methods focus on technical artifacts

Emerson & Clarke: Model checking





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

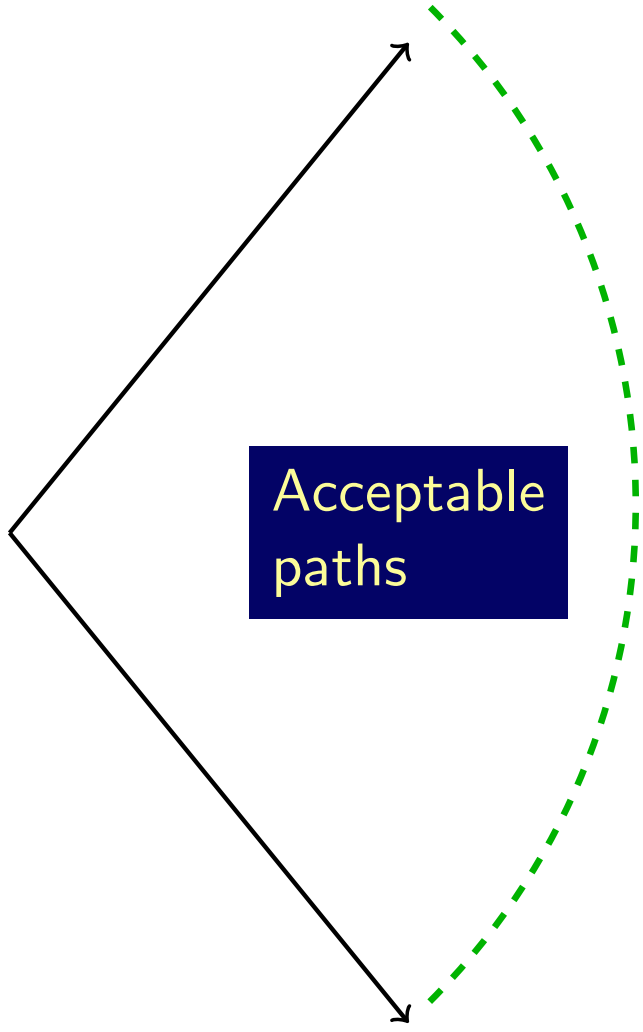
A[¬Brew U Coin]: On every path, no coffee is brewed prior to payment

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



Regimentation Ignores the Social Tier

Problems in ethics, trust, privacy, . . .

Smart contracts on blockchain have the same problem

Driver nearly crashes when her car suddenly shut down on a busy interstate because auto lender hit remote kill switch when she missed a payment

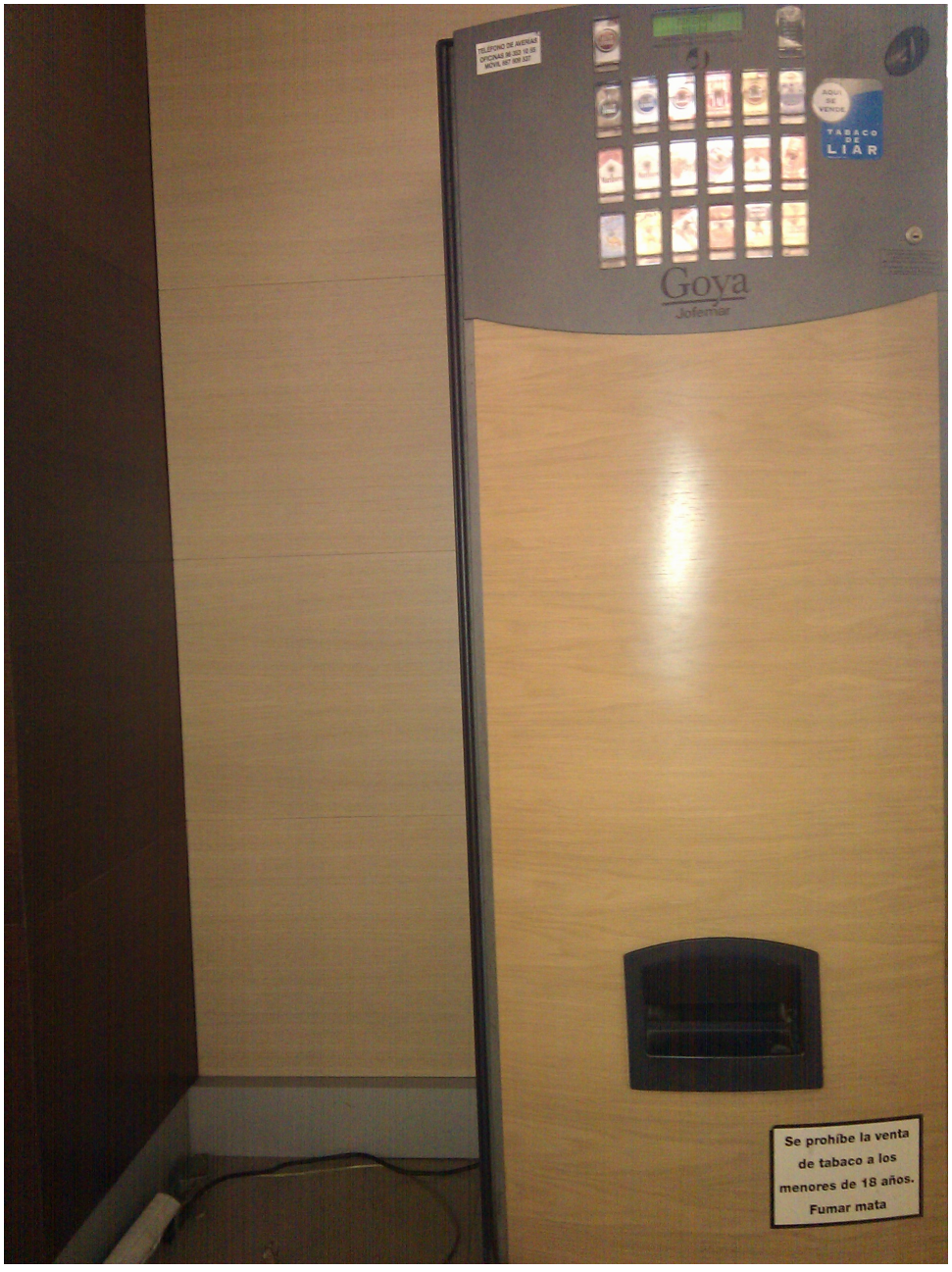
- T. Candice Smith had to have her car pushed out of on-coming traffic
- Starter Interrupt Devices allow auto lenders to 'shut down borrower's cars at any moment'
- The devices emit flashing lights, beeping noises and then shuts down the car and prevents it from starting
- These devices have been installed in more than two million vehicles

By CHARLENE ADAMS FOR MAILONLINE

PUBLISHED: 11:50 EST, 25 September 2014 | **UPDATED:** 14:21 EST, 25 September 2014

Vending Machine in Valencia

Users plus machines form a sociotechnical system



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

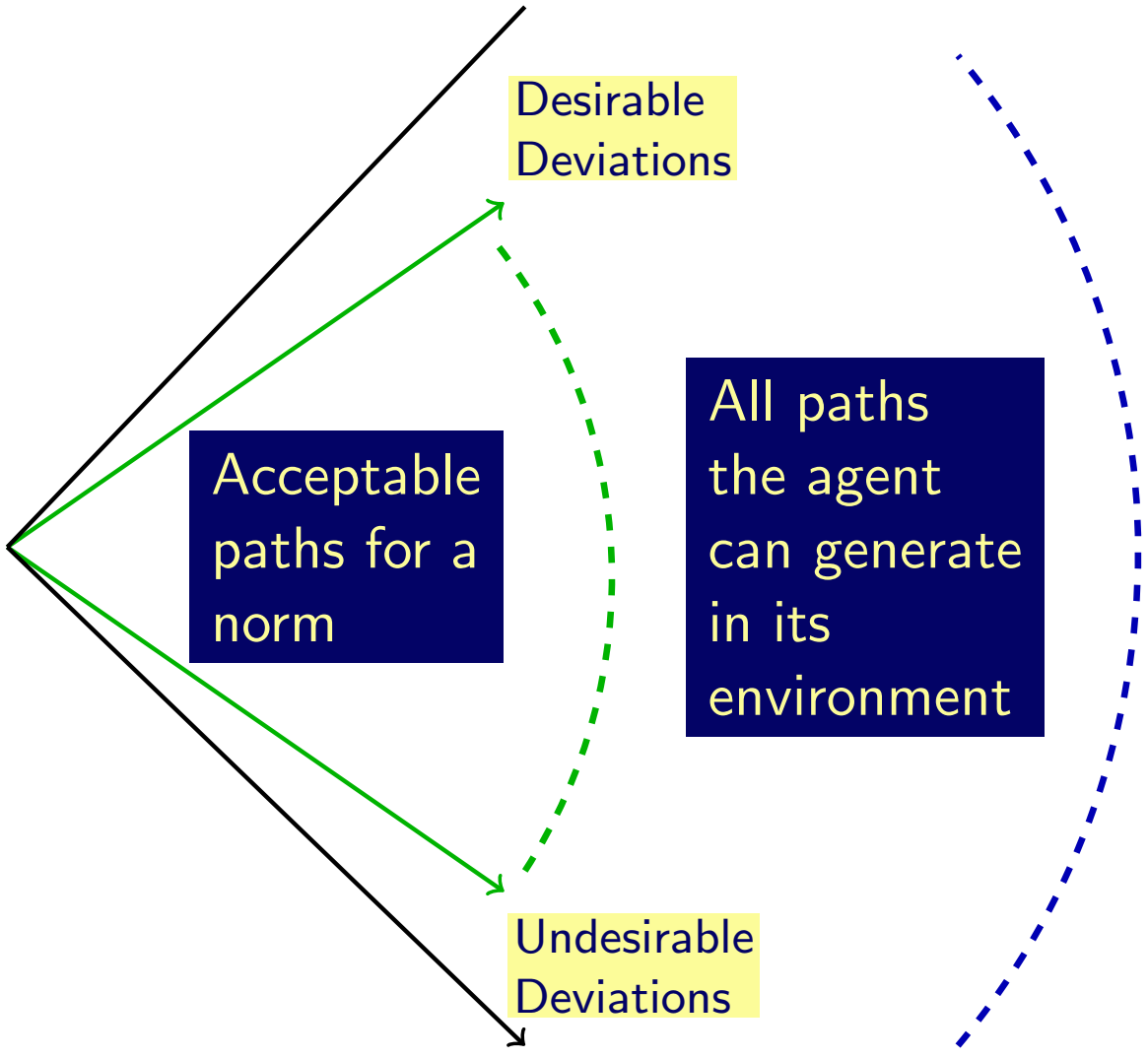
Vending Machine Close Up: Cigarettes!

Regulation: Violations are Possible

Appropriate assumption when dealing with autonomous parties

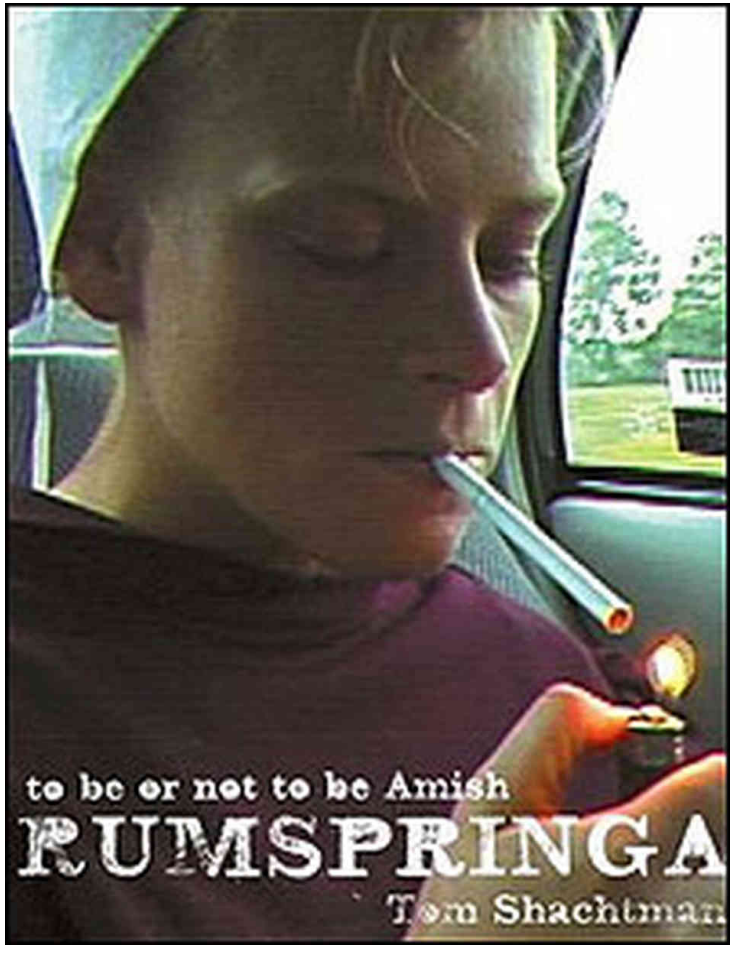
Regulation: Violations are Possible

Appropriate assumption when dealing with autonomous parties



Amish Rumspringa as a Metaphor

Kant's intuition: Autonomy is essential to ethics



- Technical architecture allows violations
- Social architecture discourages violations
  - Promotes innovation: Find new ways of behaving
  - Does it promote resilience?

Unlike Rumspringa

- Decentralization
- Continual autonomy (lifelong)

http://media.npr.org/books/images/2006/rumspringa200-d4edb2697bb547c7c12c73e2a7058289ce374ac9-s6-c30.jpg

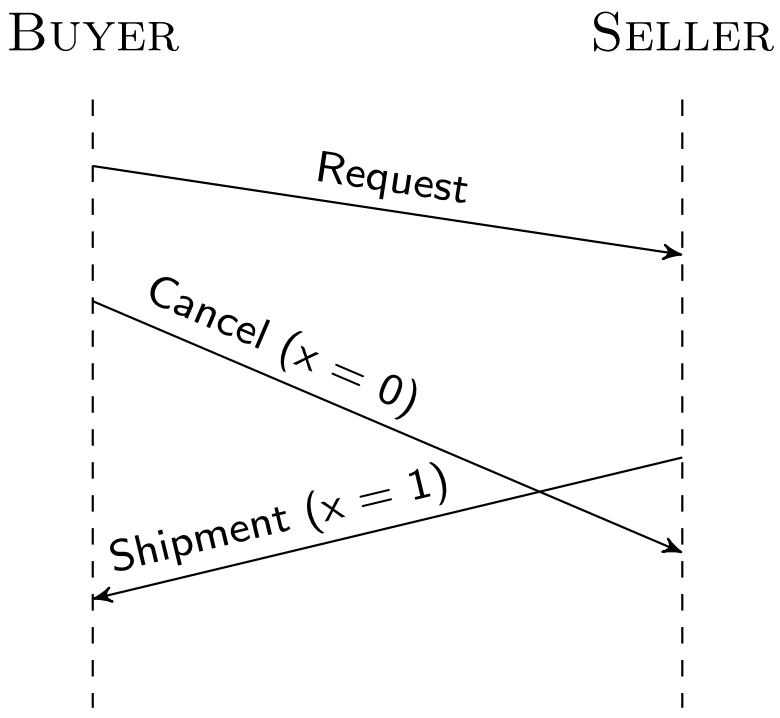
Outline

How Does Engineering Fare on Autonomy? 1 Models

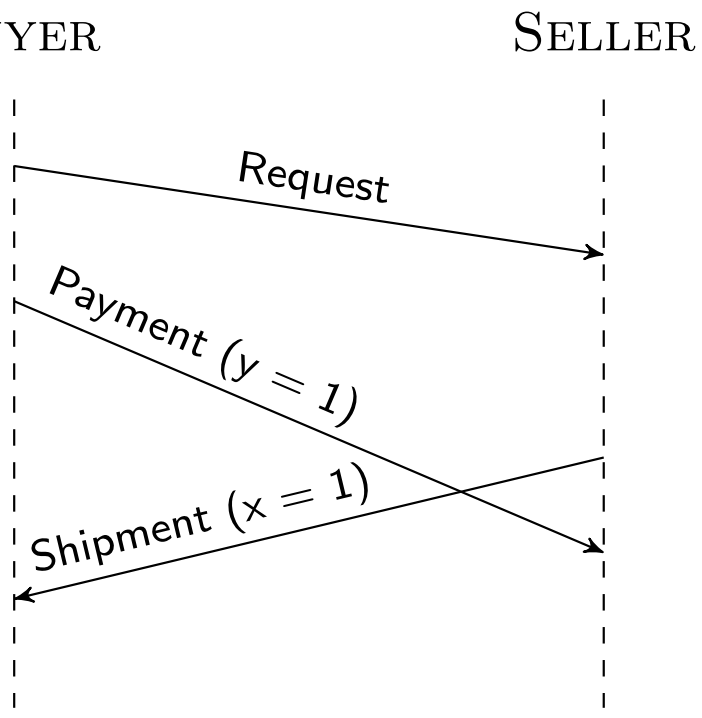
Mascardi, Ancona, Winikoff, Honda, Yoshida, . . .

Choice ≈ Order of observed events (also, what events occur)

Should disallow confusing choice



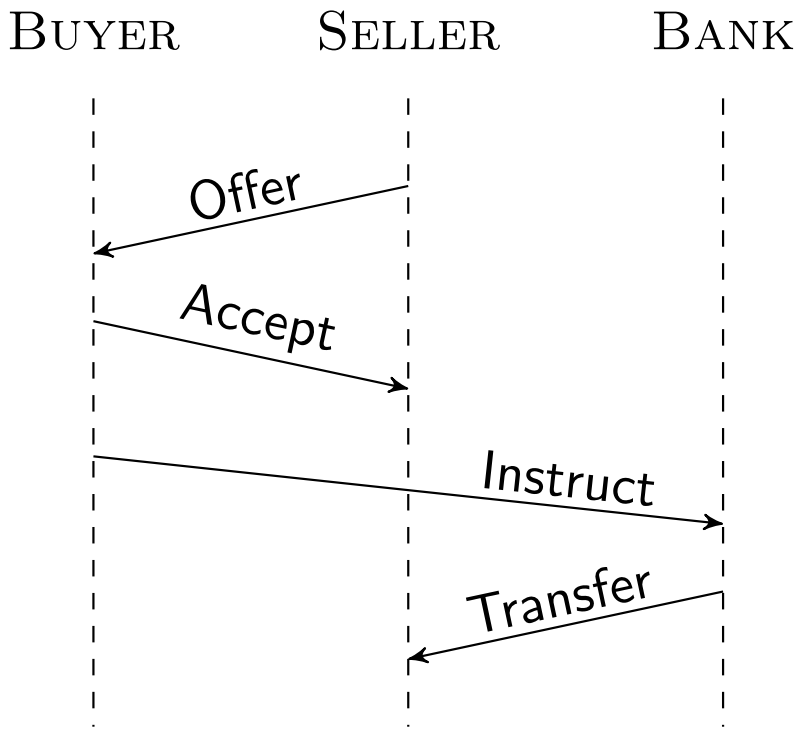
Should allow good choice

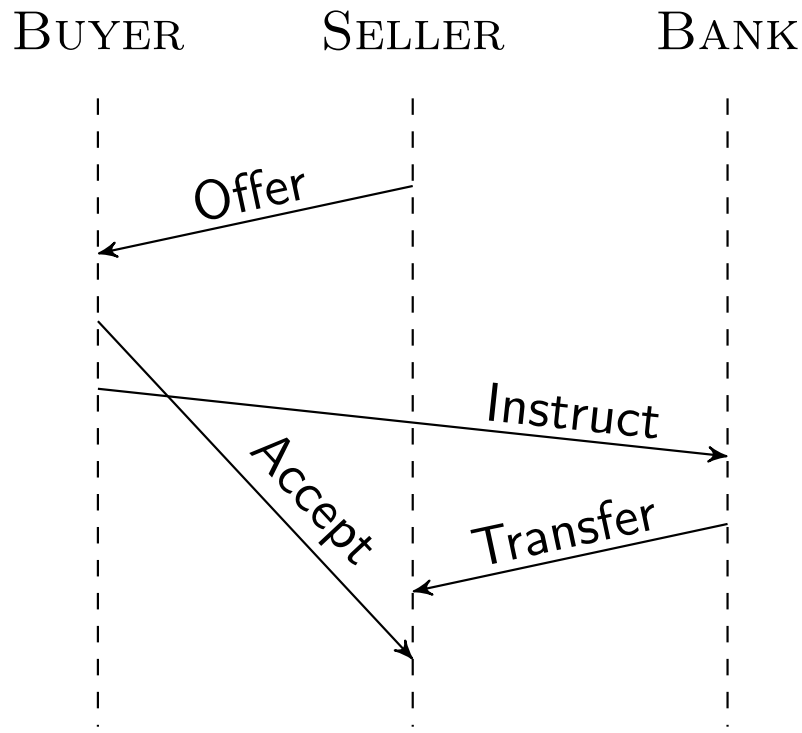


- Some choices must be consistent, so we must limit autonomy
- Traditional approaches require one agent to control each choice
- Limit autonomy of agents (other than an arbitrary one) for no reason

How Does Engineering Fare on Autonomy? 2 Models

Indirect-payment: Buyer to Bank to Seller: Must *Accept* occur before *Transfer*?
FIFO channels are inadequate
Traditional approaches force Seller to deny the contrary observation



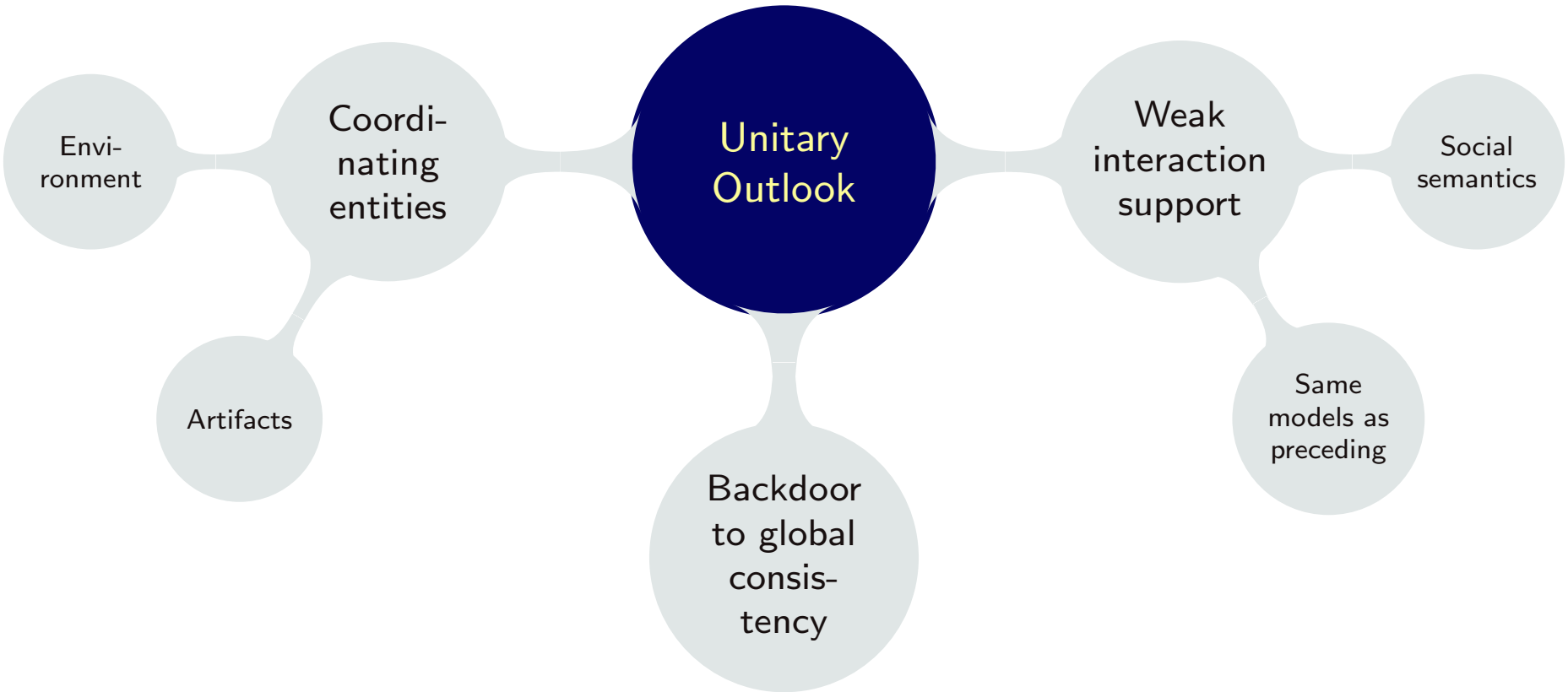


- Solipsism! Forces agents to deny observations, not exercise choice

How Does Engineering Fare on Autonomy? 3 Platforms

Bordini, Hübner, Ricci, Weyns, . . .

Great work, but . . .

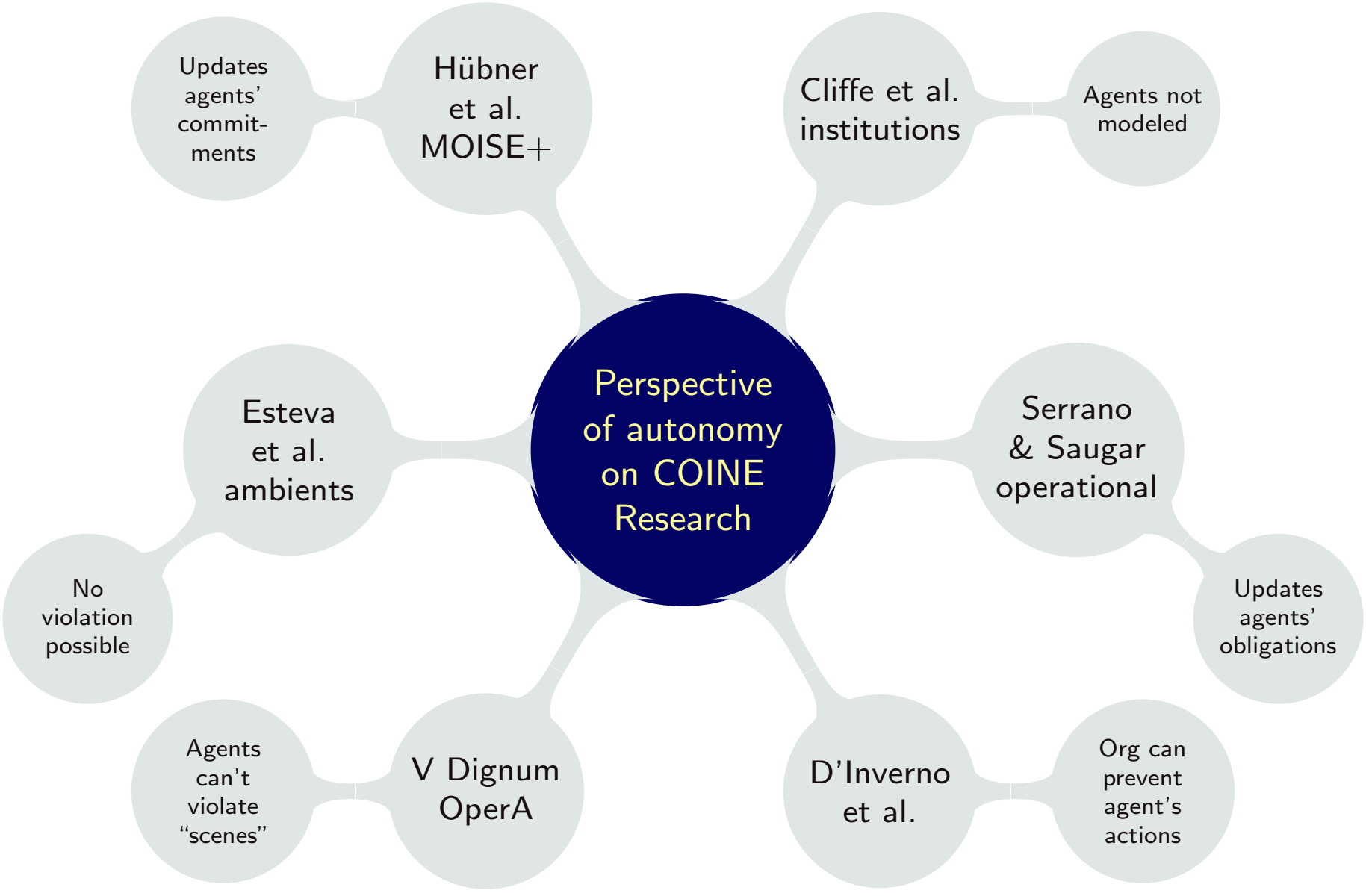


How Does Formal MAS Research Fare on Autonomy?

Tennenholtz, Ågotnes, Wooldridge, ...

How Does COINE Research Fare on Autonomy?

Luck, Sierra, Sichman, F Dignum, Padget, Rodríguez-Aguilar, . . .



Norms as Relationships help Ground Accountability

Zero-agent and one-agent obligations are inadequate

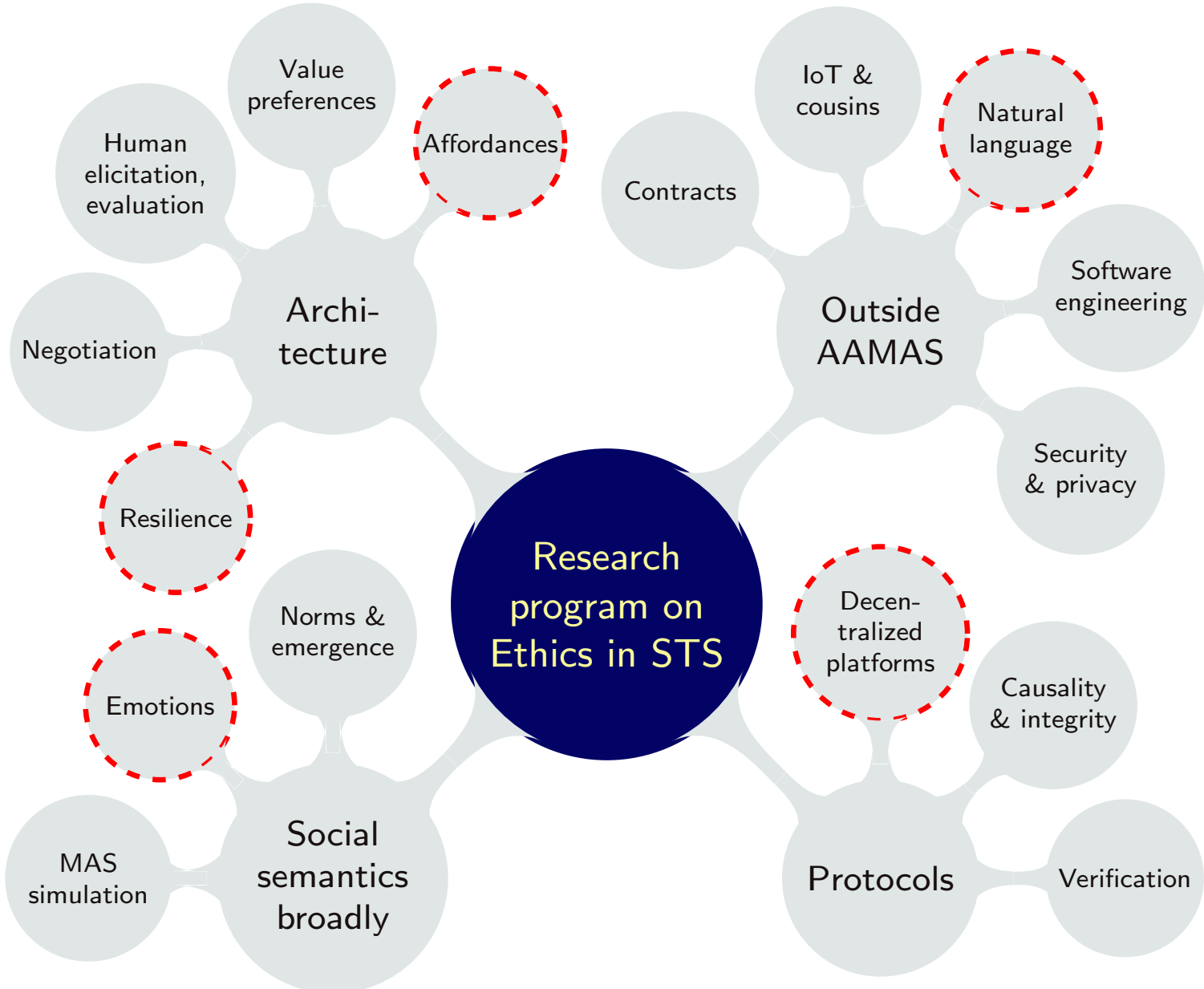
Accountability: ≥ 2 parties: one to call another to account for its actions

- Explanations and arguments to override prima facie expectations
- Not blame and sanction: subsequent to accounting
- Not traceability: a supporting mechanism

Outline

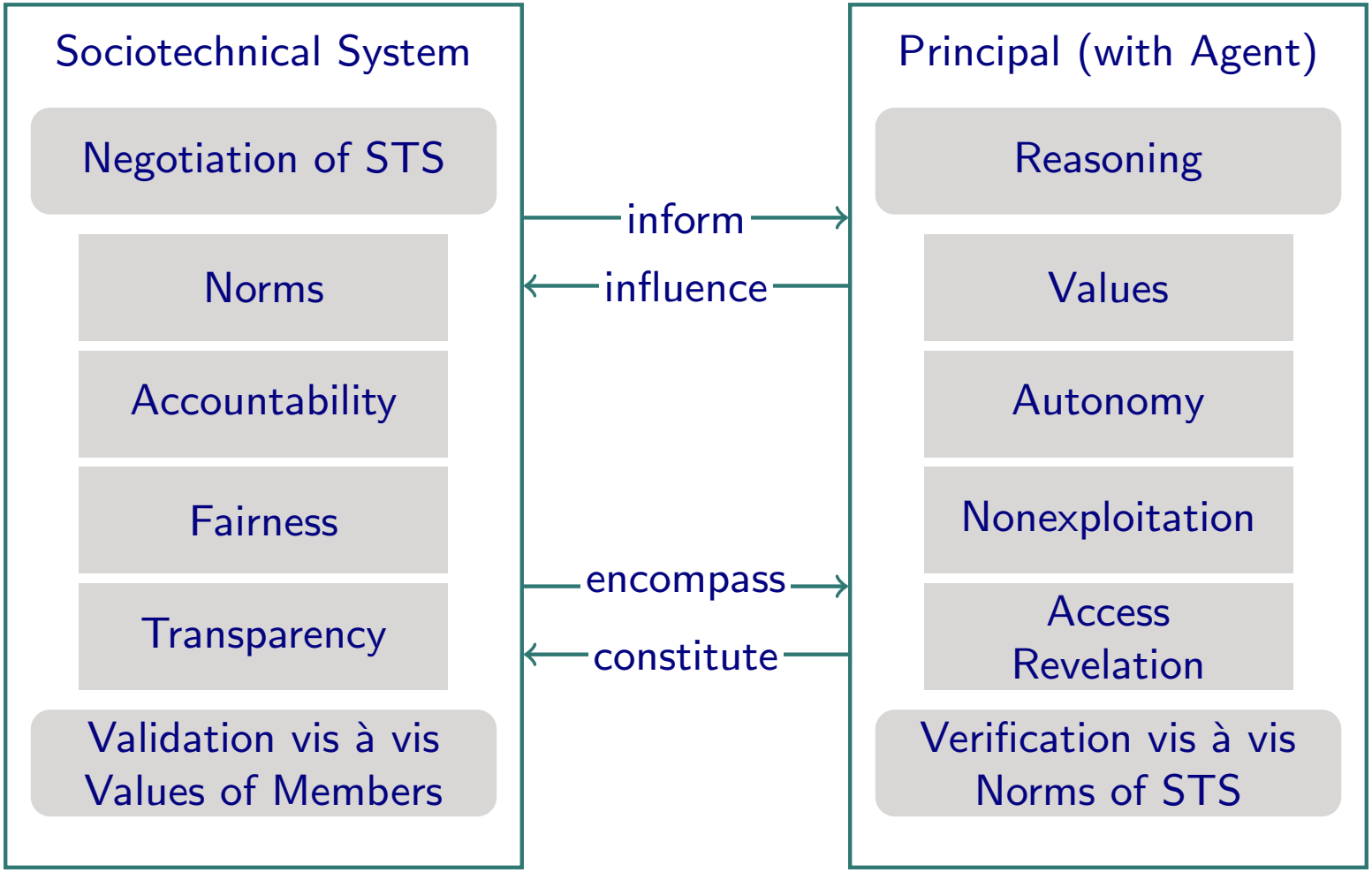
Hints at a Research Program and Recent Progress

My view: Redo everything with autonomy and decentralization!



Challenge: Governance

Continually align sociotechnical systems and principals
Judgments of ethicality of STS: Relative to principals' values
Judgments of compliance by principals: Relative to STS



Challenge: Ethics in the Large

Values, outcomes, and accountability from a sociotechnical perspective

Elements of Ethics: From Agents to Systems

	Agent Level	System Level
Scope	Individual	Individual in society
Autonomy	Intelligence and complexity	Decision making in social relationships
Transparency	Of data and algorithms	Of norms and incentives
Bases of Trust	Construction and traceability	Norms and accountability
Fairness	Preset criteria: Statistics	Reasoning about others' outcomes
Focus	Dilemmas for individuals	System properties, e.g., justice

Thanks!

- Current students: Samuel Christie, Hui Guo, Zhen Guo
- Former students: Nirav Ajmeri, Amit Chopra, Nirmit Desai, Xibin Gao, Scott Gerard, Chung-Wei Hang, Chris Hazard, Anup Kalia, Mehdi Mashayekhi, Michael Maximilien, Pradeep Murukannaiah, Derek Sollenberger, Pankaj Telang, Feng Wan, Yonghong Wang, Pınar Yolum, Bin Yu, Guangchao Yuan, Zhe Zhang
- Collaborators: Mike Huhns, Matt Arrott, Matteo Baldoni, Cristina Baroglio, Jon Doyle, Özgür Kafalı, Felipe Meneguzzi, John Mylopoulos, Simon Parsons, Jaime Sichman, Jose Such, Neil Yorke-Smith
- EMAS, Norms, Trust, Orgs, Ethics colleagues: Natasha Alechina, Gul Agha, Alex Artikis, Jamal Bentahar, Guido Boella, Rafael Bordini, Olivier Boissier, Cristiano Castelfranchi, Robin Cohen, Marco Colombetti, Vince Conitzer, Stephen Cranefield, Mehdi Dastani, Louise Dennis, Frank Dignum, Virginia Dignum, Ed Durfee, Amal El Fallah Seghrouchni, Rino Falcone, Nicoletta Fornara, Aditya Ghose, Guido Governatori, Jomi Hübner, Andrew Jones, Catholijn Jonker, Yves Lespérance, Brian Logan, Maite López-Sánchez, Emiliano Lorini, Mike Luck, Viviana Mascardi, Marco Montali, Tim Norman, Julian Padget, Jeremy Pitt, Alessandro Ricci, Juan Rodríguez-Aguilar, Tony Savarimuthu, Sandip Sen, Carles Sierra, Paolo Torroni, Leon van der Torre, Birna van Riemsdijk, Wamberto Vasconcelos, George Vouros, Michael Winikoff, Franco Zambonelli, Jie Zhang
- Sponsors: National Science Foundation, Army Research Ofice and Laboratory, DARPA, Department of Defense, Cisco, IBM, Xerox