Ethics in Multiagent Systems

Nirav Ajmeri[1]
(with help from Pradeep Murukannaiah[2] and Munindar P. Singh[3])

[1]University of Bristol

[2]Delft University of Technology

[3]North Carolina State University

July 2023

Outline and Schedule (120 minutes)

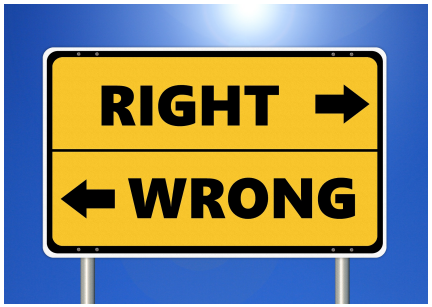
10		M	Motivation: Ethical Sociotechnical Systems
	Foundations		
10	Philosophy	M	Background on ethics (virtue, utilitarianism, Rawls . . .)
10	Law, Political Science	M	Sociotechnical systems
10	Psychology	P	Preferences and values (Rokeach, Schwartz)
	Techniques		
10	Artificial Intelligence	N	Specifying an ethical STS
10			**Q&A and Exercise**

	Techniques...		
15	Software Engineering	P	Value sensitive design
15	Operations Research	N	Reasoning about ethics (balance self and others)
	Research Directions		
3	Formal methods	N	Verification and simulation
3	Psychology	N	Emotions and equity
3	Machine Learning	P	Elicitation (surveys; active value learning; inverse RL)
3	Artificial Intelligence	P	Uniting individual and societal perspectives
3	Law	M	Law and consent
	Synthesis		
5		All	Summary and concluding remarks
10			**Q&A**

Outline and Schedule

What is Ethics?

The field of ethics involves systematizing, defending, and recommending concepts of right and wrong behavior [Fieser, The Internet Encyclopedia of Philosophy]

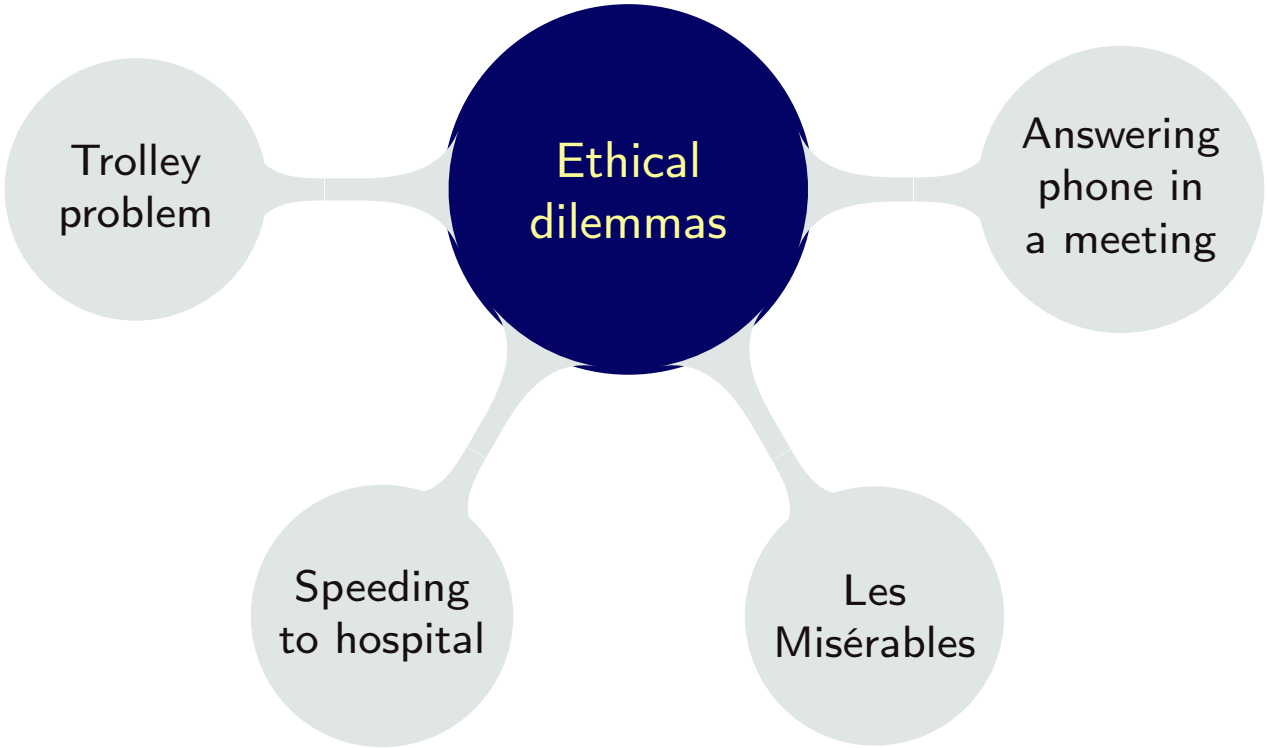


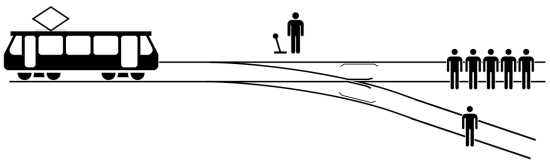
Classical ethics: Founded on economics and politics

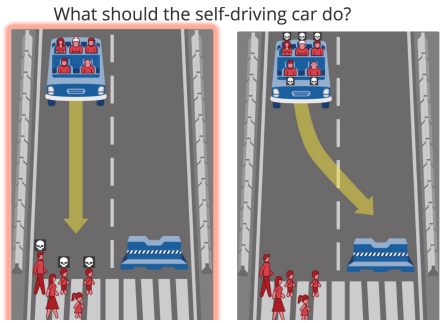
The formation of the individual character (ethos) is intrinsically related to the others, as well as to the tasks of administration of work within the family (oikos), which eventually, expands into the framework of the public space (poleis) [Ethically Aligned Design, IEEE]

Ethos
Oikos
Poleis

Ethical Dilemmas: No (Obviously) Good Choices

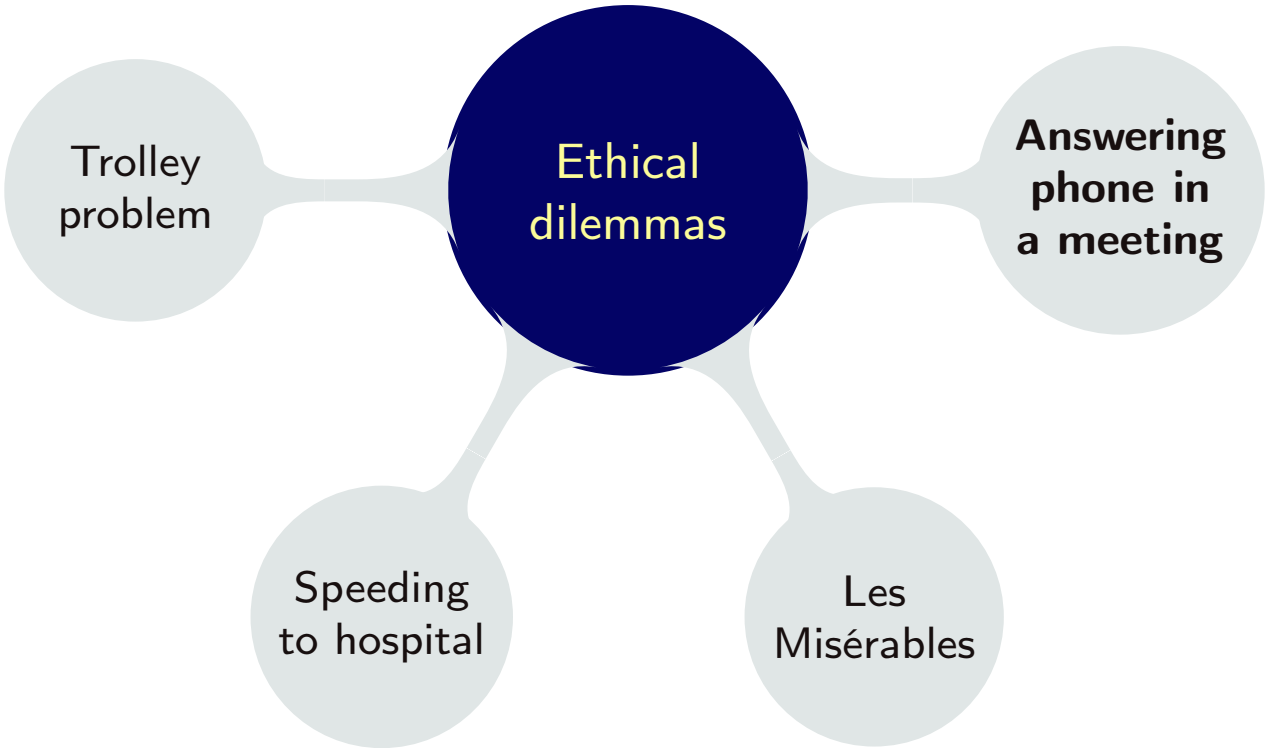






Ethical Dilemmas: No (Obviously) Good Choices

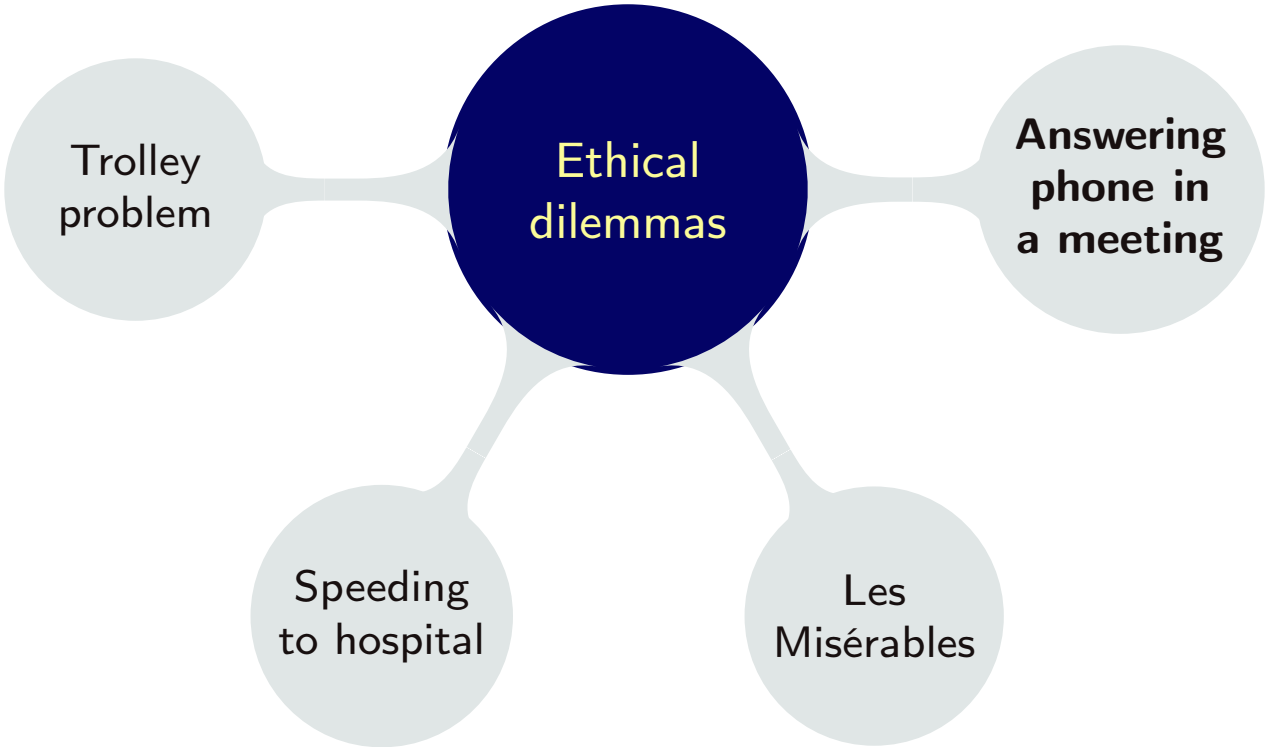
Ethical dilemmas arise not only in hypothetical or extreme scenarios but also in mundane scenarios



"Someone still has his cellphone on."

Ethical Dilemmas: No (Obviously) Good Choices

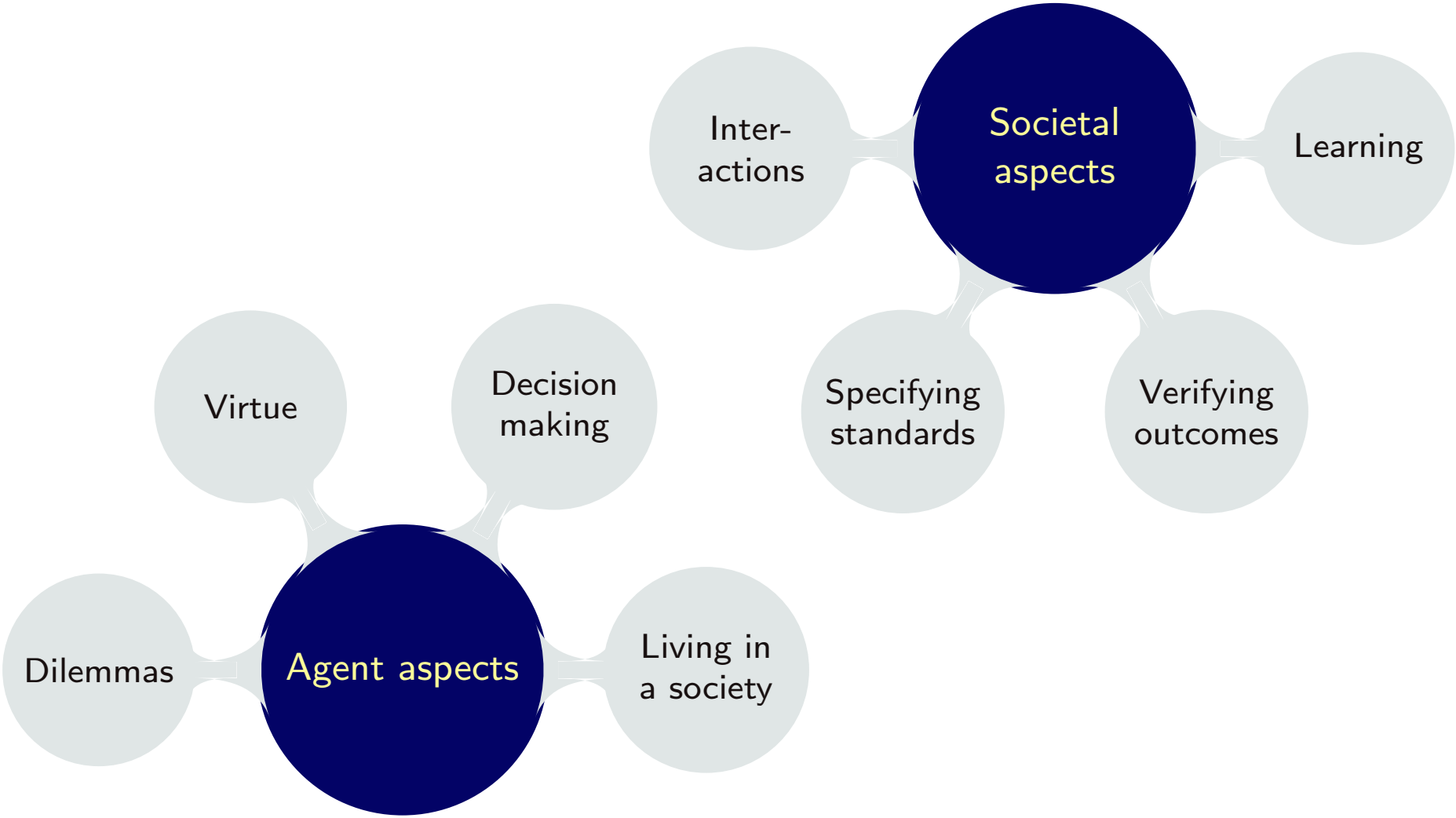
Ethical dilemmas arise not only in hypothetical or extreme scenarios but also in mundane scenarios

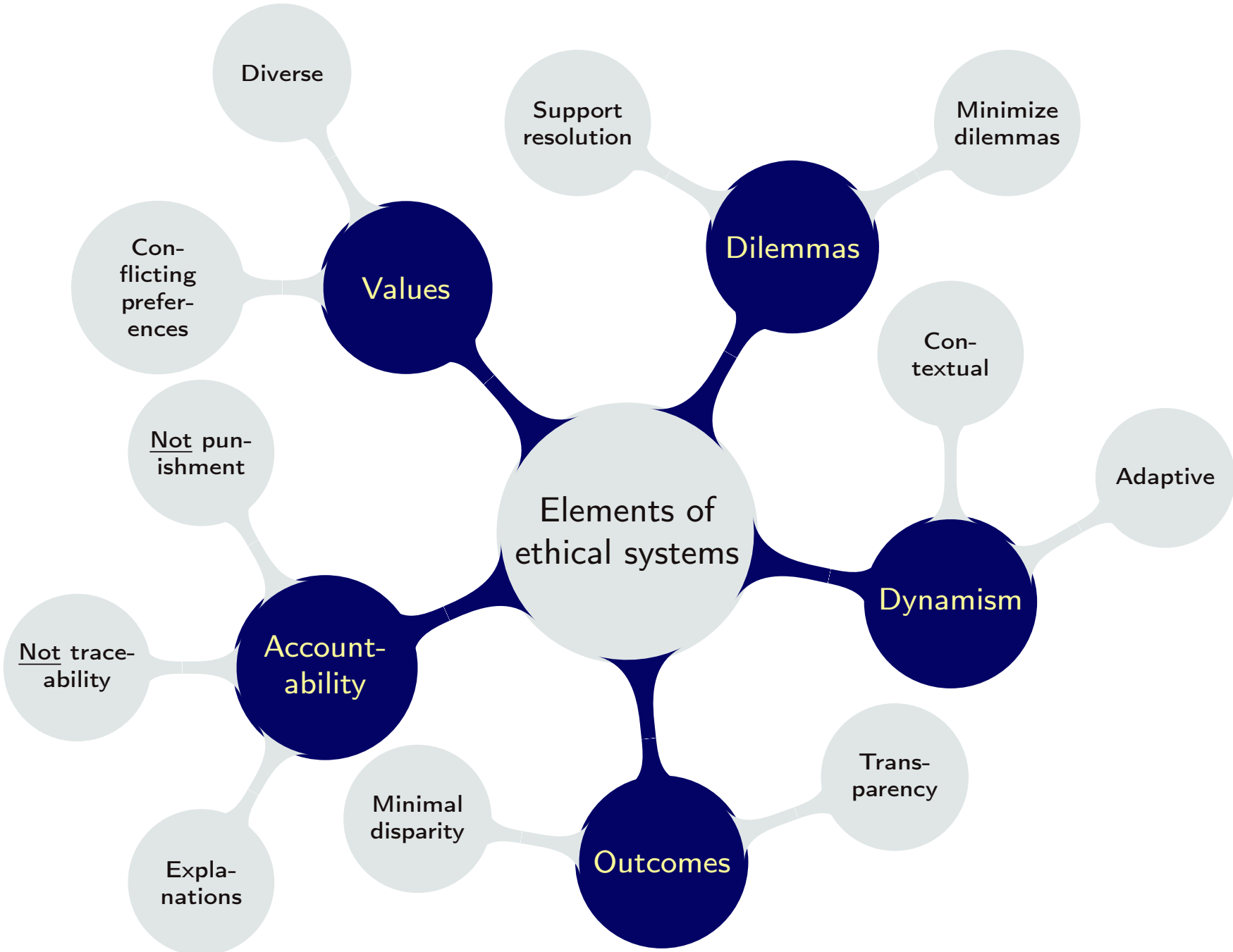


Ethics is inherently a multiagent concern

Ethics in Multiagent Systems

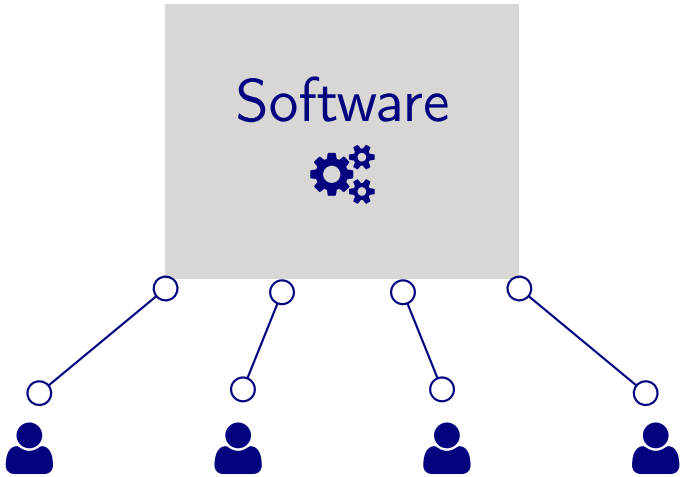
Ethics is an inherently multiagent concern, yet current approaches focus on single agents




Elements of ethical systems
Values
Diverse
Con-flicting prefer-ences
Dilemmas
Support resolution
Minimize dilemmas
Dynamism
Con-textual
Adaptive
Account-ability
Not pun-ishment
Not trace-ability
Outcomes
Minimal disparity
Trans-parency
Expla-nations

"Ethics" of a Central Technical Entity

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

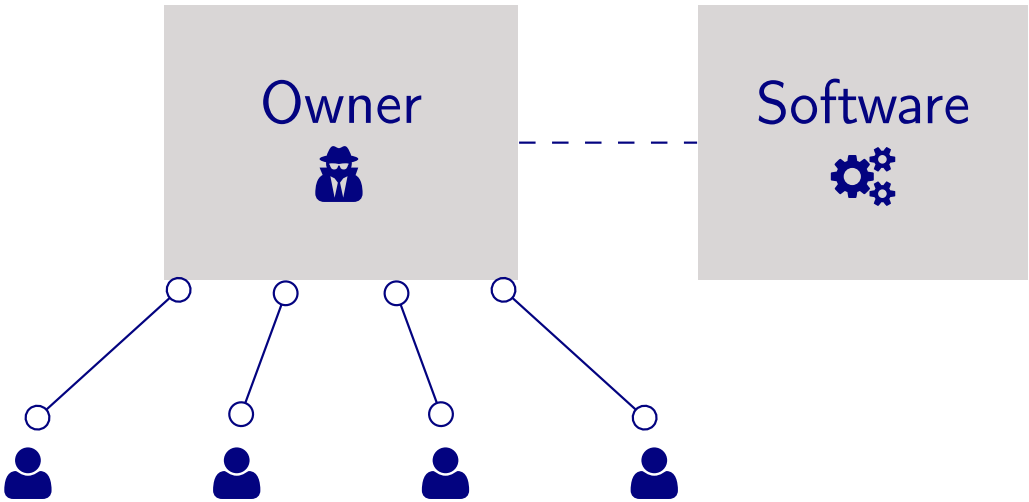


- Autonomy is defined as automation: complexity and intelligence
- Dilemmas à la trolley problems approached in an atomistic manner

Ethics of a Social Entity Equipped with Software

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

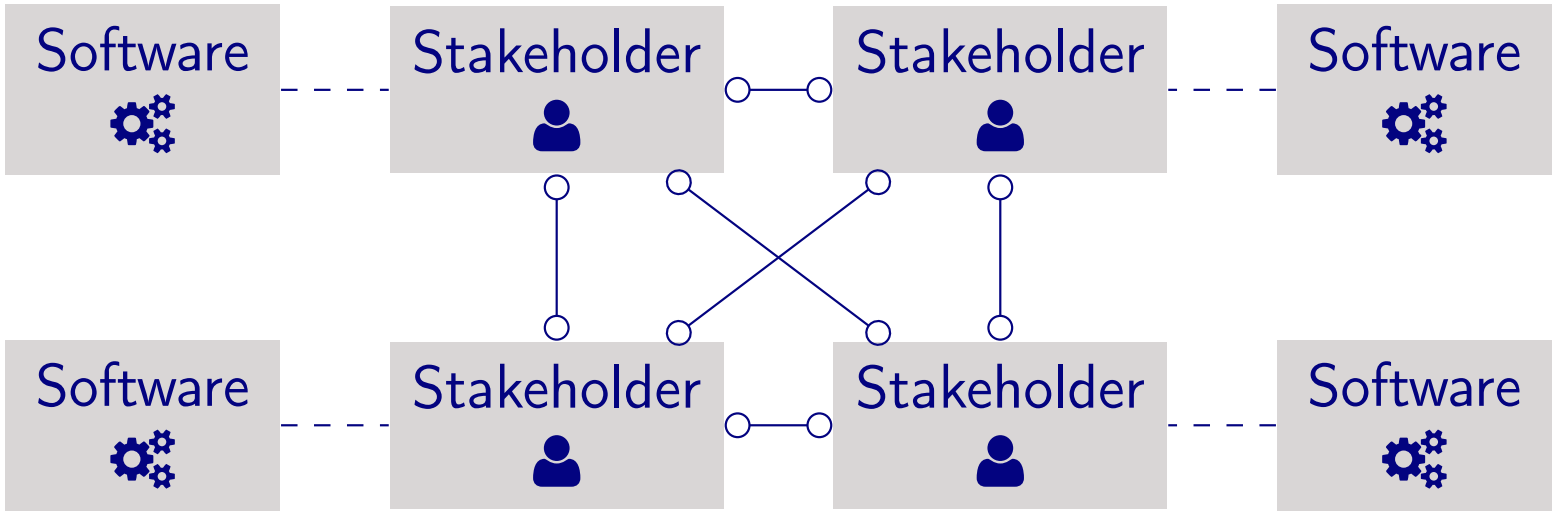
Ethical concerns focused on social entity

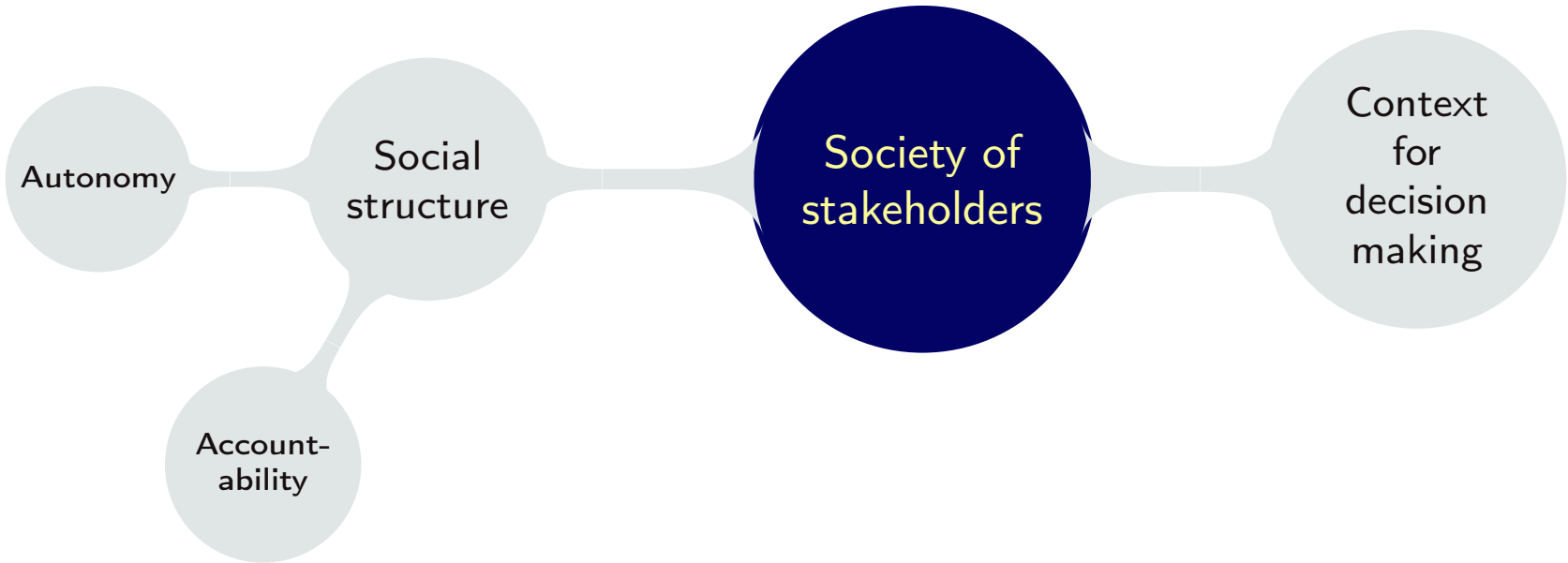


- Autonomy as a social construct; mirror of accountability
- Accountability rests with the social entity
- Powers and how they are exercised

Ethics in Society: Ethics is a Cousin of Governance

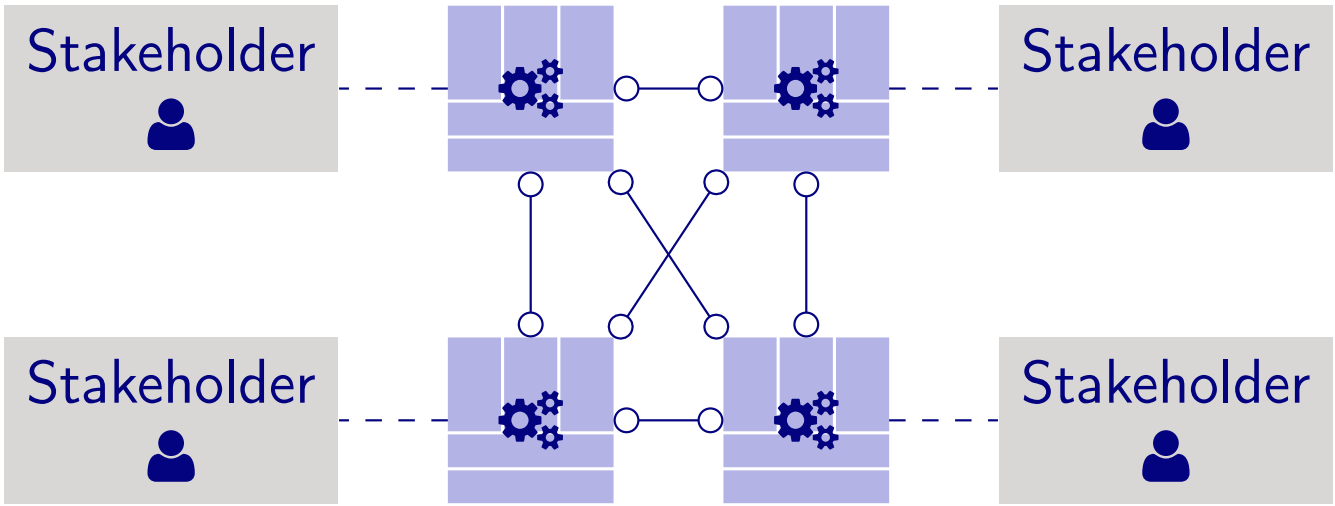
Ethical considerations and accountability arise in how social entities interact





Ethics in Society with SIPAs

SIPA: Socially intelligent (personal) agent



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

Outline and Schedule

Aristotlean Virtue Ethics

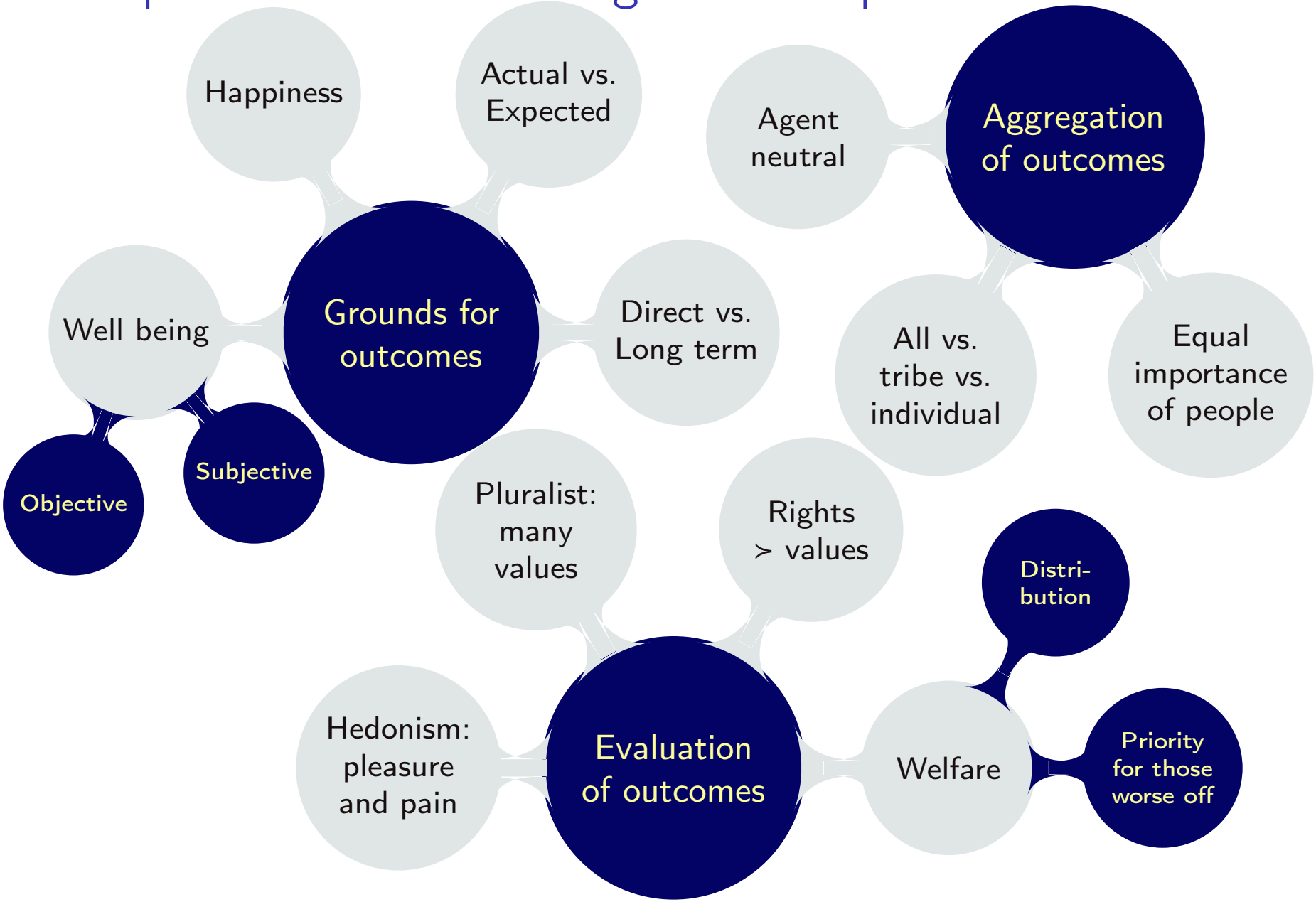
Eudaimonia (Well being; happiness; flourishing)

Teleological theory: The goal is a good life

Kantian Deontological Ethics

Thomas Hill

Consequentialism: Moral Rightness Depends on Outcomes



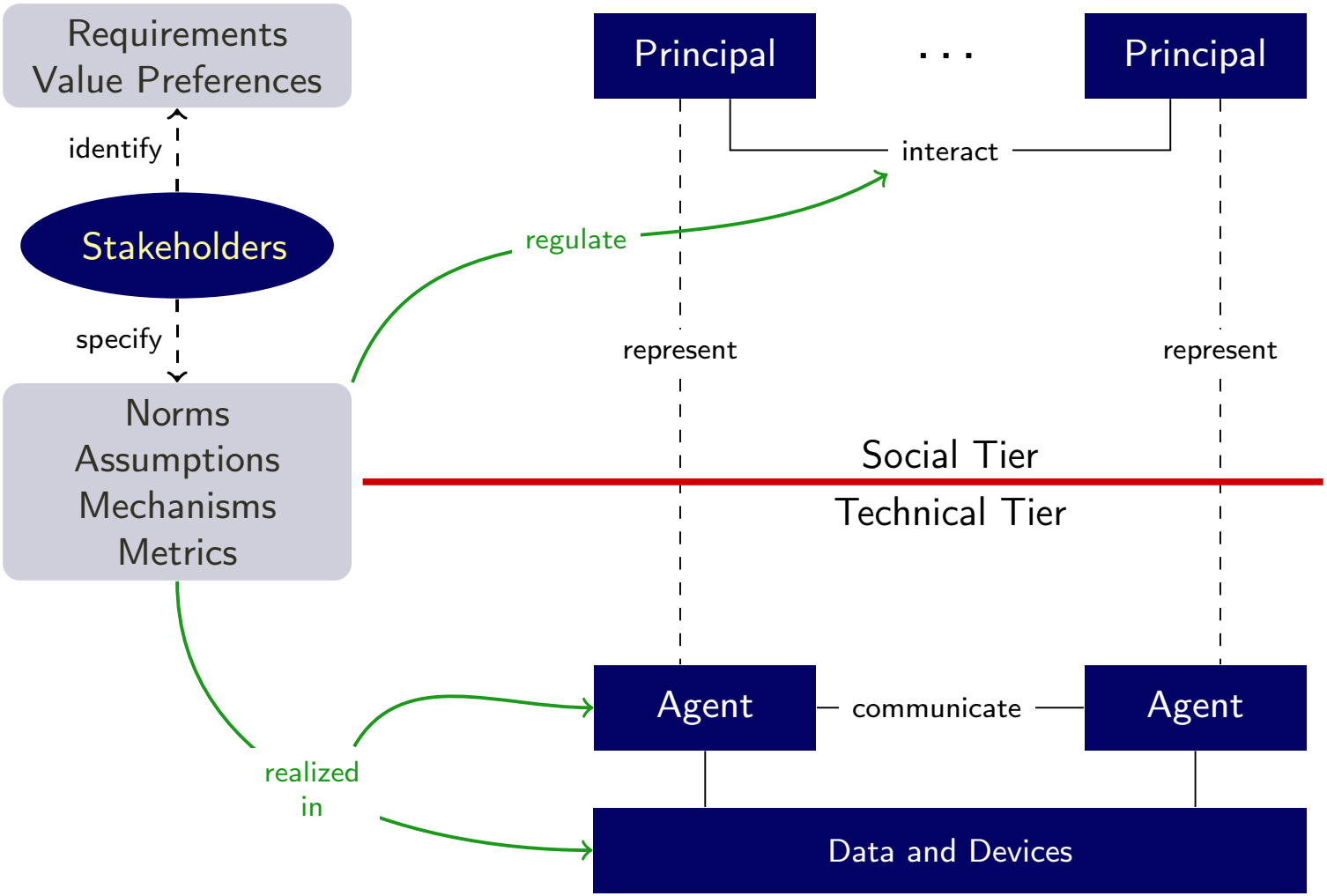
Social Welfare

Rawls ("political not metaphysical"); Adler

Outline and Schedule

Sociotechnical Systems

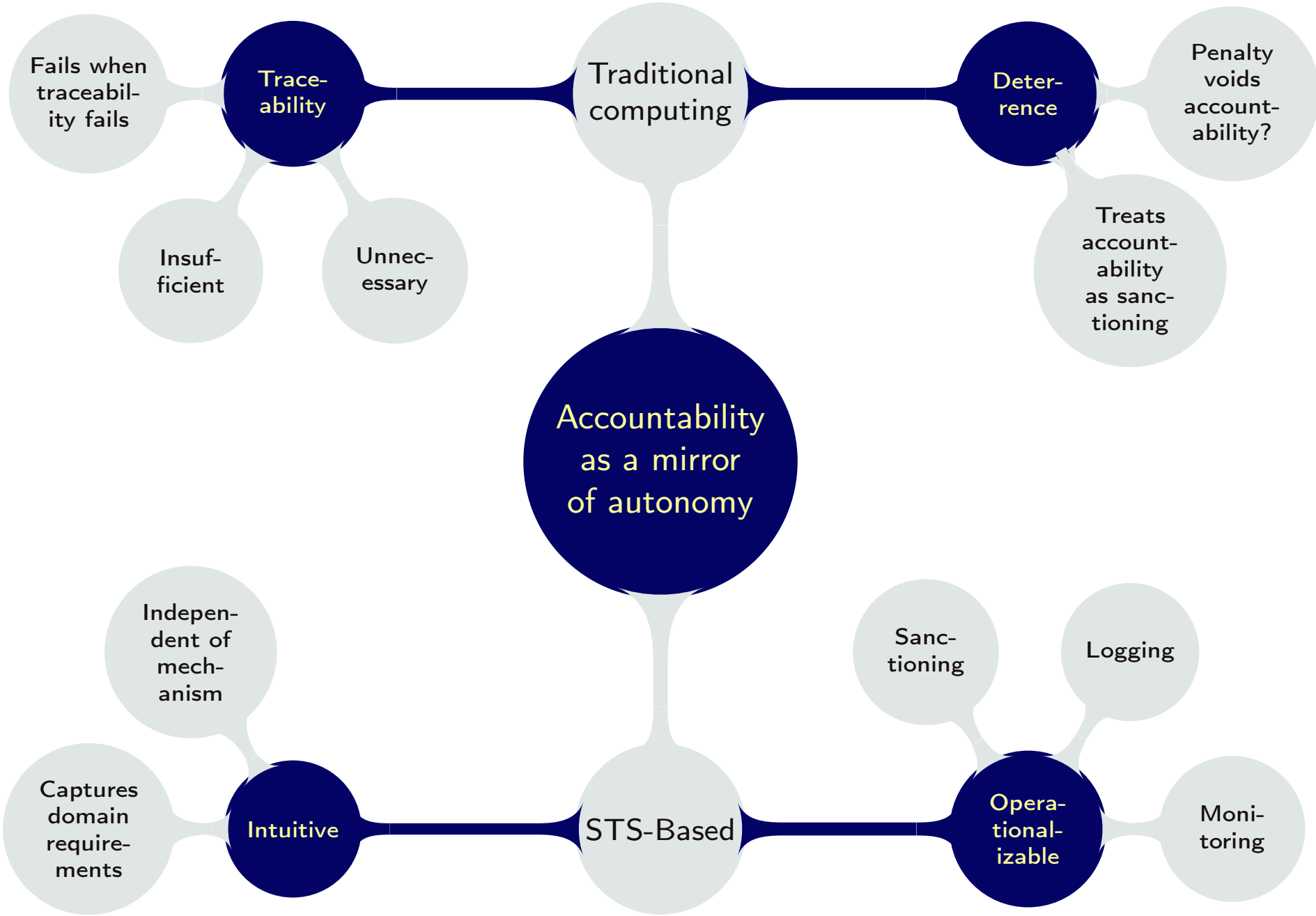
Current AI research: atomistic, single-agent decision-making focused on ethical dilemmas
Current social sciences research: Not computational in outlook

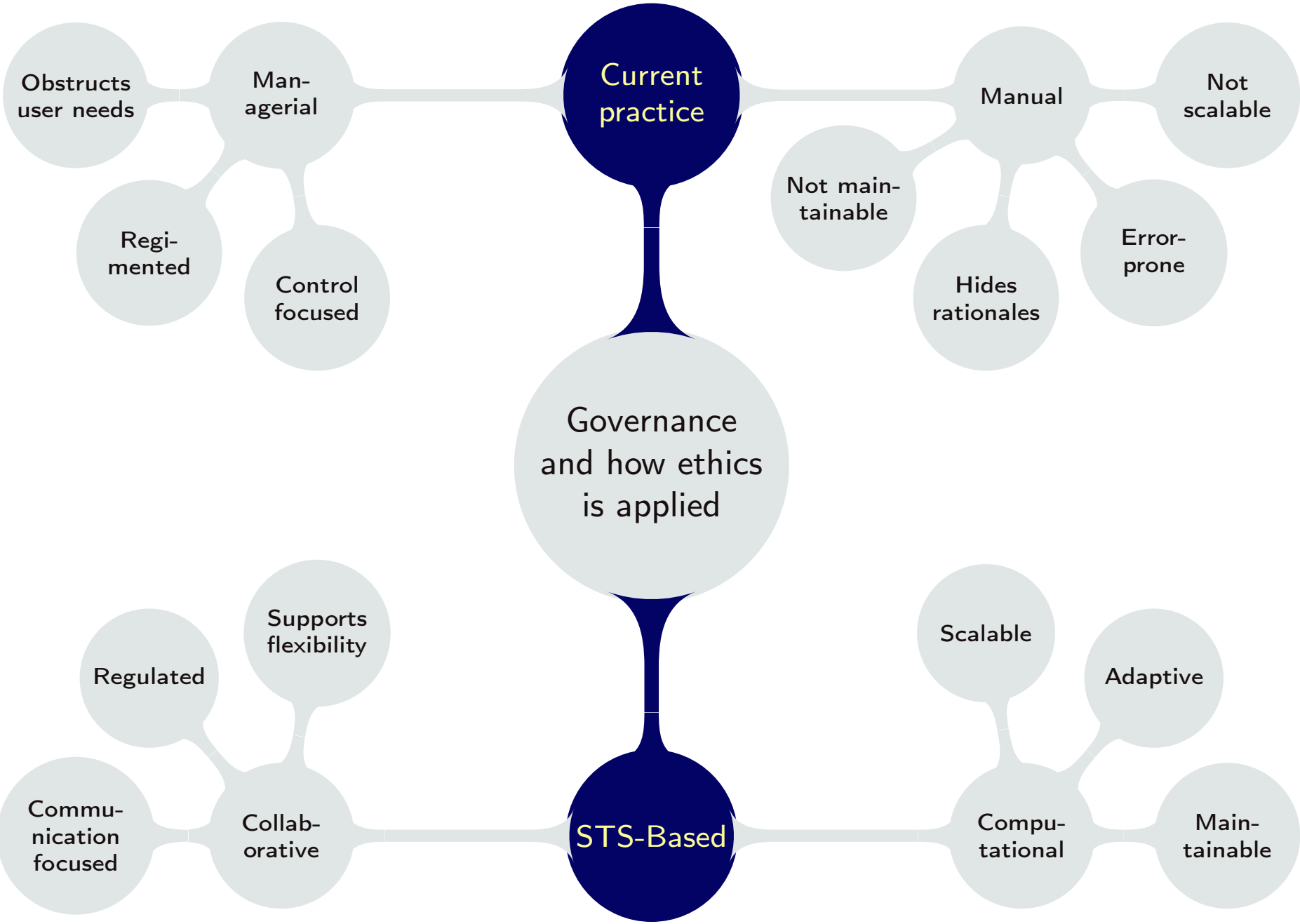


Sociotechnical Systems (STS): A Computational Norm-Based System

Context of interaction in which principals are represented by agents

- Principal: human or organization, a stakeholder who acts
- Norm: *directed* social expectation between principals
  - Types: Commitment, prohibition, authorization, power, ...
  - Standards of correctness
    - *Prima facie*, satisfaction is ethically desirable and violation undesirable
- Accountability: the power of a principal to call another to account for its actions
  - Derives from norms
  - Provides an opportunity for principals to explain their actions
    - Leading to *prima facie* judgments being reconsidered
  - Is not traceability, which is merely a supporting mechanism
  - Is not blame and sanction, which are subsequent


Accountability as a mirror of autonomy
Traditional computing
Traceability
Fails when traceability fails
Insufficient
Unnecessary
Deterrence
Penalty voids accountability?
Treats accountability as sanctioning
STS-Based
Intuitive
Independent of mechanism
Captures domain requirements
Operationalizable
Sanctioning
Logging
Monitoring


Current practice
Governance and how ethics is applied
STS-Based
Obstructs user needs
Man-agerial
Regi-mented
Control focused
Manual
Not scalable
Not main-tainable
Error-prone
Hides rationales
Regulated
Supports flexibility
Commu-nication focused
Collab-orative
Scalable
Adaptive
Compu-tational
Main-tainable

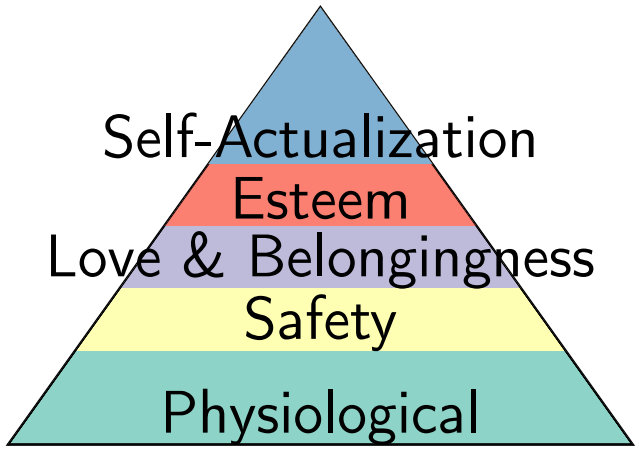
Outline and Schedule

Values: Motivations and Goals

When we think of values, we think of what is important to us in life [Schwartz, 2012]

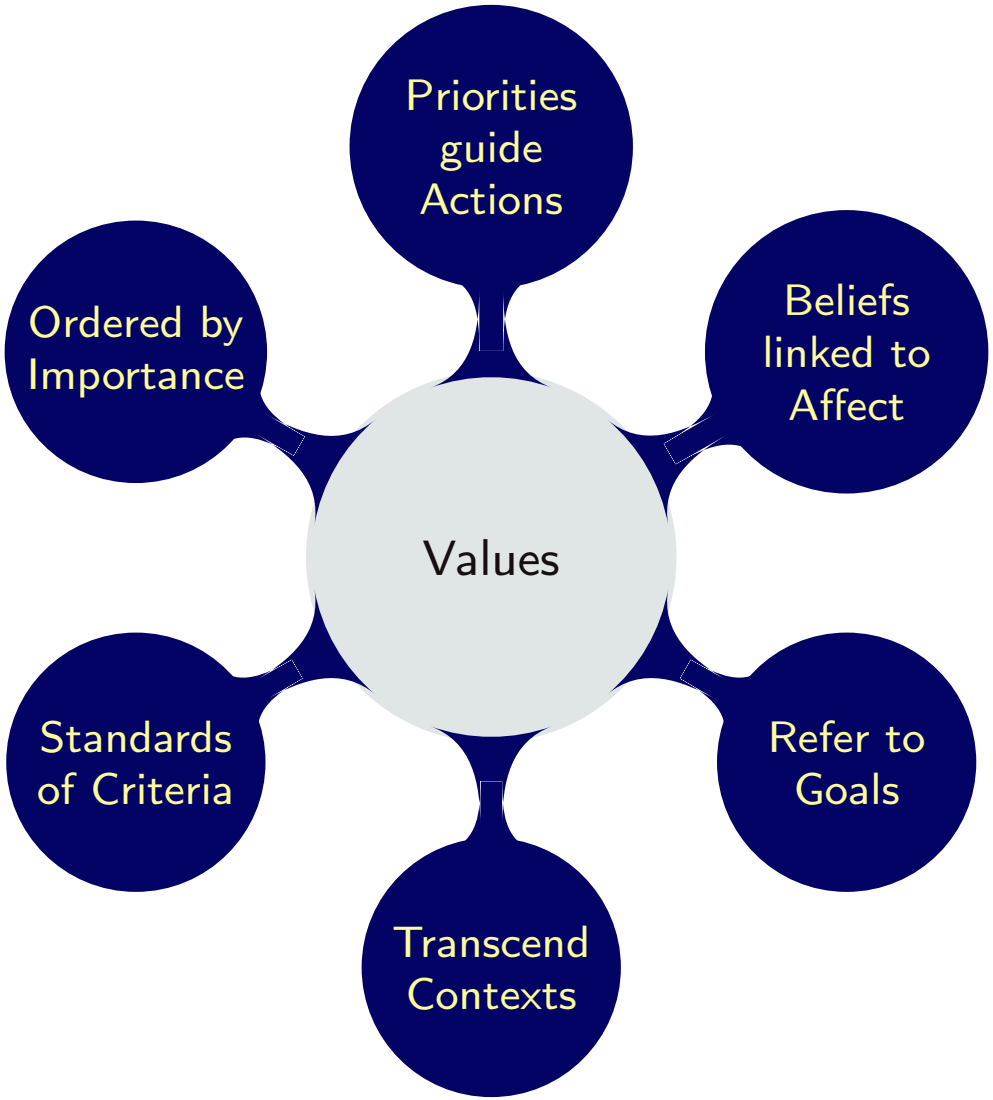
Basic values are likely to be universal because they are grounded in one or more universal requirements of human existence

1. Needs of individuals as biological organisms
2. Requisites of coordinated social interaction
3. Survival and welfare needs of groups



- People articulate appropriate goals to cope with these requirements, communicate them to others, and gain cooperation in their pursuit
- Values are constructs used to represent such goals mentally and vocabulary used to express them in social interaction

The Nature and Features of All Values [Schwartz, 2012]



Preferences: Values vs. Interests

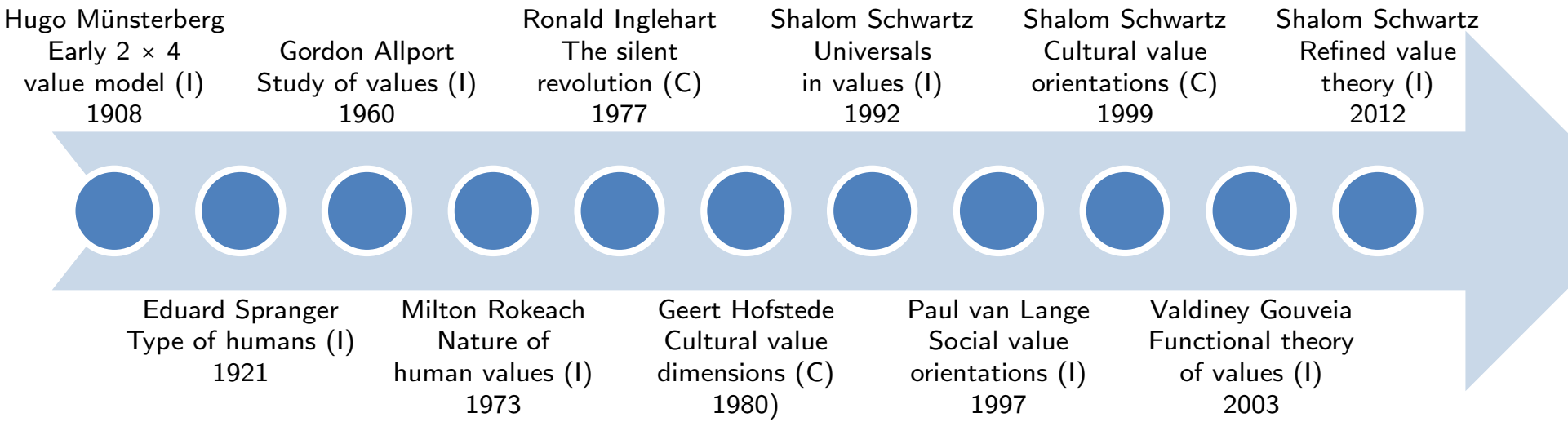
A preference means a more positive attitude (*leaning*) toward one alternative over another (or other) alternative(s) [Dawis, 1991]

- Preferences can be over values, interests, or other arbitrary choices

Values vs. Interests

- An interest is manifested as sustained attention involving cognition of the interest object, accompanying positive affect
- A value is manifested as affective valuation
- Both values and interests influence behavior
- When judgment in preference is based on liking (i.e., attraction), it is an interest; when the basis is importance (i.e., significance or meaning), the preference is a value

A Timeline of (Selected) Value Models [Hanel et al., 2018]



Individual value model (I): Describe and measure the values of an individual

Cultural value model (C): Describe and measure the values of a culture

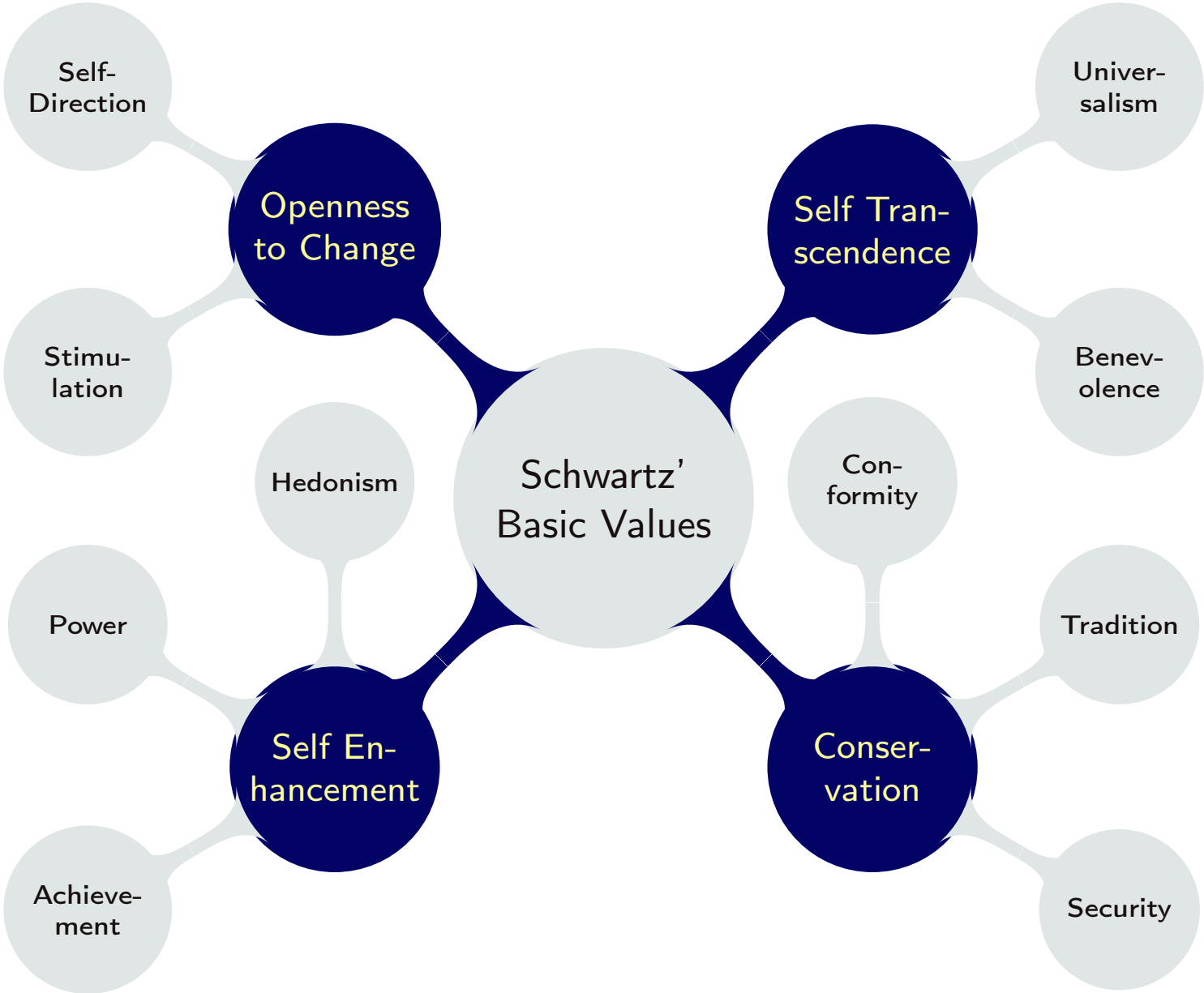
Rokeach Value System [Rokeach, 1973]

18 Terminal Values: Ends (Desirable end-states of existence)

- True Friendship, Mature Love, Self-Respect, Happiness, Inner Harmony, Equality, Freedom, Pleasure, Social Recognition, Wisdom, Salvation, Family Security, National Security, A Sense of Accomplishment, A World of Beauty, A World at Peace, A Comfortable Life, An Exciting Life

18 Instrumental Values: Means (Preferable modes of behavior)

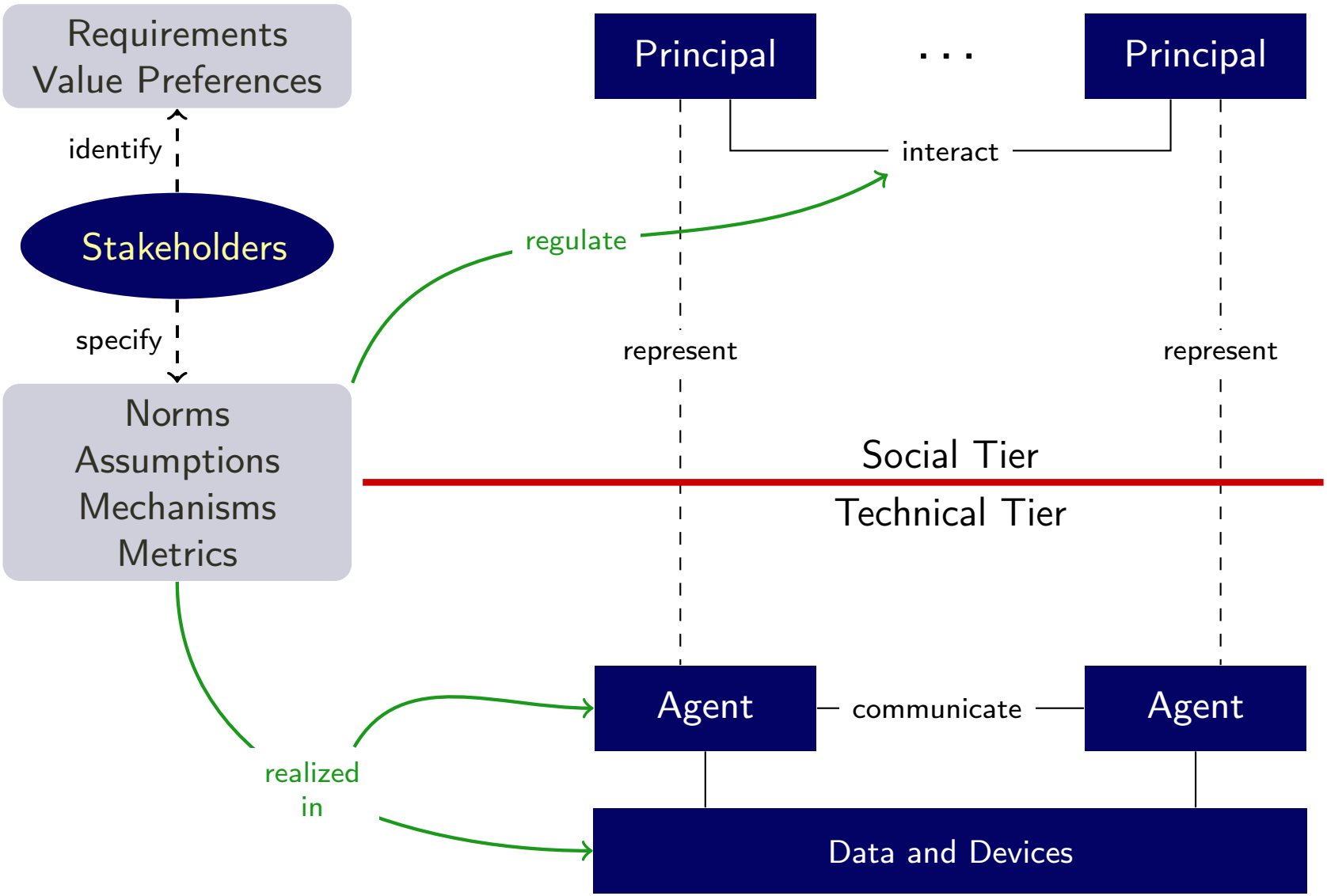
- Cheerfulness, Ambition, Love, Cleanliness, Self-Control, Capability, Courage, Politeness, Honesty, Imagination, Independence, Intellect, Broad-Mindedness, Logic, Obedience, Helpfulness, Responsibility, Forgiveness


Schwartz' Basic Values
Openness to Change
Self-Direction
Stimu-lation
Hedonism
Self Tran-scendence
Univer-salism
Benev-olence
Con-formity
Self En-hancement
Power
Achieve-ment
Conser-vation
Tradition
Security

Value	Motivational Goals
Self-Direction	Independent thought and action, self-respect, privacy
Stimulation	Excitement, novelty, and challenges in life
Benevolence	Welfare of those in frequent contact
Universalism	Welfare of all people and nature
Security	Safety, harmony, and stability of self and others
Conformity	Restraint of actions that violate social norms
Tradition	Conforming to cultural and religious customs and ideas
Achievement	Personal success, competence
Power	Social status, control over people and resources
Hedonism	Pleasure or sensuous gratification for oneself

Outline and Schedule

Sociotechnical Systems



Ethics in the Large: Accountability and Adaptivity

An ethical STS presupposes good governance

An adaptive methodology undertaken by stakeholders of an STS

- Identify each stakeholder's value preferences
- Specify the norms that support those value preferences
  - Norms are operational refinements of value preferences
  - Norms make accountability concrete
- A stakeholder's SIPA
  - Adopts one or more roles
  - Carries out its part of an enactment
  - Evaluates outcomes on its (primary and secondary) stakeholders
    - Whether values are promoted in alignment with the preferences
    - Which norms are satisfied
- Iterate

Social Norms

Norms govern the interactions between principals

Formally, a norm is a tuple $\langle n, sbj, obj, ant, con \rangle$, where

- n, its type, is one of {c, p, a};
- $sbj \in \mathbb{R}$ is its subject;
- $obj \in \mathbb{R}$ is its object;
- $ant \in Expr$ is its antecedent; and
- $con \in Expr$ is its consequent.

We write a norm as n(sbj, obj, ant, con)

Types of Social Norms

Examples of a commitment, a prohibition, and an authorization

- Commitment: *A meeting attendee is committed to other attendees that he or she will keep his or her phone on silent during the meeting*
- Prohibition: *A library visitor is prohibited by the library to answer any phone calls when the visitor is in the silent reading area of the library*
- Authorization: *A library staff member is authorized by the library to make any personal phone calls during lunch hours*

Requirements of a Healthcare STS

Healthcare emergency scenario. Trade-off between values of privacy and safety

- R-Publish: Patient's personally identifying information (PHI) should not be published online under any circumstances
- R-External: Except in emergencies, hospital physicians should not share a patient's PHI with outside physicians
- R-Family: In emergencies, hospital physicians may share patient's PHI with family members to inform family members or gather new information to help with treatment

Normative Specification of an STS

Healthcare emergency scenario
Initial specification, to be refined

- R-Publish: Patient's personally identifying information (PHI) should not be published online under any circumstances
  prohibition(physician, hospital, true, publish_PHI_online)
- R-External: Except in emergencies, hospital physicians should not share a patient's PHI with outside physicians
  prohibition(physician, hospital, true, share_PHI_outside_phy)
- R-Family: In emergencies, hospital physicians may share patient's PHI with family members to inform family members or gather new information to help with treatment
  authorization(physician, hospital, true, share_PHI_family)

Refining a Specification of an STS

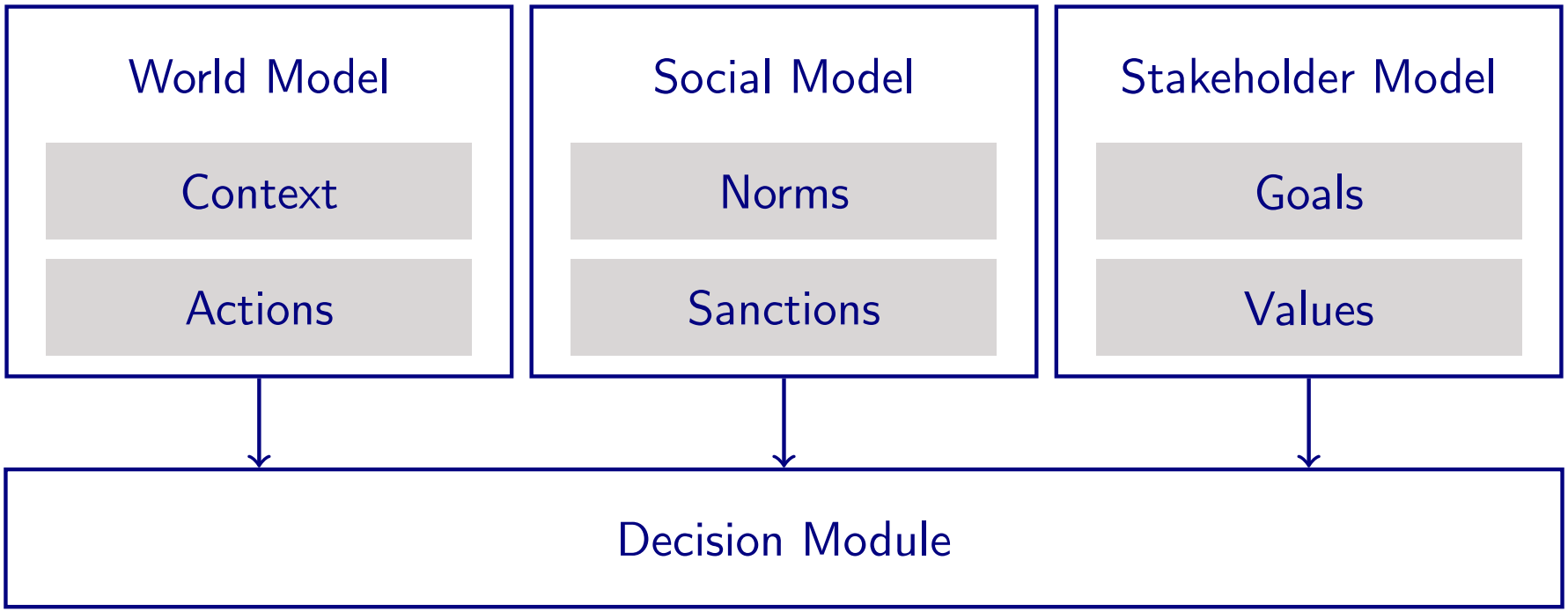
Healthcare emergency scenario

- R-Publish: Don't publish PHI
  prohibition(physician, hospital, true, publish_PHI_online)
  - Further refinement: Include a mechanism to not allow publishing PHI online
- R-External: Except in emergencies, don't share PHI with outside physicians
  ~~prohibition(physician, hospital, true, share_PHI_outside_phy)~~
  prohibition(physician, hospital, ¬emergency, share_PHI_outside_phy)
- R-Family: Share PHI in emergencies
  ~~authorization(physician, hospital, true, share_PHI_family)~~
  authorization(physician, hospital, emergency, share_PHI_family)
  - Further refinement: Include a commitment from physician to family commitment(physician, family, emergency, share_PHI_family)

A SIPA: Schematically

What must a SIPA represent and reason about to participate ethically in a multiagent system?

A SIPA's decision making takes into account its stakeholders, primary and secondary



Outline and Schedule

Value Sensitive Design [Friedman et al., 2017]

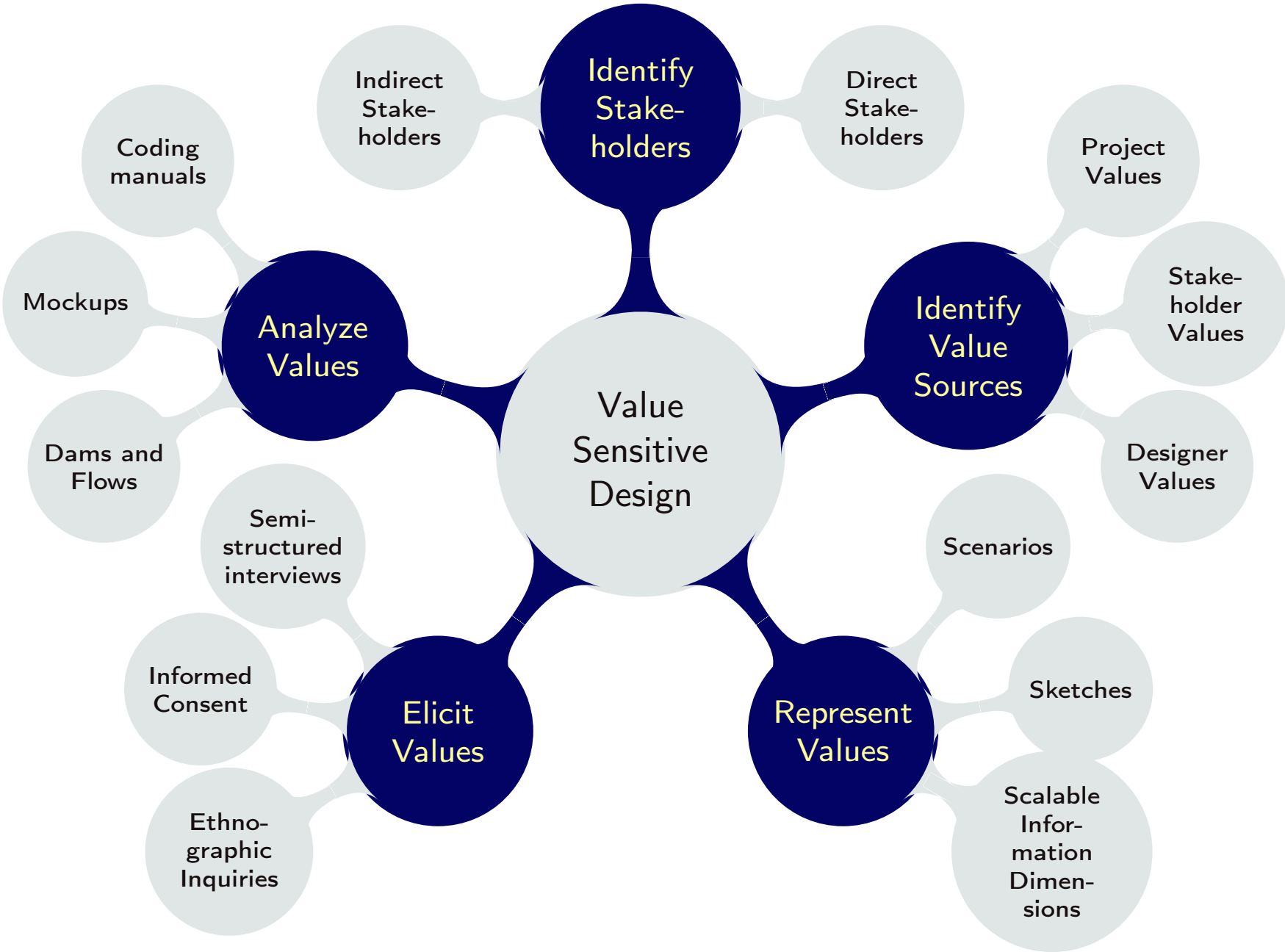
A theoretically grounded approach to the design of technology that accounts for human values in a principled and systematic manner

- How can we explore the sociotechnical design space from the perspective of values?
- How can we identify stakeholders, and their values?
- How can we resolve value tensions among stakeholders?
- How can we translate stakeholders' values into technical design?

Not one method ...

... but a class of methods faithful to value sensitive design principles

- Intended to be integrated with other methods and processes for technical design and development
- Starting points: A value, technology, policy, or context of use

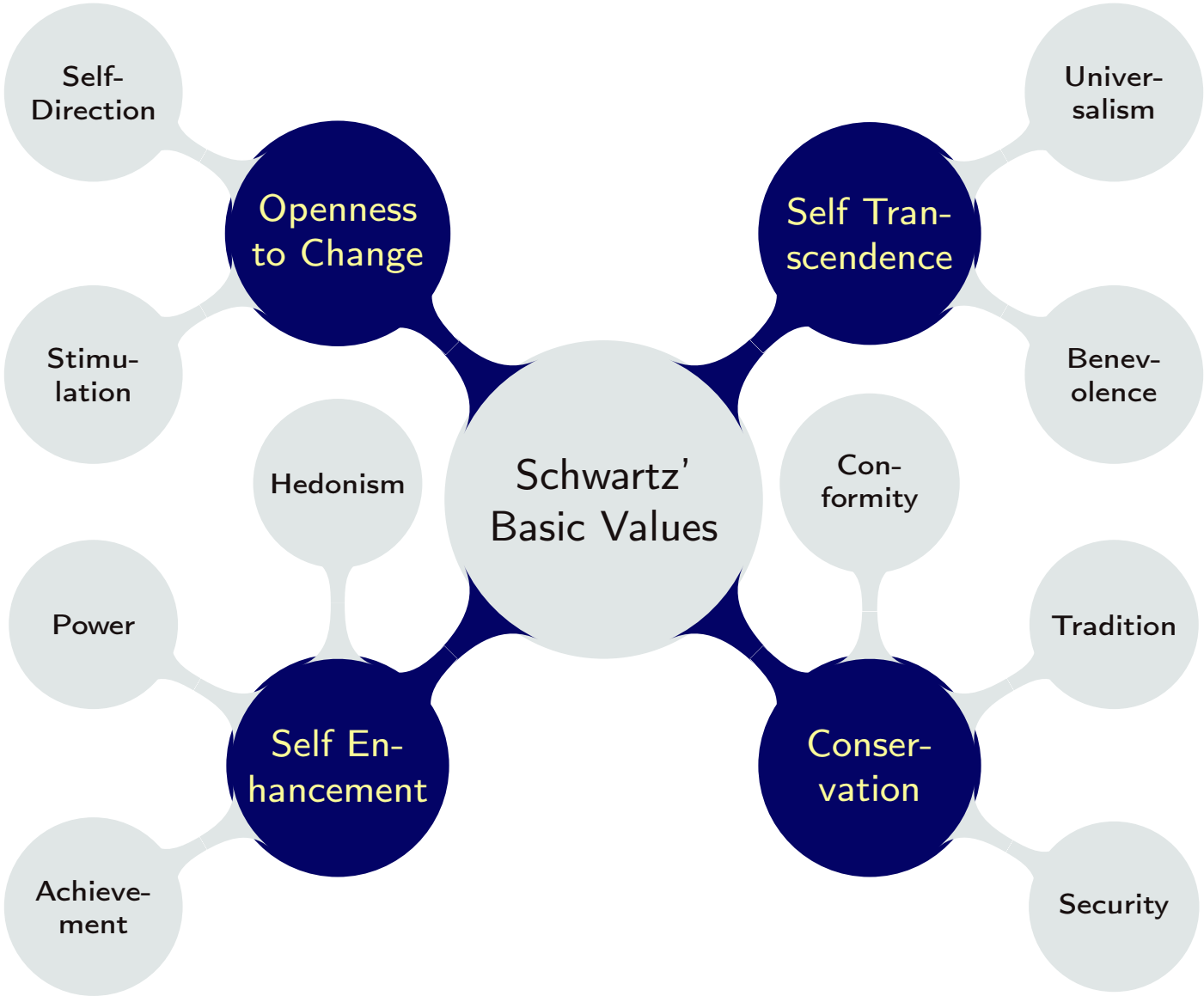

Value Sensitive Design
Identify Stake-holders
Indirect Stake-holders
Direct Stake-holders
Identify Value Sources
Project Values
Stake-holder Values
Designer Values
Represent Values
Scenarios
Sketches
Scalable Infor-mation Dimen-sions
Elicit Values
Semi-structured interviews
Informed Consent
Ethno-graphic Inquiries
Analyze Values
Coding manuals
Mockups
Dams and Flows

Value Sensitive Design (VSD) of Multiagent Systems

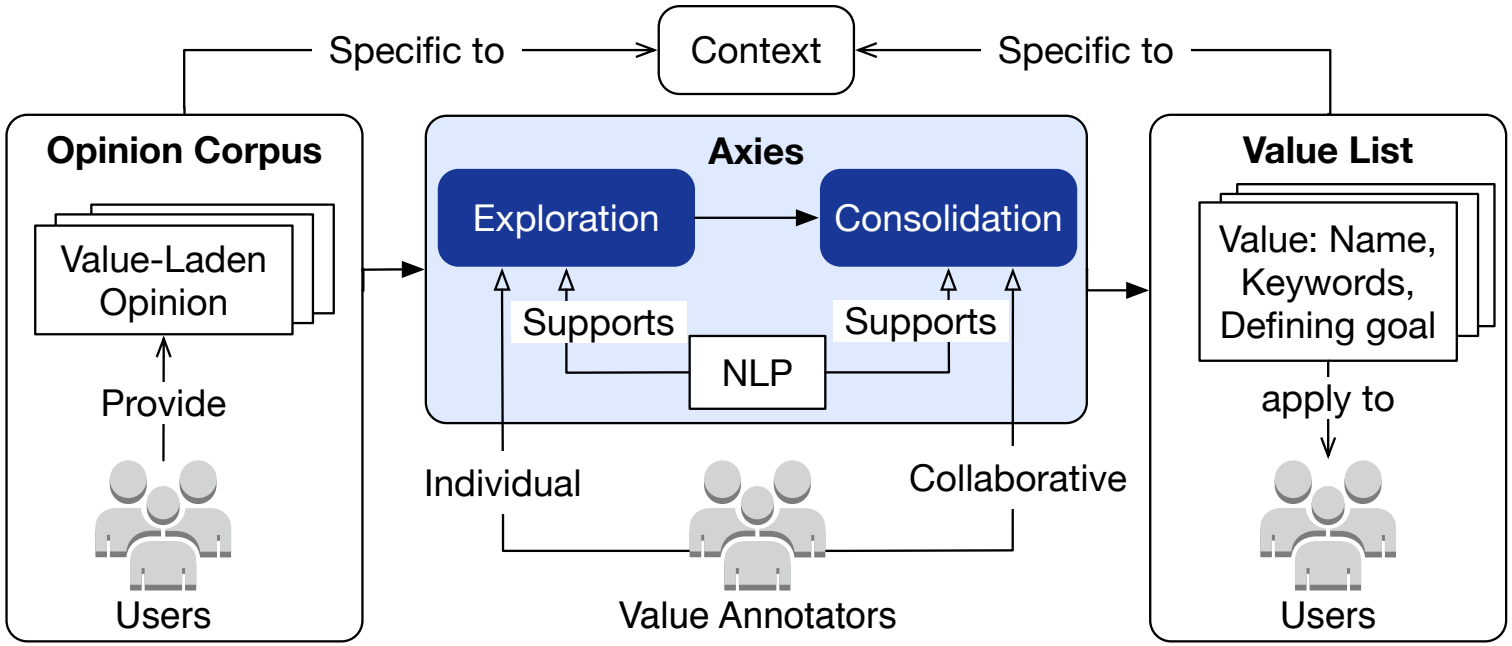
Integrating VSD and Agent-Oriented Software Engineering (AOSE)

Axies [Liscio et al., 2021, 2022]	Identify contextually relevant values
Xipho [Murukannaiah and Singh, 2014]	Incorporate values in agent-oriented models
Arnor; Valar [Ajmeri et al., 2017, 2018a]	Relate and reason about values and norms
XSIGA; Poros [Agrawal et al., 2022; Ajmeri et al., 2018b]	Communicate values
Elessar [Ajmeri et al., 2020]	Reason about value value tensions and conflicts

Identifying Values of Interest to an Agent or an MAS



Identifying Values of Interest to an Agent or an MAS



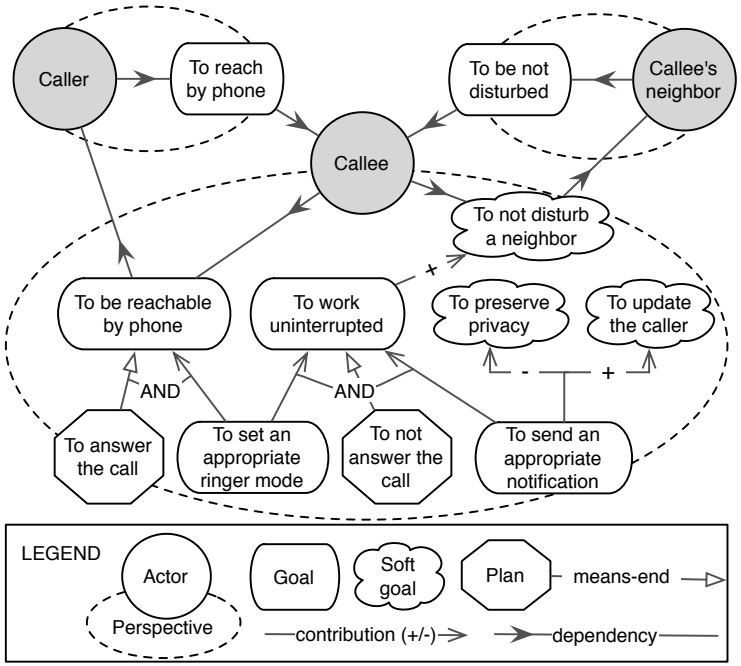
Incorporating Values in an Agent Model

AOSE provides high-level technical abstractions to represent values

Example abstrations from Tropos

- **Actor**: A social, physical, or software agent
- **Goal**: A strategic interest of an actor
- **Plan**: An abstraction of action
- **Belief**: An actor's representation of the world
- **Dependency**: A relationship between actors

A Tropos model of an Intelligent Ringer



Understanding Values in Context

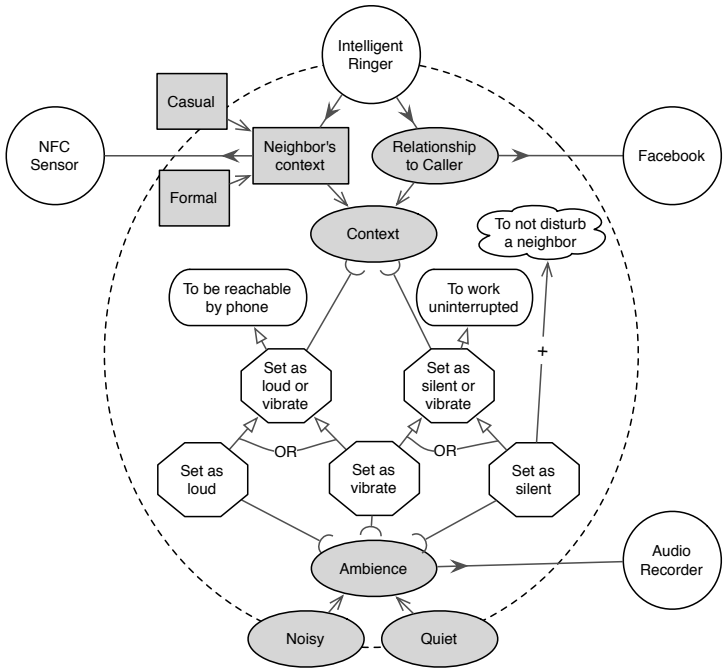
Xipho provides systematic steps to contextualize agent capabilities

- To be reachable:
  Welfare of others $\Uparrow$
- To work uninterrupted:
  Ambition $\Uparrow$
- Welfare of others $\succ$ Ambition?

Xipho can yield a specification of value preferences grounded in contexts, e.g.,

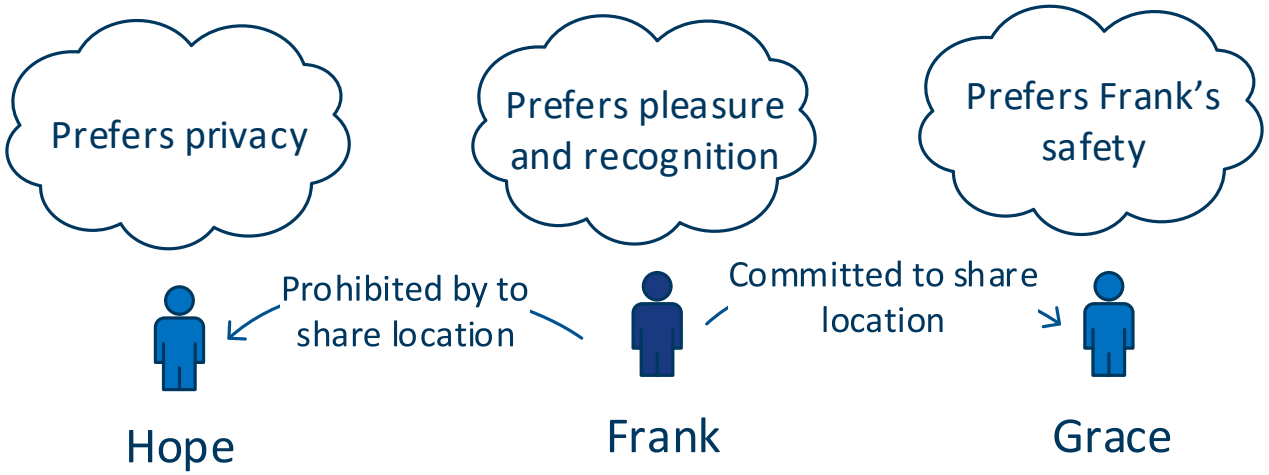
$$\textit{Relationship} = ?R_1 \wedge$$
$$\textit{Neighbor's context} = ?N_1 \rightarrow$$
$$\text{Welfare of others} \succ \text{Ambition}$$

A contextual model of Intelligent Ringer



Reasoning about Values to Revise Norms

Arnor & Valar model social expectations (norms), considering values



Frank's dilemma: Which sharing policy to select?

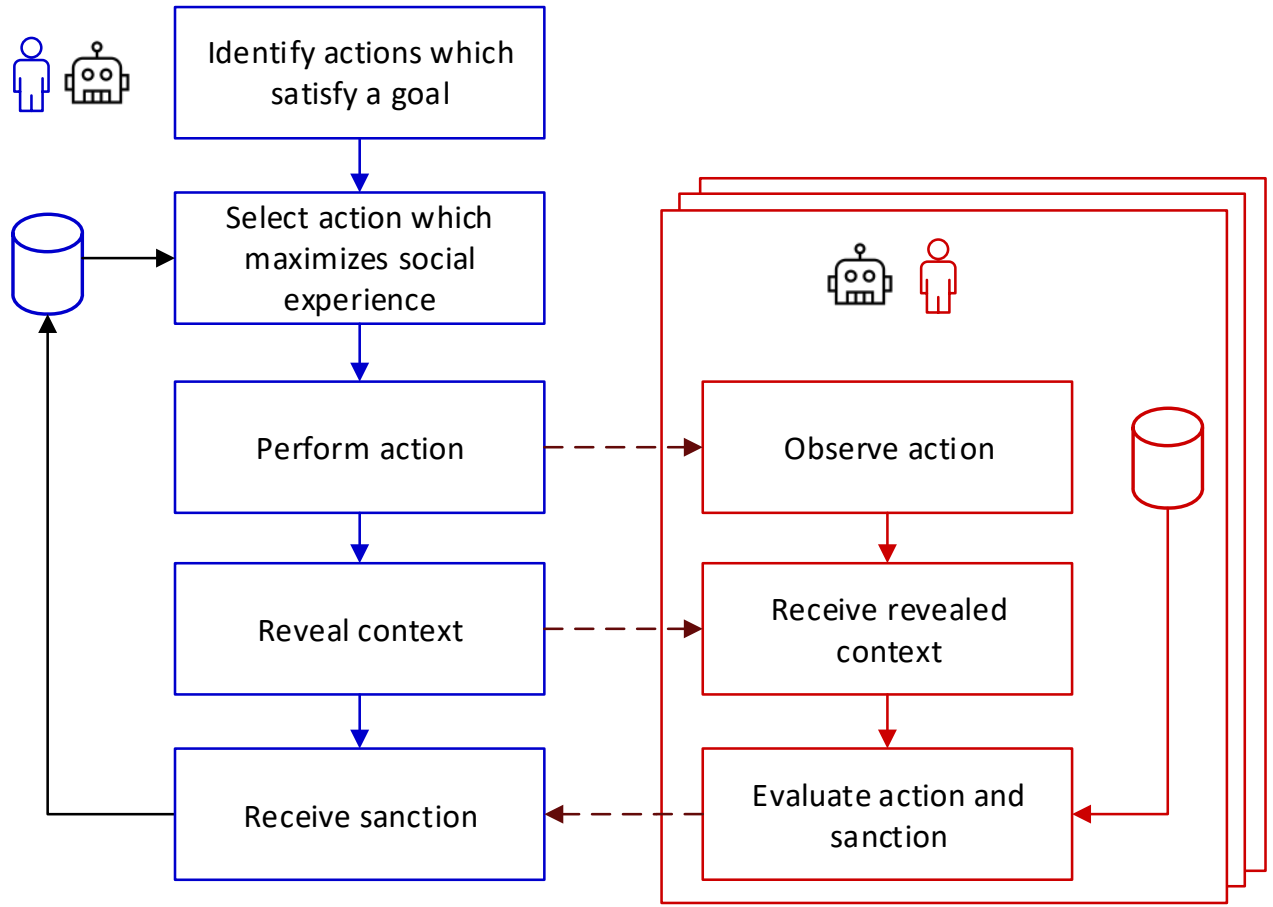
Share with all: Pleasure for Frank ⇑

Share only with Grace: Safety for Grace ⇑

Share with no one: Privacy for Hope ⇑

Communicating Values by Revealing Contexts

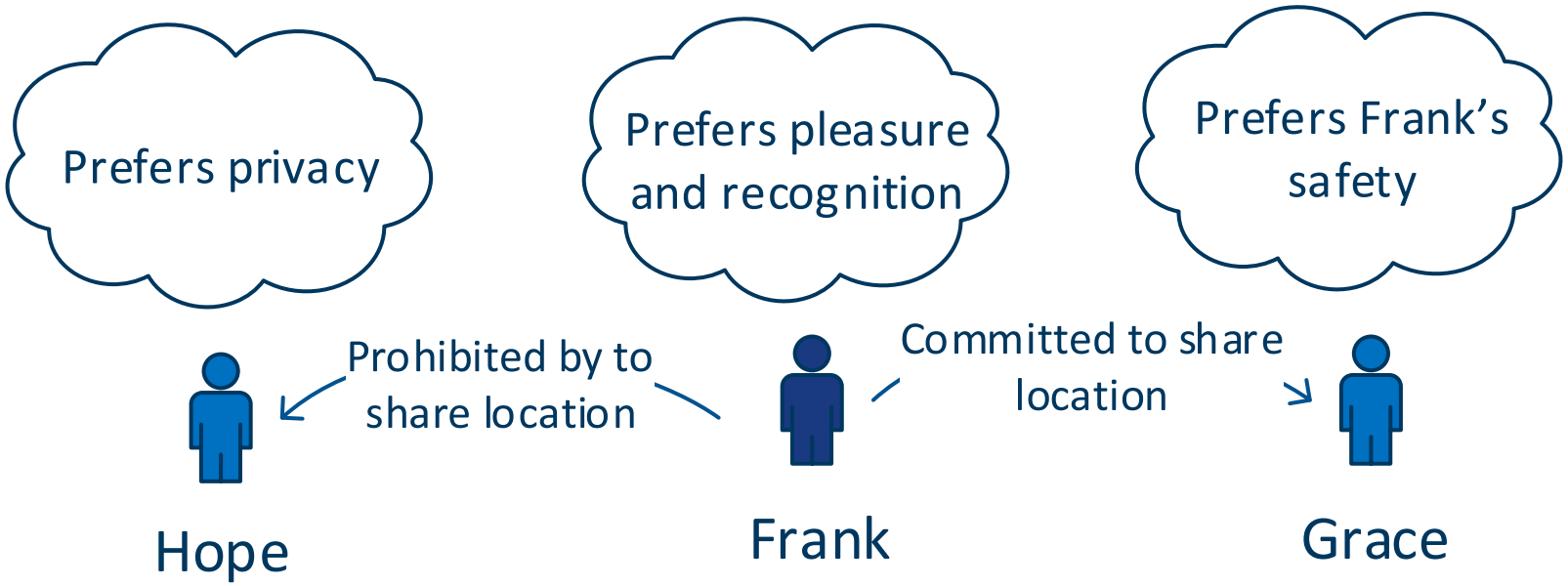
Poros helps agents communicate values by revealing context



Outline and Schedule

Example of Context Sharing Setting

Frank: committed to his mother Grace to share his location; visits aunt Hope in NYC



Frank's dilemma: Which sharing policy to select?

Share with all: Pleasure for Frank $\Uparrow$

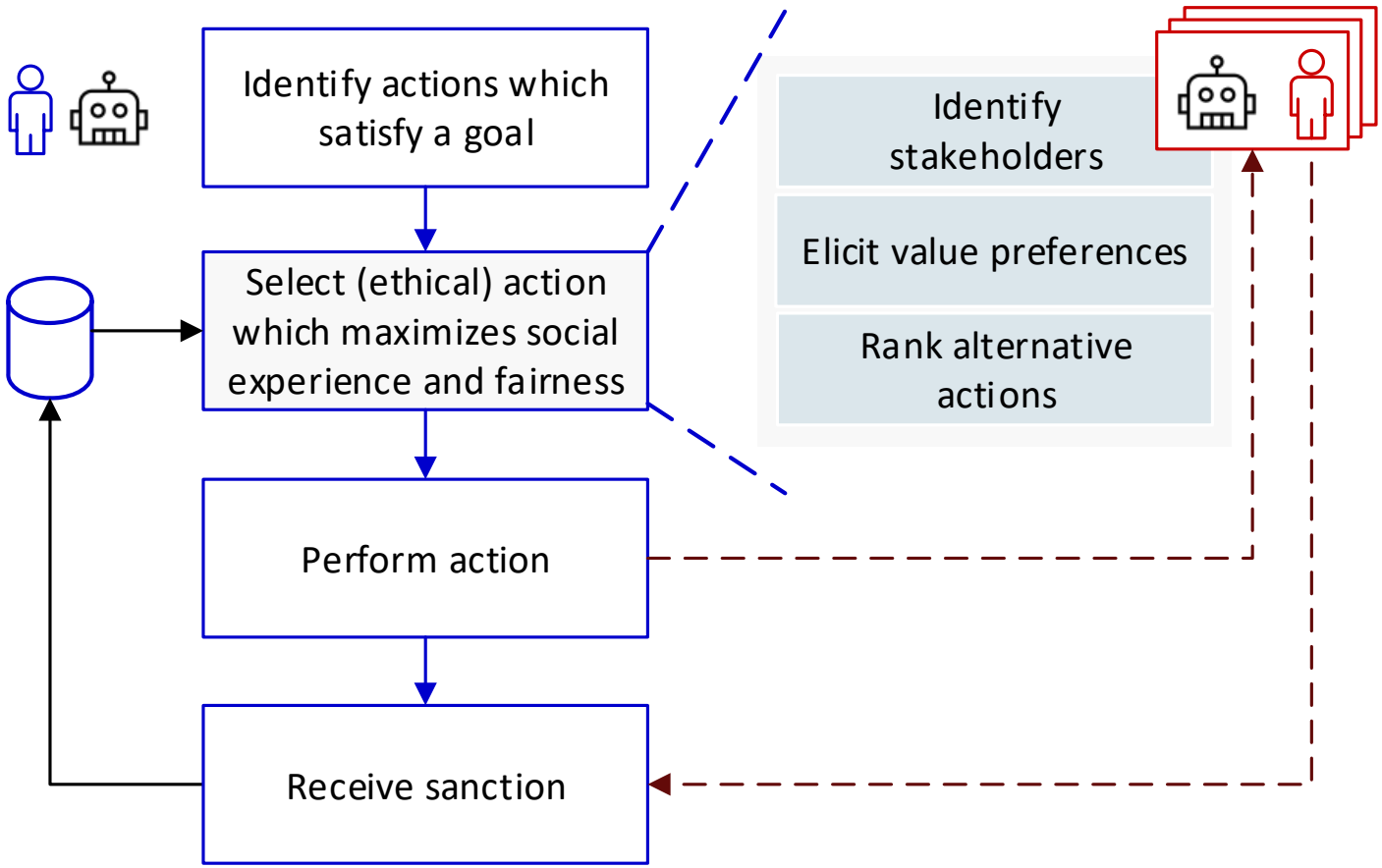
Share only with Grace: Safety for Grace $\Uparrow$

Share with no one: Privacy for Hope $\Uparrow$

Reasoning about Stakeholders' Value Preferences

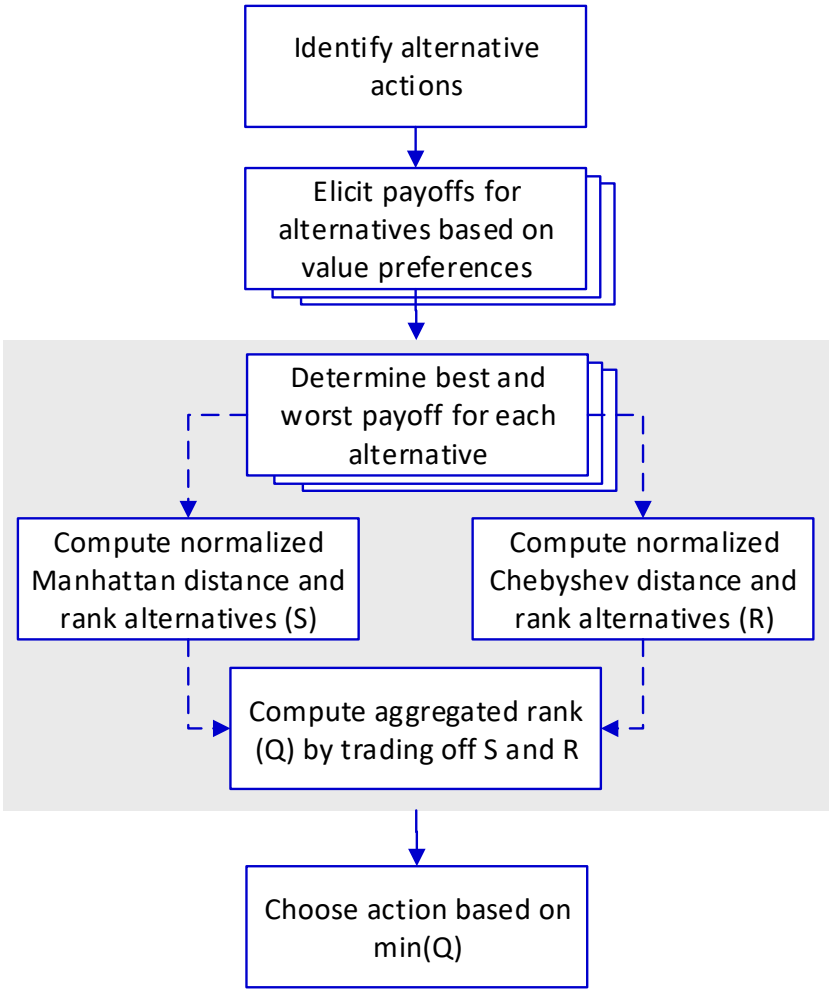
How can SIPAs aggregate value preferences of their stakeholders to select an ethical action?

A SIPA's secondary stakeholders can change with the context



Choosing an Ethical Action

SIPAs adapt a multicriteria decision making method (VIKOR) to select ethically appropriate action—balancing *social welfare* and *egalitarianism*



Choosing an Ethical Action

Selecting appropriate context-sharing policy using VIKOR

Alternatives	Frank's Values				Hope's Values				S_a	R_a	Q_a
	Pleasure	Privacy	Recognition	Security	Pleasure	Privacy	Recognition	Security			
a_1 All	1.0	0.5	1.0	0.5	0.5	0.0	0.5	0.5	**2.50**	2.00	0.50
a_2 Grace	0.5	0.5	0.5	1.0	0.5	0.5	0.5	0.5	3.00	**1.00**	**0.10**
a_3 No one	0.0	0.5	0.0	0.0	0.5	1.0	0.5	0.5	5.00	2.00	1.0
w_v: Value preferences	2	1	2	1	1	2	1	1			
f_v^*	1.0	0.5	1.0	1.0	0.5	1.5	0.5	0.5			
f_v^-	0.0	0.5	0.0	0.0	0.5	0.0	0.5	0.5			

VIKOR calculations for context sharing example:
`https://go.ncsu.edu/vikor-context-sharing-example`

Restaurant Example: Where should Jess, Dan, and Alex Go?

Contrasting various ethical principles

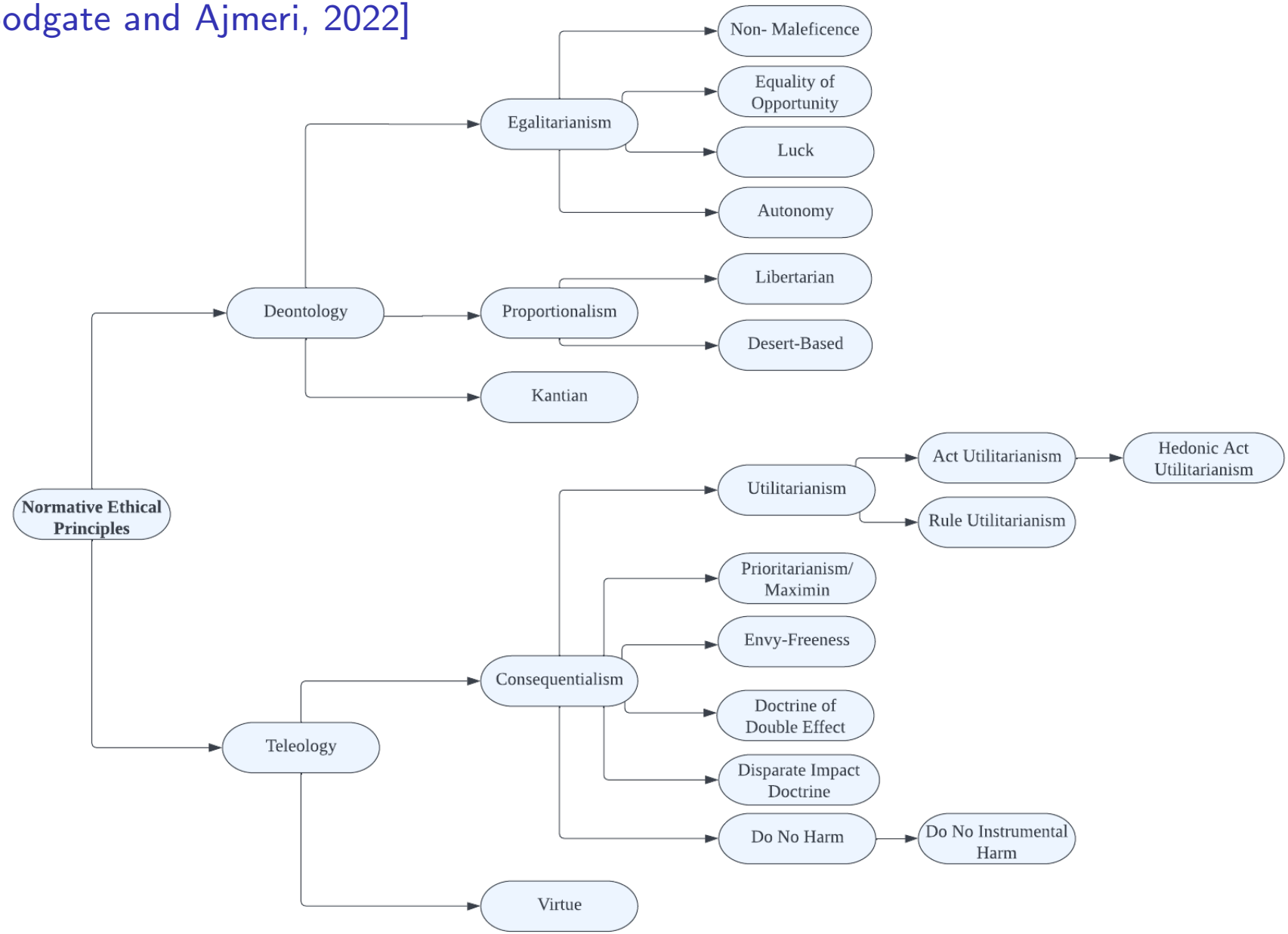
	Jess	Dan	Alex
Pancake restaurant	10	10	2
Pasta restaurant	7	7	7
Pizza restaurant	5	5	10

Aggregate Happiness

- Pancake: 22
- Pasta: 21
- Pizza: 20

Taxonomy of Ethical Principles

[Woodgate and Ajmeri, 2022]

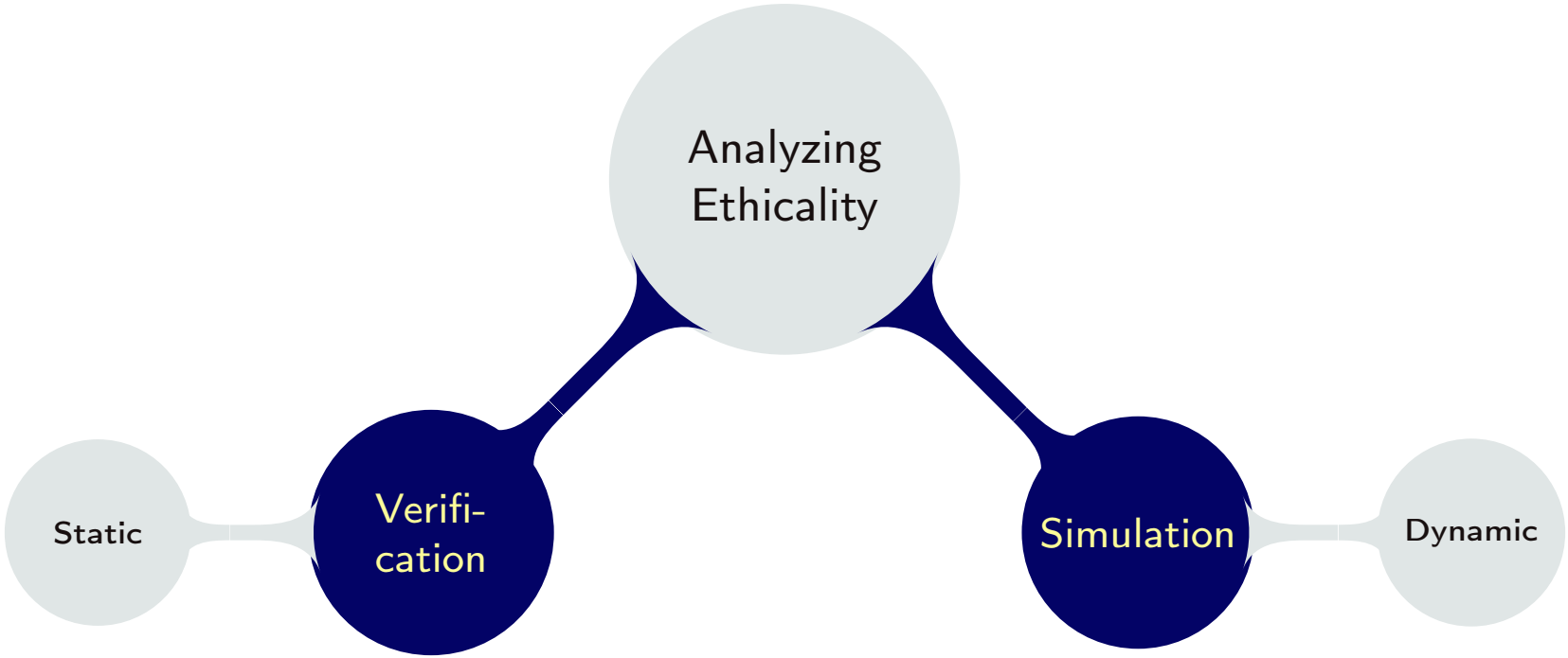


Outline and Schedule

Analyzing Ethicality

How do we analyze if an STS specification satisfies the stakeholders' requirements with respect to their value preferences and ethical criteria such as social welfare and egalitarianism?

- Liveness: something good happens
- Safety: nothing bad happens
- Robustness: how long something good keeps happening
- Resilience: how soon we recover from something bad



Verification and Simulation: Challenges and Opportunities

Verification

How can we verify an STS specification for ethicality?

Simulation

How can we enable an STS stakeholder to assess runtime outcomes of an STS specification?

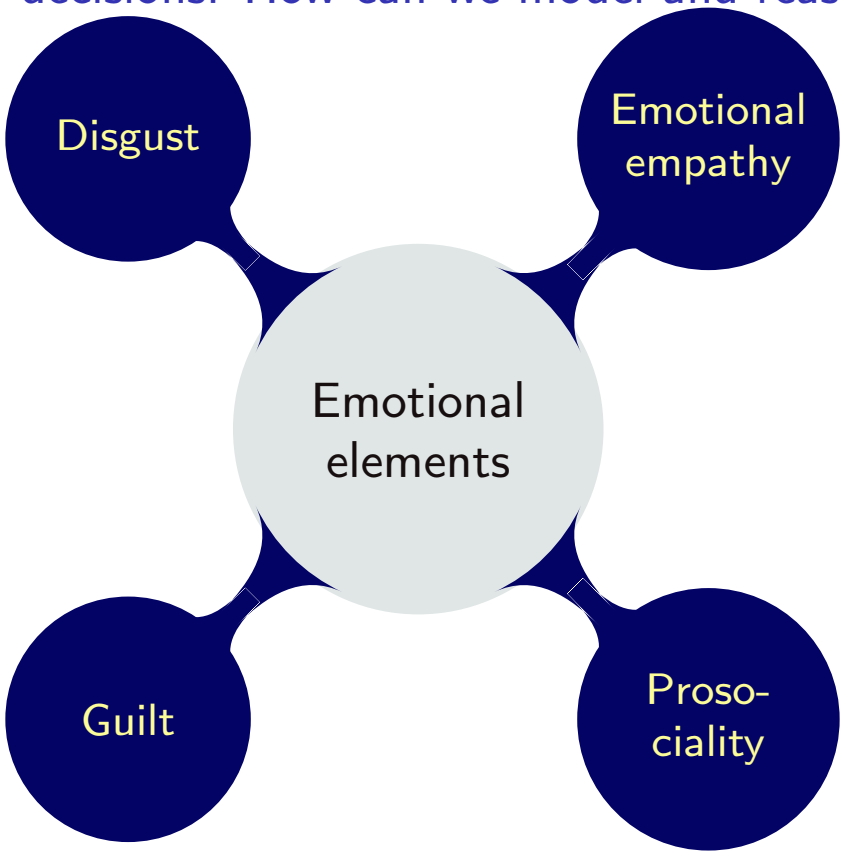
Opportunities

- Develop new model checking approaches that consider value preferences of stakeholders
- Enable stakeholders to guide simulations at runtime and help understand the simulation outcomes

Outline and Schedule

Incorporating Understanding of Emotions

Emotions influence social decisions. How can we model and reason about emotions?



Opportunities

- New techniques to model emotional elements
- Enable SIPAs to understand these emotional elements

Outline and Schedule

Value Elicitation: Instruments for Measuring Value Priorities

Rokeach Value Survey

Arrange the 18 terminal values, followed by the 18 instrumental values, into an order "of importance to YOU, as guiding principles in YOUR life."

Schwartz Value Survey

- 57 items about potentially desirable end states or ways of acting
- Rate the importance of each item "as a guiding principle in MY life"
- Nine point asymmetric rating scale

Portrait Values Questionnaire

- Short (gender-matched) verbal portraits of 40 different people
- Question: How much like you is this person?
- Six-point rating scale (*very much like me* to *not at all like me*)

Value Elicitation: Challenges and Opportunities

Learning

How can an agent elicit user and context specific value preferences unintrusively?

Negotiation

How can we enable stakeholders to create an STS specification that accords with their value preferences?

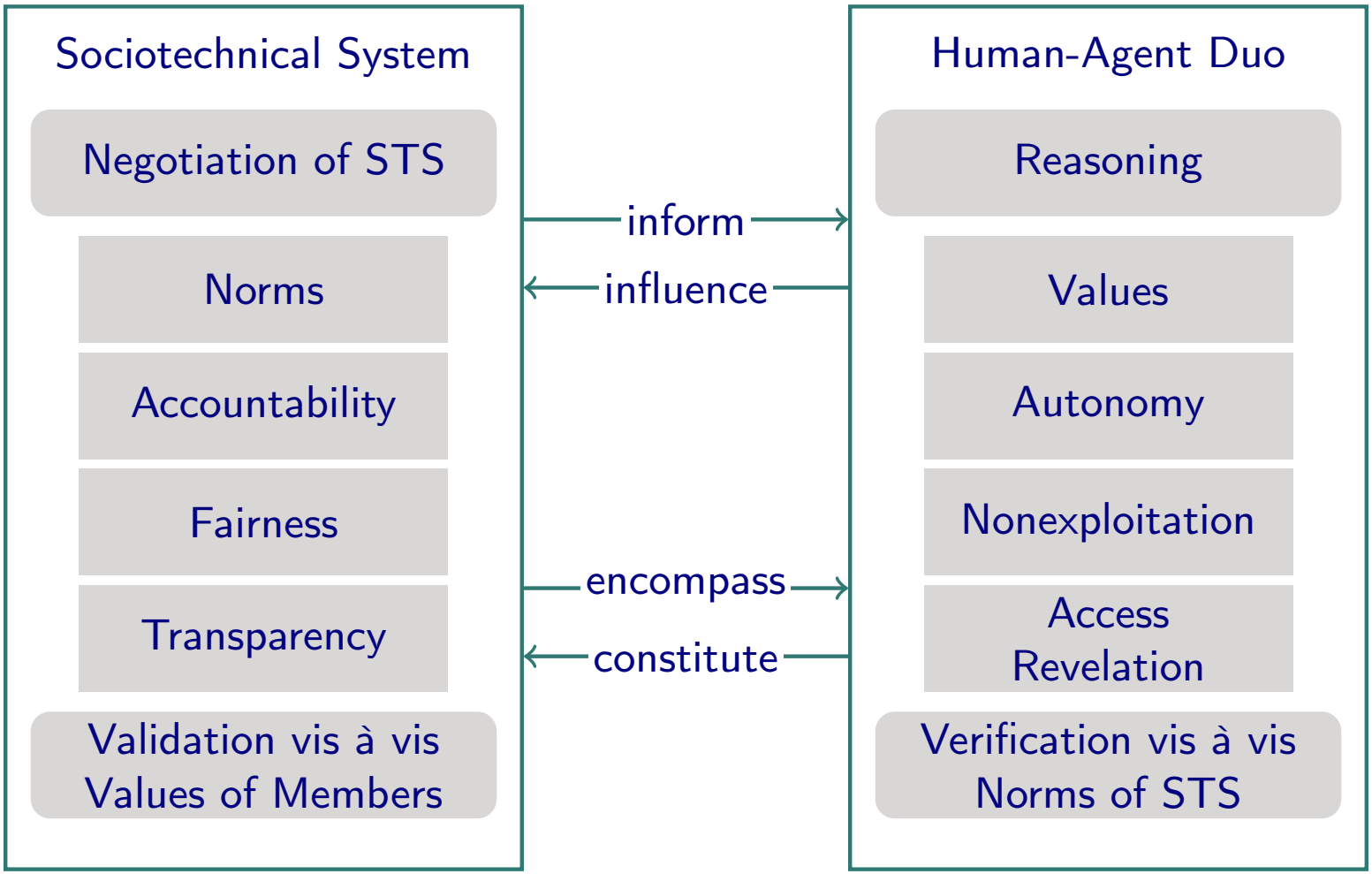
Opportunities

- Learn value preferences by observing the principals' actions as well as the (positive or negative) sanctions they receive
- Support stakeholders with conflicting requirements but similar value preferences in generating an acceptable STS specification

Outline and Schedule

Alignment of Systems and Human-Agent Duos

Duo: A user and an agent working together

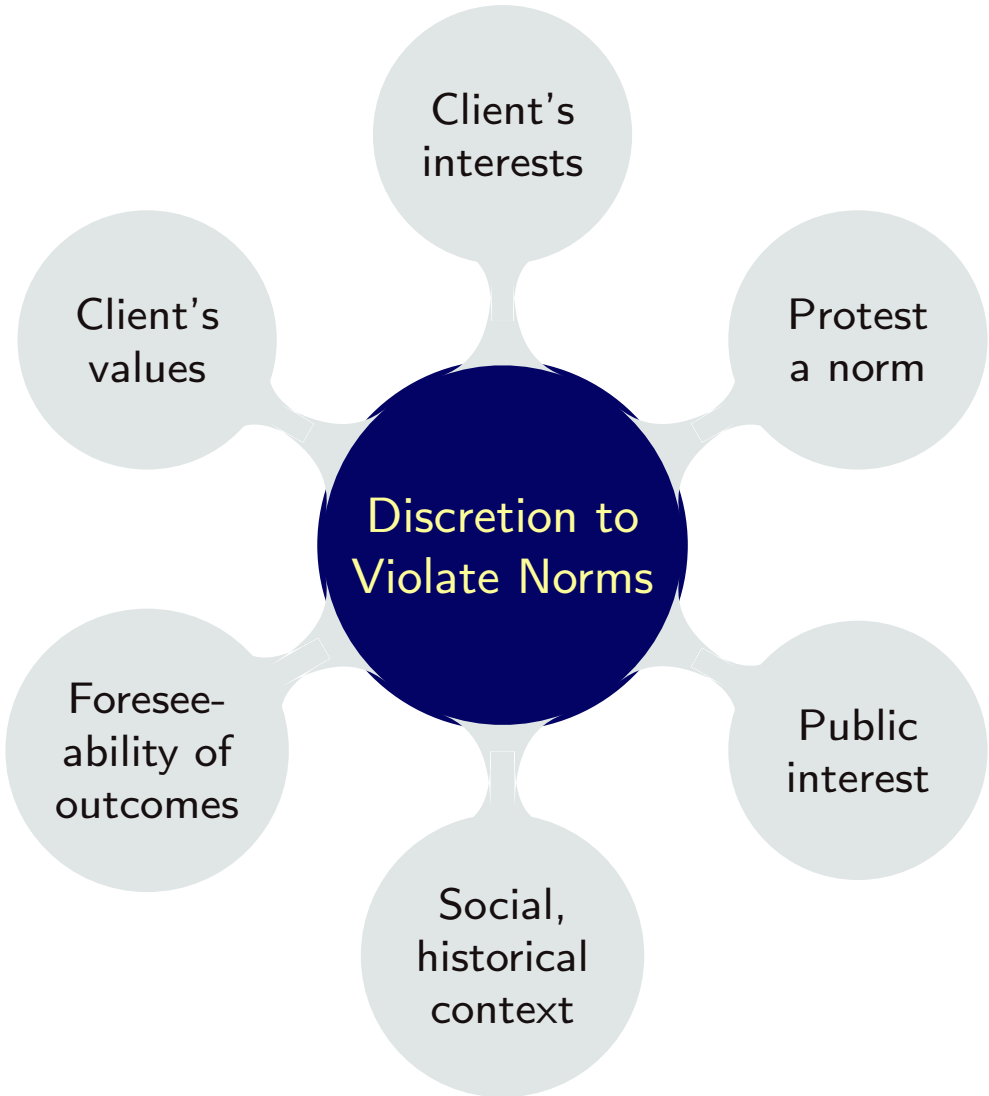


Outline and Schedule

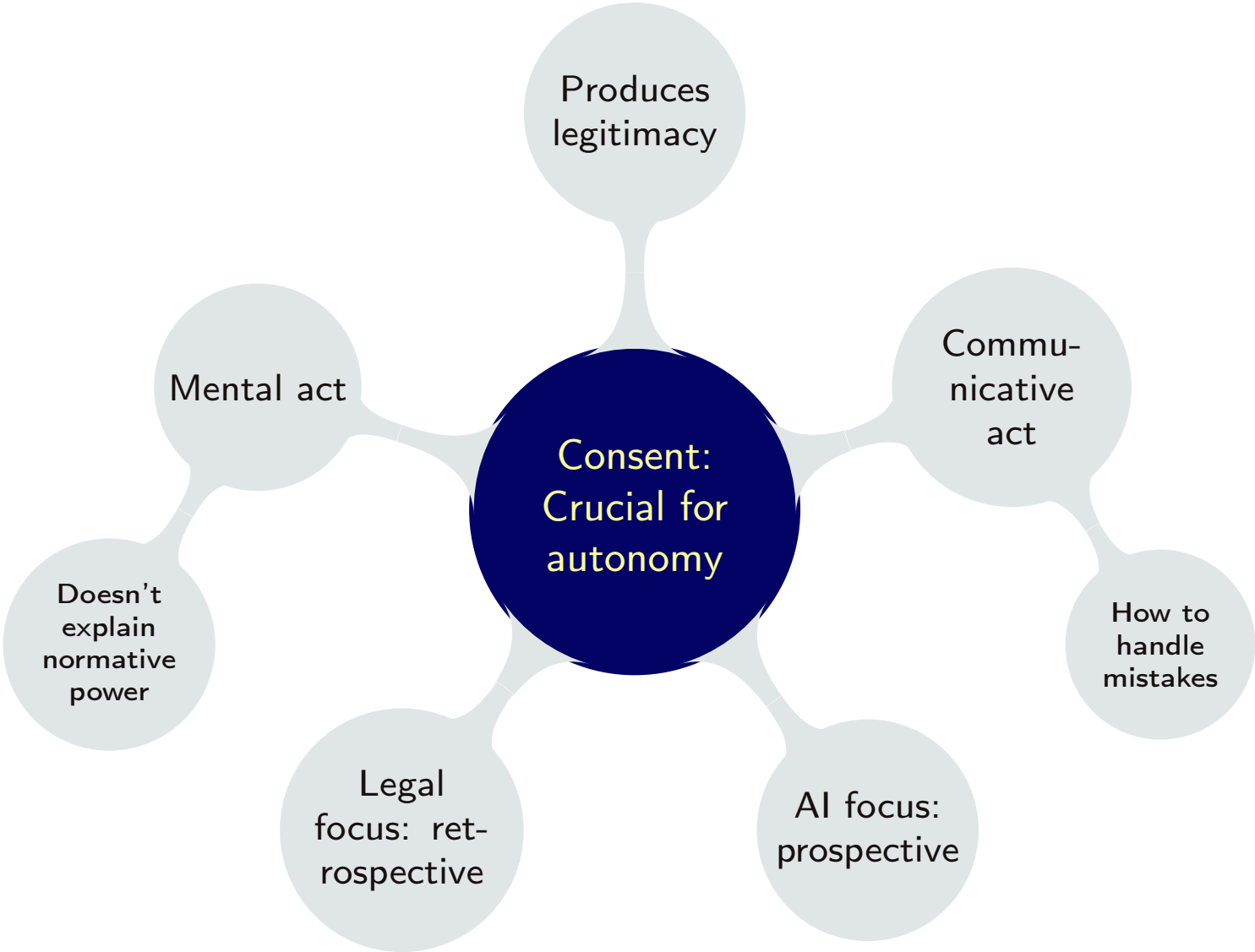
Law and Ethics

Often at odds with each other because laws can deviate from values

Law applies existing norms; ethics critically evaluates existing norms



Consent as a Challenge



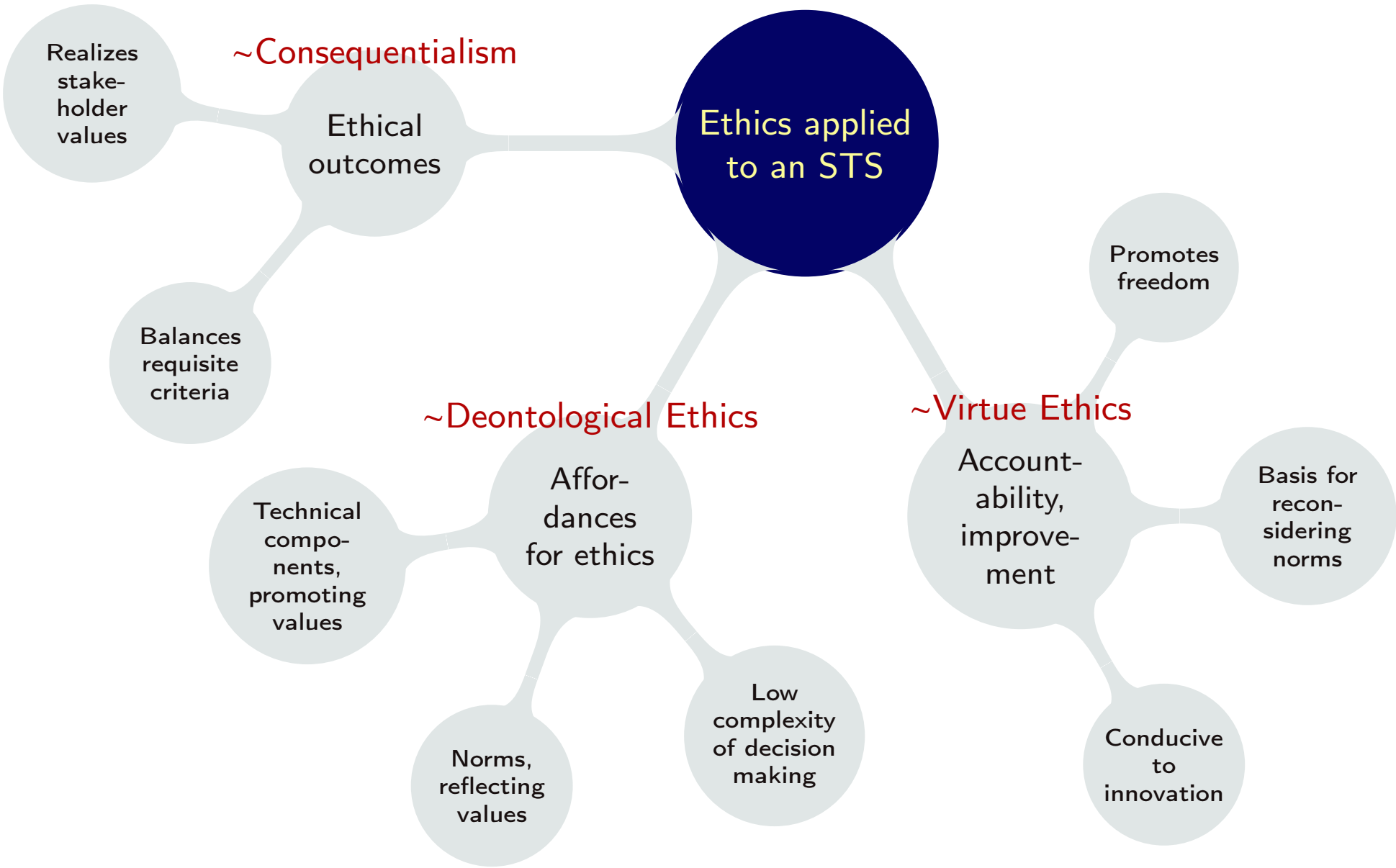
Privacy Law: A Well-Developed Theme Relating to Ethics

Margot Kaminski and Casey, Farhangi, & Vogl on GDPR

How can we incorporate these concerns in a computational framework for STS?

Outline and Schedule

Ethics in the Large: Values and Outcomes



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	About data and algorithms	About 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

Thanks!

- Amit Chopra, Hui Guo, Catholijn Jonker, Özgür Kafalı
- National Science Foundation (IIS-2116751)
- Science of Security Lablet

https://sites.google.com/view/ai-ethics

https://niravajmeri.github.io
https://ii.tudelft.nl/ pradeep/
https://www.csc.ncsu.edu/faculty/mpsingh/
https://research.csc.ncsu.edu/mas

Bibliography I

Matthew D. Adler. *Measuring Social Welfare: An Introduction*. Oxford University Press, New York, 2019.

Rishabh Agrawal, Nirav Ajmeri, and Munindar P. Singh. Socially intelligent genetic agents for the emergence of explicit norms. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 10–16, Vienna, July 2022. IJCAI. doi: 10.24963/ijcai.2022/2.

Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Arnor: Modeling social intelligence via norms to engineer privacy-aware personal agents. In *Proceedings of the 16th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 230–238, São Paulo, May 2017. IFAAMAS. doi: 10.5555/3091125.3091163.

Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Designing ethical personal agents. *IEEE Internet Computing (IC)*, 22(2):16–22, March 2018a. doi: 10.1109/MIC.2018.022021658.

Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Robust norm emergence by revealing and reasoning about context: Socially intelligent agents for enhancing privacy. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 28–34, Stockholm, July 2018b. IJCAI. doi: 10.24963/ijcai.2018/4.

Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Elessar: Ethics in norm-aware agents. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 16–24, Auckland, May 2020. IFAAMAS. doi: 10.5555/3398761.3398769.

Bibliography II

Paul Bremner, Louise A. Dennis, Michael Fisher, and Alan F. T. Winfield. On proactive, transparent, and verifiable ethical reasoning for robots. *Proceedings of the IEEE*, 107(3): 541–561, March 2019. doi: 10.1109/JPROC.2019.2898267.

Paolo Bresciani, Anna Perini, Paolo Giorgini, Fausto Giunchiglia, and John Mylopoulos. Tropos: An agent-oriented software development methodology. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 8(3):203–236, May 2004. doi: 10.1023/B:AGNT.0000018806.20944.ef.

Bryan Casey, Ashkon Farhangi, and Roland Vogl. Rethinking explainable machines: The GDPR's "right to explanation" debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law Journal*, 34(1):143–188, May 2019. doi: 10.15779/Z38M32N986.

Amit K. Chopra and Munindar P. Singh. Sociotechnical systems and ethics in the large. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, pages 48–53, New Orleans, February 2018. ACM. doi: 10.1145/3278721.3278740.

René V Dawis. Vocational interests, values, and preferences. In Marvin D. Dunnette and Leaetta M. Hough, editors, *Handbook of Industrial and Organizational Psychology*, volume 2, pages 833–871. Consulting Psychologists Press, 1991.

Veljko Dubljević, Sebastian Sattler, and Eric Racine. Deciphering moral intuition: How agents, deeds, and consequences influence moral judgment. *PLOS ONE*, 13(10):1–28, 2018. doi: 10.1371/journal.pone.0204631.

Bibliography III

Philippa Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5: 5–15, 1967.

Batya Friedman, David G. Hendry, and Alan Borning. A survey of value sensitive design methods. *Foundations and Trends in Human-Computer Interaction*, 11(2):63–125, November 2017. ISSN 1551-3955. doi: 10.1561/1100000015. URL `https://doi.org/10.1561/1100000015`.

Paul H. P. Hanel, Lukas F. Litzellachner, and Gregory R. Maio. An empirical comparison of human value models. *Frontiers in Psychology*, 9:1643:1–1643:14, 2018. ISSN 1664-1078. doi: 10.3389/fpsyg.2018.01643. URL `https://www.frontiersin.org/article/10.3389/fpsyg.2018.01643`.

Thomas E. Hill Jr. *Virtue, Rules, and Justice: Kantian Aspirations*. Oxford University Press, New York, 2012.

IEEE. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. `https://ethicsinaction.ieee.org`, March 2019. IEEE Global Initiative for Ethically Aligned Design.

Margot E. Kaminski. The right to explanation, explained. *Berkeley Technology Law Journal*, 34 (1):189–218, May 2019. doi: 10.15779/Z38TD9N83H.

Bibliography IV

Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. Axies: Identifying and evaluating context-specific values. In *Proceedings of the 20th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, pages 799–808, London, 2021.

Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. What values should an agent align with? an empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems*, X(Y):1–36, 2022. To appear.

Pradeep K. Murukannaiah and Munindar P. Singh. Xipho: Extending Tropos to engineer context-aware personal agents. In *Proceedings of the 13th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 309–316, Paris, May 2014. IFAAMAS. doi: 10.5555/2615731.2615783.

Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. New foundations of ethical multiagent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1706–1710, Auckland, May 2020. IFAAMAS. doi: 10.5555/3398761.3398958. Blue Sky Ideas Track.

Milton Rokeach. *The nature of human values*. Free press, 1973.

Shalom H. Schwartz. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1):11:1–11:20, 2012. ISSN 2307-0919. doi: 10.9707/2307-0919.1116.

Bibliography V

Munindar P. Singh. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):21:1–21:23, December 2013. doi: 10.1145/2542182.2542203.

Walter Sinnott-Armstrong. Consequentialism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Stanford, Summer 2019 edition, 2019. URL `https://plato.stanford.edu/entries/consequentialism/`.