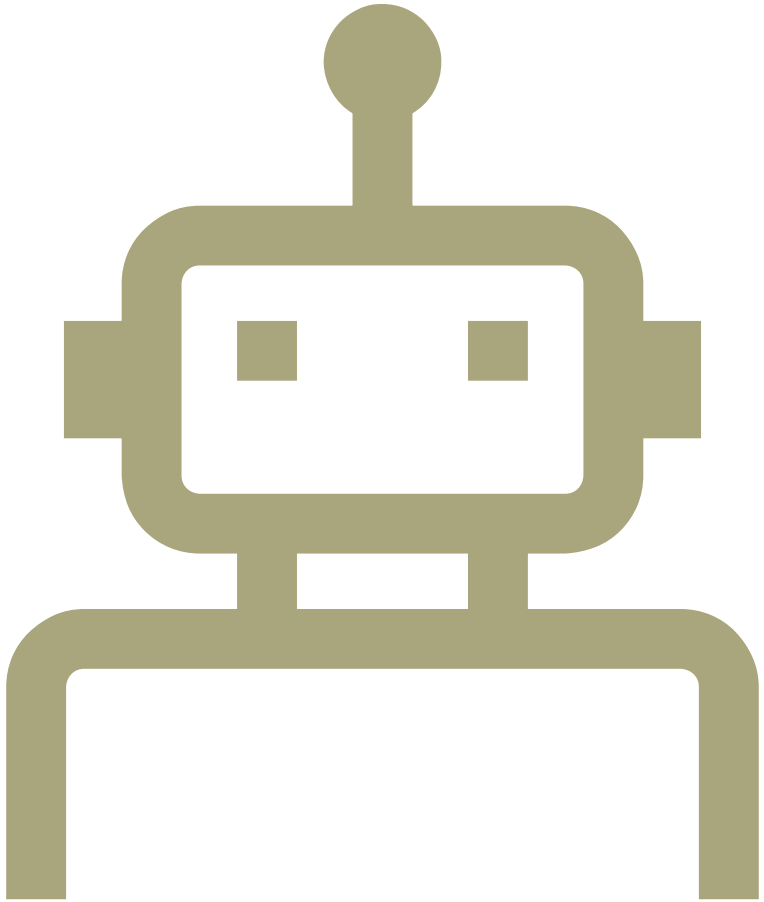
Machine Ethics

Why is it challenging?

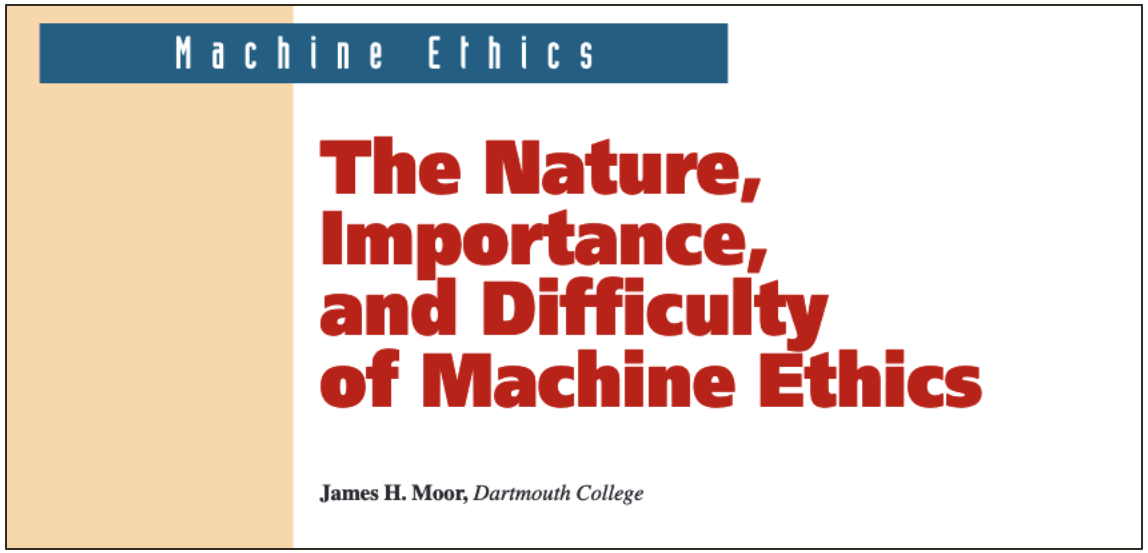


What is AI? (an agent-based definition)

- As per Poole and Mackworth (2017) "Artificial Intelligence is the field that studies the synthesis and analysis of computational agents that act intelligently".
- Agent = an entity that acts in an environment
- Computational agent = an agent whose decisions about its actions can be explained in terms of computation
- We will look at how computational agents could make ethical decisions.

Machine Ethics

- How to automate moral reasoning for computational agents?



* Moor, James H.: The Nature, Importance, and Difficulty of Machine Ethics. In: IEEE Intelligent Systems (2006), Juli, Nr. 4, S. 18–21.

Humans are machines and humans have ethics.

Machine ethics does not exist because ethics is simply emotional

Could a computer operate ethically because it is internally ethical in some way?

* *Moor, James H.: The Nature, Importance, and Difficulty of Machine Ethics. In: IEEE Intelligent Systems (2006), Juli, Nr. 4, S. 18–21.*

Machine Ethics -- Ethical Agents

Ethical-impact agents: Designing a machine solution for a specific task, which impacts ethical issues. (ex: loan system)

Implicit ethical agents: Constraining the machine's actions to avoid unethical outcomes. (ex: banking agents)

Explicit ethical agents: Representing ethics explicitly. (ex: modeling privacy preferences as logic-based rules)

Full ethical agents: Making judgments with justifications while having features such as consciousness, intentionality and free will.

Developing Explicit Ethical Agents

- They fall short of being full ethical agents, BUT they could prevent help prevent unethical outcomes.
- Why is Machine Ethics important?
  - We want machines to treat us well!
  - Future machines will likely have increased control and autonomy. They will need more powerful machine ethics.
  - We should also understand ethics. Programming or teaching a machine to make ethical decisions is also good for us!

** Moor, James H.: The Nature, Importance, and Difficulty of Machine Ethics. In: IEEE Intelligent Systems (2006), Juli, Nr. 4, S. 18–21.*

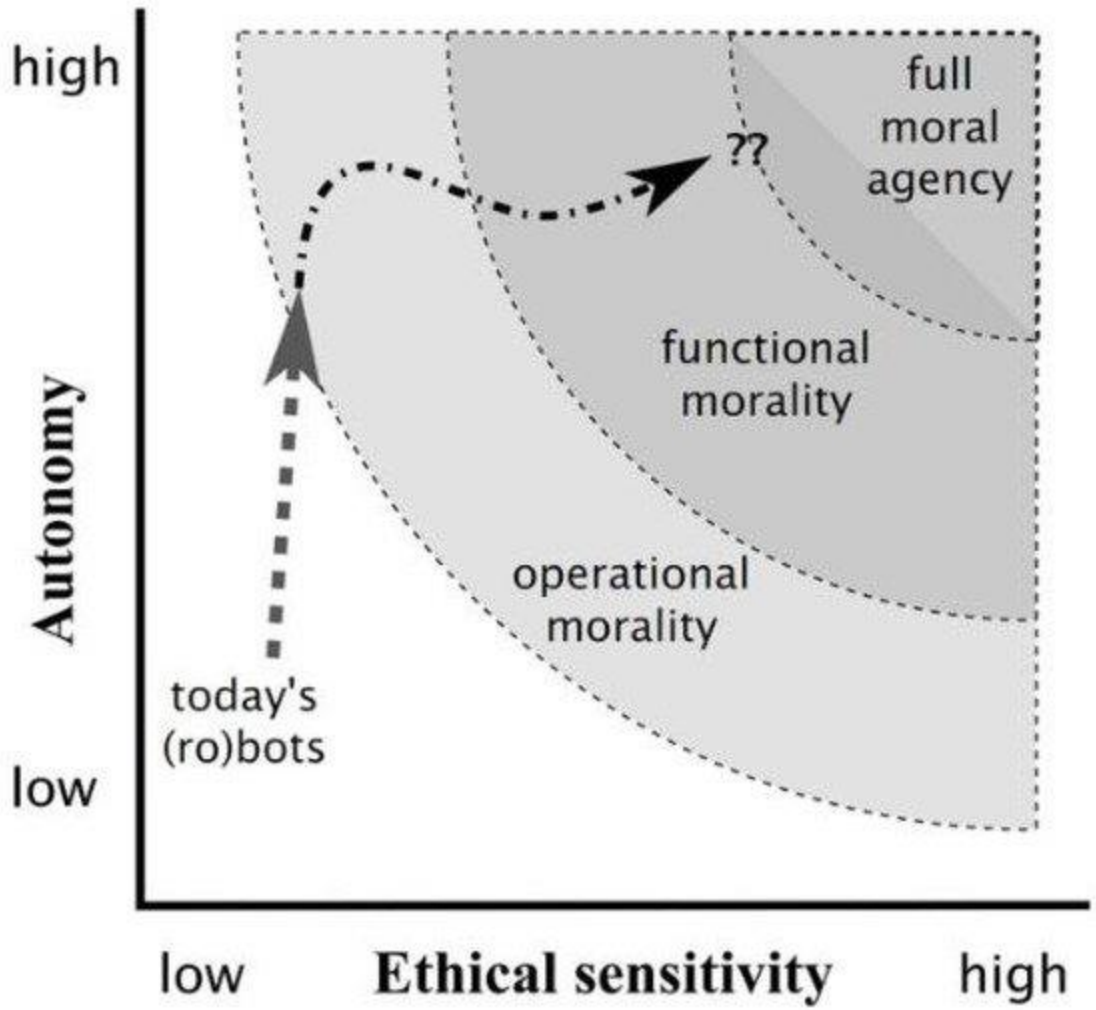
Why is Machine Ethics a "myth"?

- We have a limited understanding of ethical theories.
  - Disagreement on the subject
  - Conflicting ethical intuitions and beliefs
  - Different than programming an agent to do some complex task where moves are well defined (e.g., chess)
- We need to understand learning better (e.g., machine learning etc.)
- Computers have limited commonsense knowledge.

[PS: all three items are still hot topics in research!]

** Moor, James H.: The Nature, Importance, and Difficulty of Machine Ethics. In: IEEE Intelligent Systems (2006), Juli, Nr. 4, S. 18–21.*

Another Categorization of Machine Ethics



Wallach and Allen. 2008. Moral machines: Teaching robots right from wrong. Oxford University Press, Oxford, UK.

How to implement Machine Ethics?

- Top-Down
  - Start with an ethical theory, identify smaller problems and solve them.
  - Pros: no need to identify additional problems
  - Cons: Not clear from the beginning if subproblems are solvable
- Bottom-Up
  - Start with data, and learn ethical behavior from data.
  - Pros: Subproblems are solvable
  - Cons: Non-necessary subproblems may be dealt with.

Wallach and Allen. 2008. Moral machines: Teaching robots right from wrong. Oxford University Press, Oxford, UK.

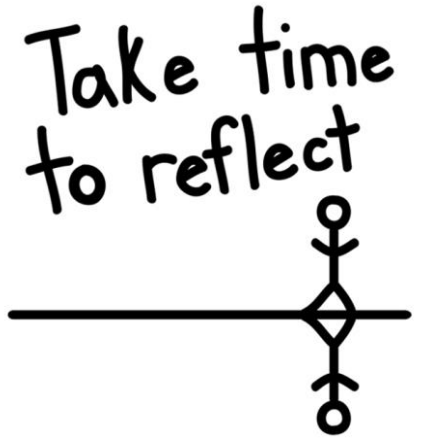
Another taxonomy by Louise Dennis

- Constraint-Based Ethical Systems
  - Ethics is placed on some sub-system that guides/constrains the actions of other parts of the system.
  - Other parts of the system can guide the decision-making process of the agents.
- Global Ethical Systems
  - All decisions are ethical.

Give me the taxonomy!

- The truth is there is no clear taxonomy.
- Let's think about the following question according to Moore's agent types:

Could you find example AI systems that could be assigned to more than one ethical agent type? Justify your response.



Social Choice and Machine Ethics

- We often talk about implementing values or obligations.
- We are now interested in the question of whose values/obligations a machine should implement.
- Once we know what we want to implement, we can develop algorithms to verify machine ethics systems (e.g., Isabelle).

Consequentialist Theories (revisited)

- Ethical Egoism
  - Focuses on own best interests
- Utilitarianism
  - Focuses on everyone
  - Act-utilitarianism:
    - from individual to society
  - Rule-utilitarianism:
    - A rule to follow to achieve overall good

Social Choice Ethics in AI

AI & Soc (2020) 35:165–176
DOI 10.1007/s00146-017-0760-1

ORIGINAL ARTICLE

Social choice ethics in artificial intelligence

Seth D. Baum[1]

Received: 17 July 2016 / Accepted: 16 September 2017 / Published online: 30 September 2017

Social Choice Ethics in AI

- Goal: Designing AI to act according to the aggregate views of society (i.e., bottom-up).
- AI faces three sets of decisions:
  - Standing (whose ethics views)
  - Measurement (identifying views)
  - Aggregation (combining to a single view)
- Non-social ethics could be even more challenging
  - Considering future generations, or the AI itself

Seth D. Baum. 2020. Social choice ethics in artificial intelligence. AI & Society 35, 1 (2020), 165–176.

Explicit ethical agents: Representing ethics explicitly.

AI for Privacy:
A Multiagent Perspective

Cybersecurity

Preserving Privacy in an Online World

How to represent the actual privacy preferences of users?

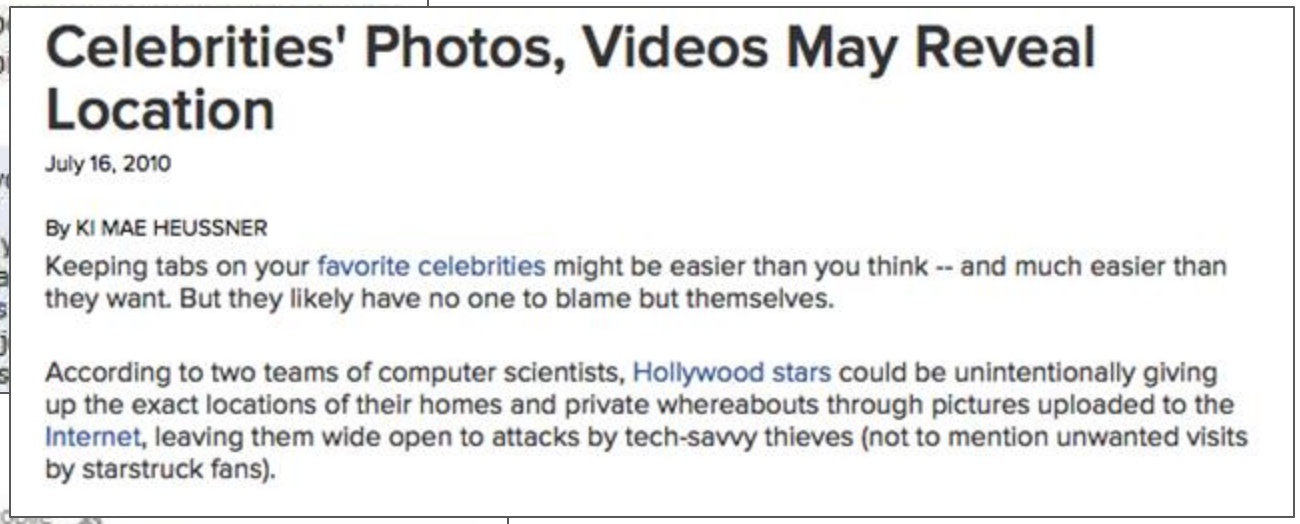
How to elicit the privacy preferences from users?

How to advise the users to take actions to preserve their privacy?

How to agree on how a co-owned content will be shared?

How to explain privacy decisions?

Real Life Scenarios*



Celebrities' Photos, Videos May Reveal Location

July 16, 2010

By KI MAE HEUSSNER

Keeping tabs on your favorite celebrities might be easier than you think -- and much easier than they want. But they likely have no one to blame but themselves.

According to two teams of computer scientists, Hollywood stars could be unintentionally giving up the exact locations of their homes and private whereabouts through pictures uploaded to the Internet, leaving them wide open to attacks by tech-savvy thieves (not to mention unwanted visits by starstruck fans).

Our online survey with 330 participants shows that more than 90% of privacy violations occur through inference.

** sources: http://www.oddee.com/item_96937.aspx; http://abcnews.go.com/Technology/celebrity-stalking-online-photos-give-location/story?id=11162352*

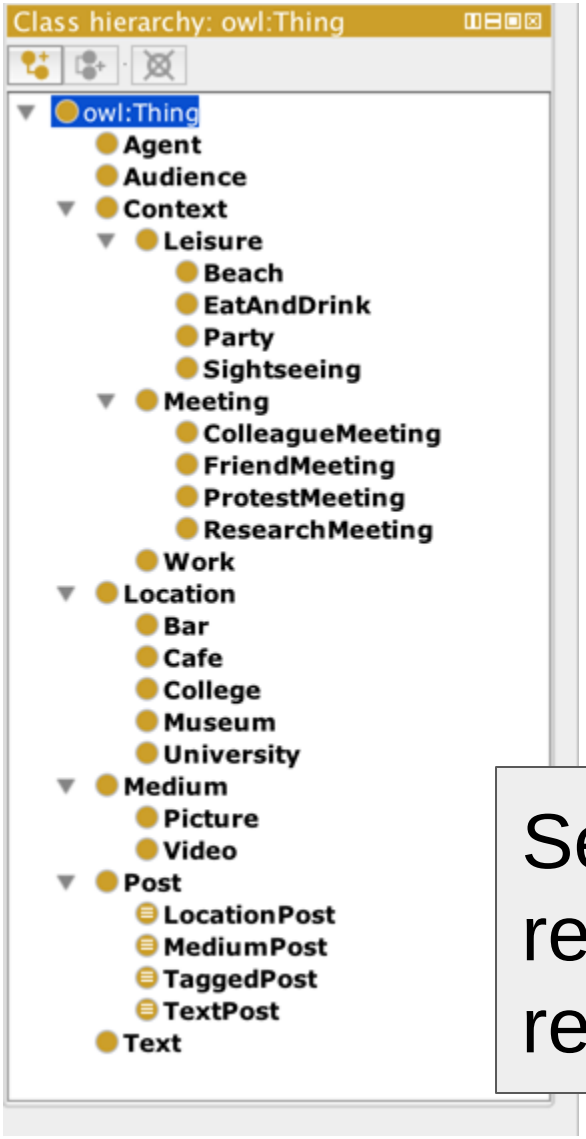
Understanding privacy violations

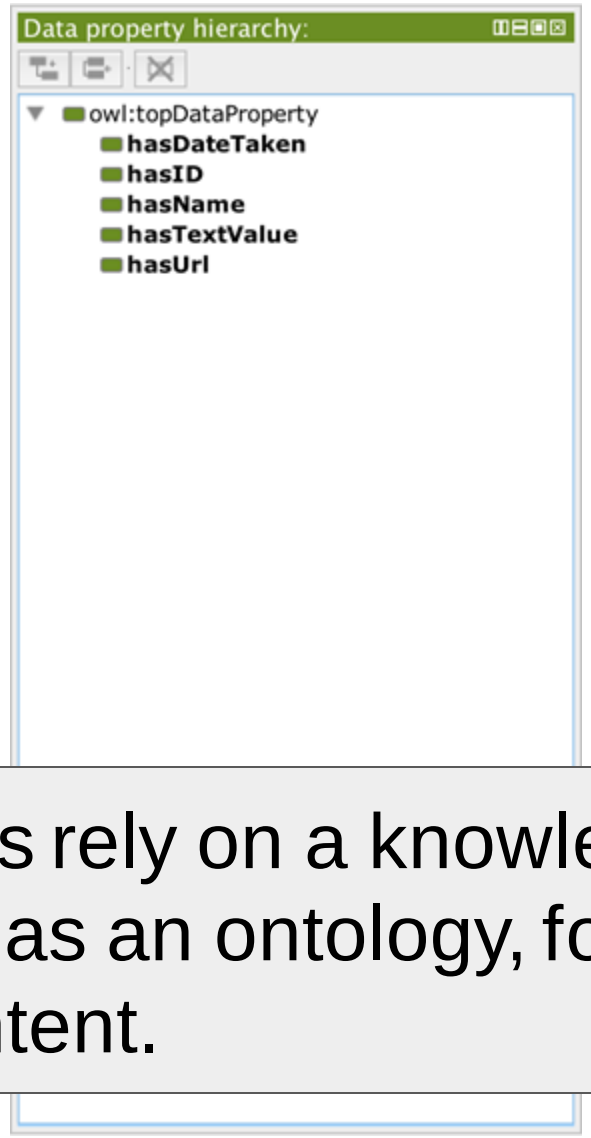
Privacy Concerns of Dennis

Dennis wants his friends to see his pictures but not his location.

	No inference	**Inference**
User	(i) Dennis checks in at a restaurant.	(iii) Dennis shares a picture without declaring his location. It turns out that his picture is geo-tagged.
Others	(ii) Charlie shares a picture with everyone. He tags Dennis in it as well.	(iv) Charlie checks in at a restaurant. At the same time, Dennis shares a picture of Charlie.

Content Ontology



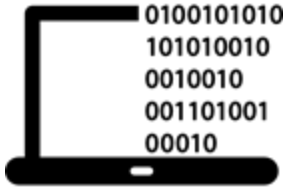


Semantic approaches rely on a knowledge representation, such as an ontology, for reasoning on the content.

Privacy Preferences

P_{E_2}: *hasMedium*(?pr, ?m), *taggedPerson*(?m, :eve),
isInContext(?pr, ?ctx), Work(?ctx) → *rejects*(:eve, ?pr)
[Eve rejects posts that are in work context.]

We can build software agents that can reason on users' privacy preferences.



Detection of Privacy Violations: PriGuard

- We represent the social network as an *agent-based online social network* (ABSN).
- Agents know the privacy preferences of their users.
- We develop a sound and complete algorithm to detect privacy violations.
- We show the scalability of the approach on real-life social networks.

PriGuard can detect privacy violations and notify the users to take actions.

Nadin Kökciyan and Pinar Yolum. 2016. "PriGuard: A Semantic Approach to Detect Privacy Violations in Online Social Networks," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2724-2737

Prevention of Privacy Violations: PriArg

- Agents discuss on a post *before* it is shared.
- We develop a framework that enables agents to carry out a dialogue with other agents.
- We adapt **computational argumentation** to enable privacy decision-making.

Argumentation serves as a useful technique to mimic how humans deal with privacy disputes.

Nadin Kökciyan, Nefise Yaglikci, and Pinar Yolum. 2017. An Argumentation Approach for Resolving Privacy Disputes in Online Social Networks. *ACM Transactions on Internet Technology* 17, 3, Article 27 (June 2017), 22 pages.

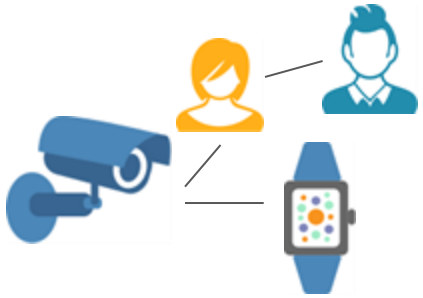
Prevention of Privacy Violations: PriNego

- PriNego is a negotiation-based approach where agents negotiate with each other on their privacy preferences.
- Agents use different negotiation strategies to preserve their users' privacy.
- It exploits reciprocity as a heuristic (e.g., this time you help me, next time I help you).

Agreement can be established over multiple posts.

Dilara Kekulluoglu, Nadin Kökciyan, and Pinar Yolum. 2018. Preserving Privacy as Social Responsibility in Online Social Networks. *ACM Transactions on Internet Technology* 18, 4, Article 42 (April 2018), 22 pages.

Proactive Agents: an IoT example



- Each IoT entity follows **contextual norms** to calculate the appropriateness of sharing information.
- **Computational argumentation** enables the agent to reason on its knowledge and belief bases under **uncertainty**.
- To make inference based on others' information, a **trust model** needs to be in place.

Agents can choose to violate privacy for a better outcome!

Nadin Kökciyan and Pinar Yolum, 2017, August. Context-Based Reasoning on Privacy in Internet of Things. In *IJCAI* (pp. 4738-4744).

Preserving Privacy in an Online World

How to represent the actual privacy preferences of users?

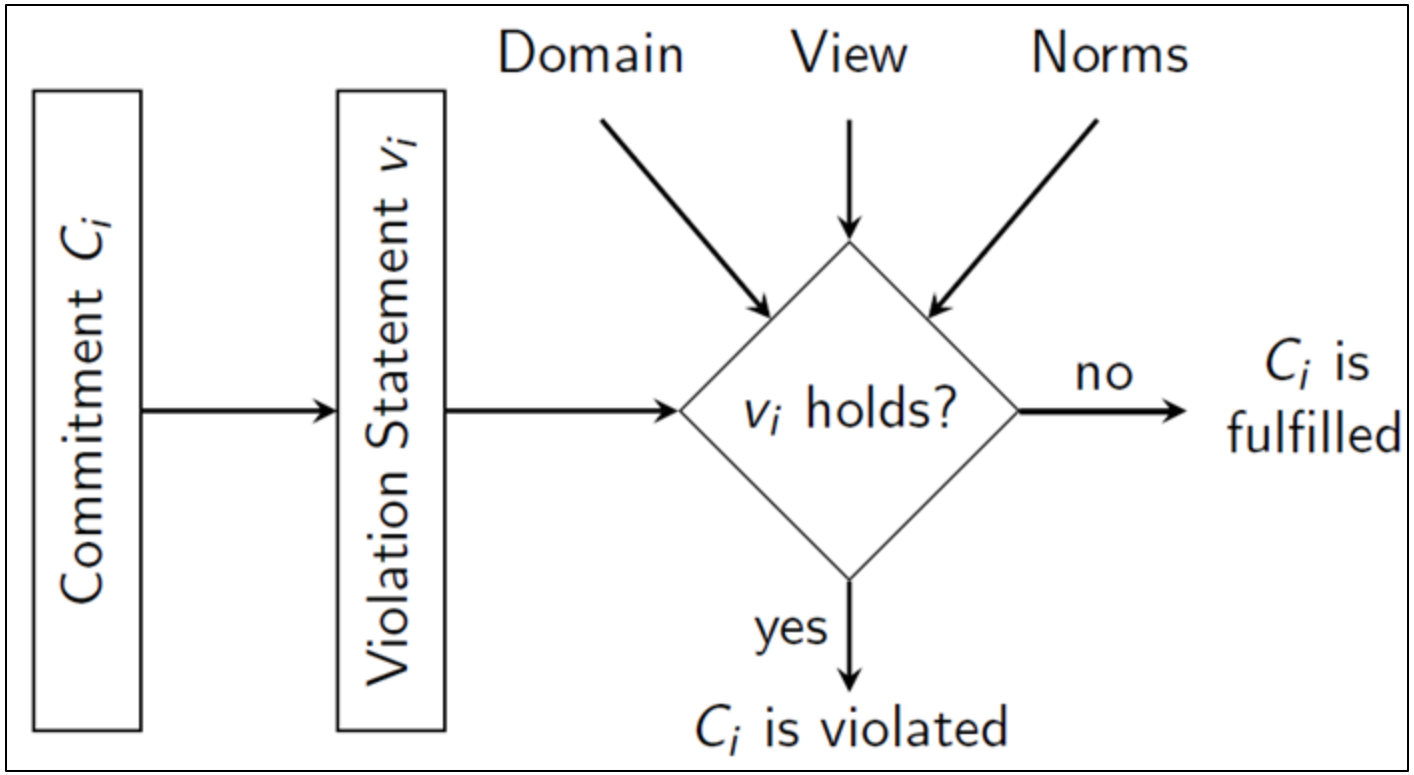
How to elicit the privacy preferences from users?

How to advise the users to take actions to preserve their privacy?

How to agree on how a co-owned content will be shared?

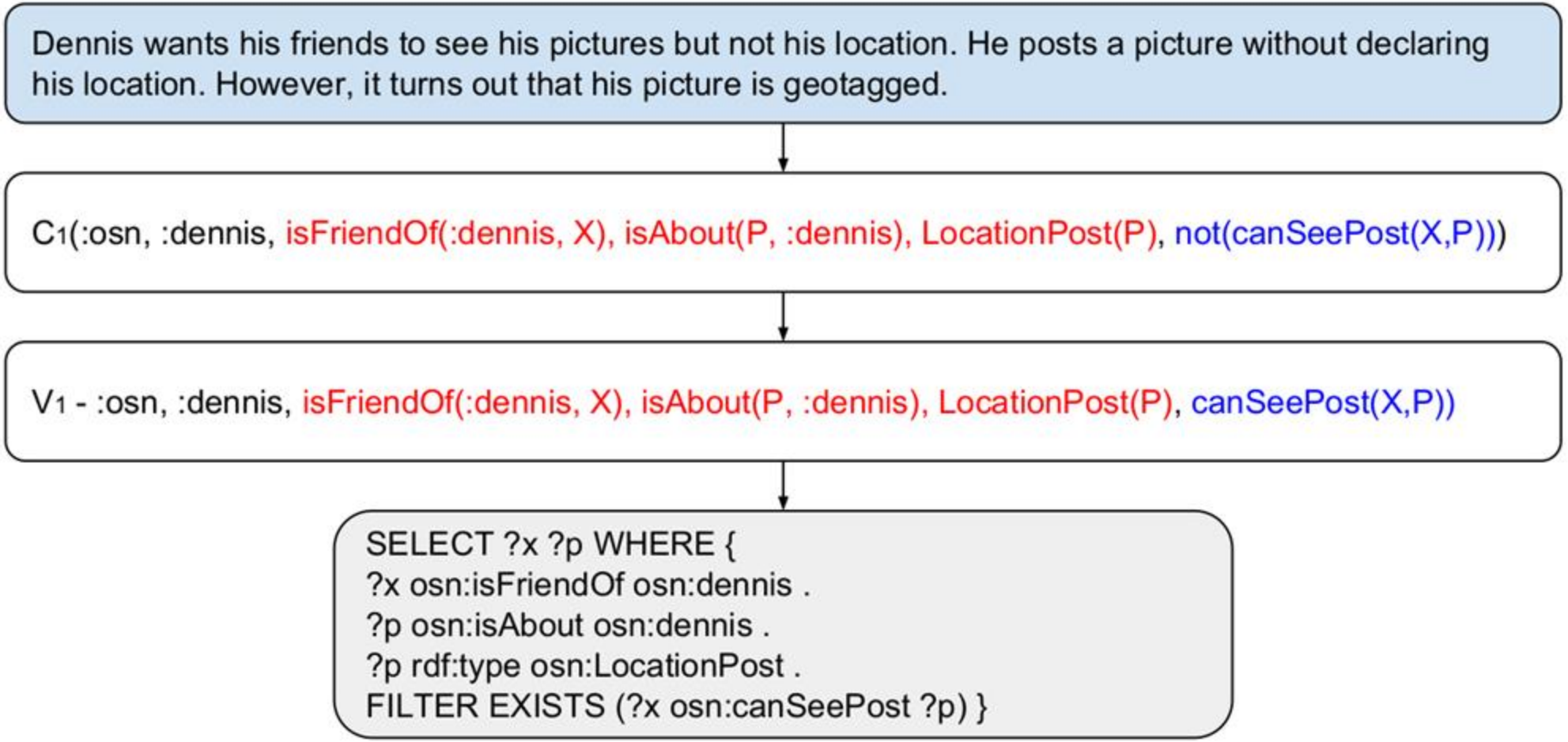
How to explain privacy decisions?

PriGuard: Detection of Privacy Violations



Nadin Kökciyan and Pinar Yolum. 2016. "PriGuard: A Semantic Approach to Detect Privacy Violations in Online Social Networks," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2724-2737

An Example


Dennis wants his friends to see his pictures but not his location. He posts a picture without declaring his location. However, it turns out that his picture is geotagged.
C1(:osn, :dennis, isFriendOf(:dennis, X), isAbout(P, :dennis), LocationPost(P), not(canSeePost(X,P)))
V1 - :osn, :dennis, isFriendOf(:dennis, X), isAbout(P, :dennis), LocationPost(P), canSeePost(X,P))
SELECT ?x ?p WHERE {
?x osn:isFriendOf osn:dennis .
?p osn:isAbout osn:dennis .
?p rdf:type osn:LocationPost .
FILTER EXISTS (?x osn:canSeePost ?p) }

The Social Network Domain

$\texttt{Agent}, \texttt{Post}, \texttt{Audience}, \texttt{Context}, \texttt{Content} \sqsubseteq \top$	$\texttt{Leisure}, \texttt{Meeting}, \texttt{Work} \sqsubseteq \texttt{Context}$
$\texttt{Beach}, \texttt{EatAndDrink}, \texttt{Party}, \texttt{Sightseeing} \sqsubseteq \texttt{Leisure}$	$\texttt{Bar}, \texttt{Cafe}, \texttt{College}, \texttt{Museum}, \texttt{University} \sqsubseteq \texttt{Location}$
$\texttt{Picture}, \texttt{Video} \sqsubseteq \texttt{Medium}$	$\texttt{Medium}, \texttt{Text}, \texttt{Location} \sqsubseteq \texttt{Content}$
$\texttt{Post} \sqcap \exists \mathit{sharesPost}^{-}.\texttt{Agent} \equiv \exists \mathit{R\_sharedPost}.\mathit{Self}$	$\texttt{LocationPost} \equiv \exists \mathit{R\_locationPost}.\mathit{Self}$
$\texttt{LocationPost} \equiv \texttt{Post} \sqcap \exists \mathit{hasLocation}.\texttt{Location}$	$\texttt{MediumPost} \equiv \texttt{Post} \sqcap \exists \mathit{hasMedium}.\texttt{Medium}$
$\texttt{TaggedPost} \equiv \texttt{Post} \sqcap \exists \mathit{isAbout}.\texttt{Agent}$	$\texttt{TextPost} \equiv \texttt{Post} \sqcap \exists \mathit{hasText}.\texttt{Text}$

Norms

N_1: $sharesPost(X,P) \rightarrow canSeePost(X,P)$

[Agent can see the posts that it shares.]

N_2: $sharesPost(X,P) \wedge hasAudience(P,A) \wedge hasMember(A,M) \rightarrow canSeePost(M,P)$

[Audience of a post can see the post.]

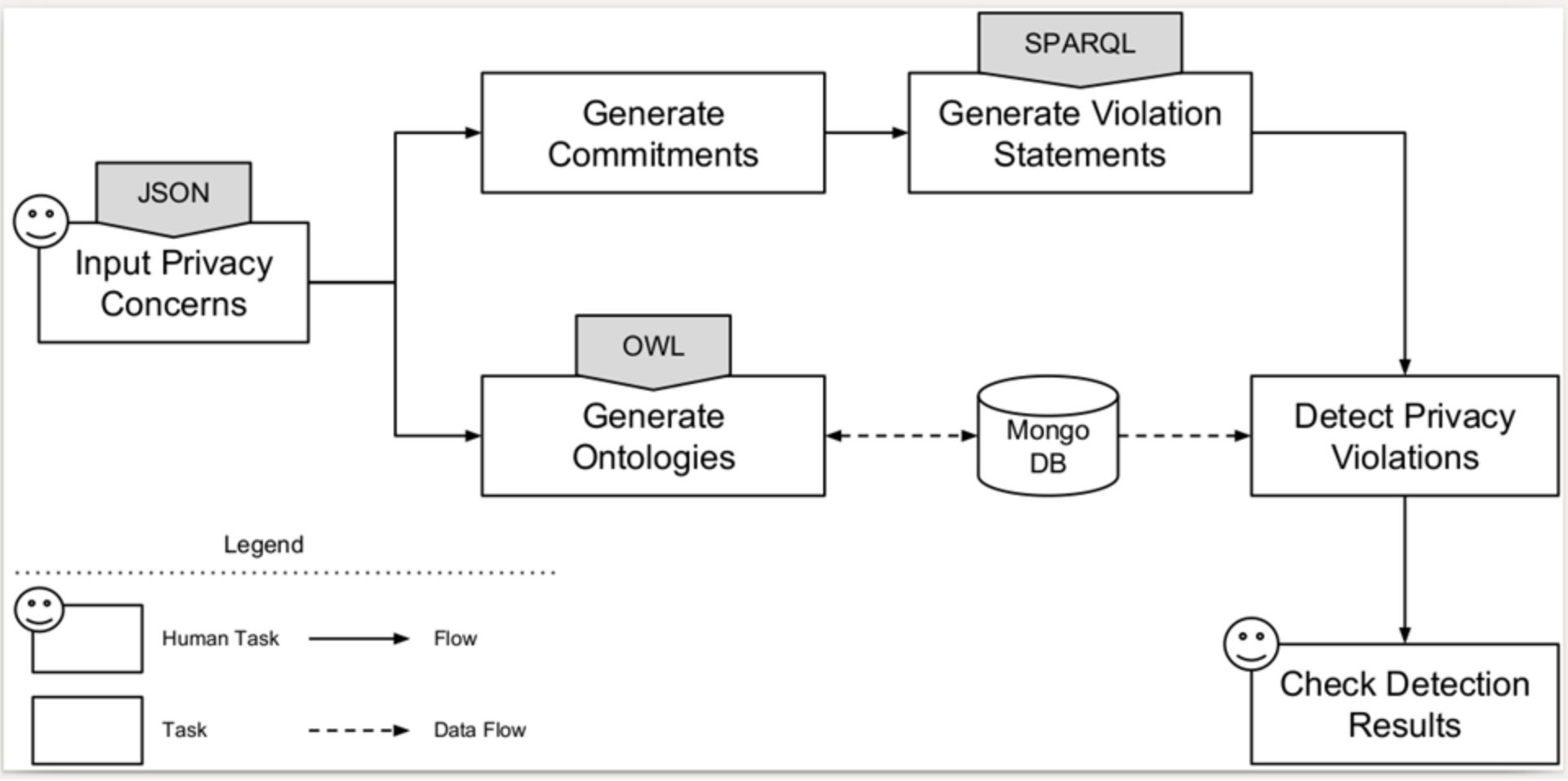
N_3: $hasMedium(P,M) \wedge taggedPerson(M,X) \rightarrow isAbout(P,X)$

[Post is about agents tagged in a medium.]

N_4: $\texttt{Post}(P) \wedge hasMedium(P,M) \wedge hasGeotag(M,T) \rightarrow \texttt{LocationPost}(P)$

[Geotagged medium gives away the location.]

A Facebook Application: PriGuardTool



Kökciyan N., Yolum P. (2016) PriGuardTool: A Web-Based Tool to Detect Privacy Violations Semantically. In: Baldoni M., Müller J., Nunes I., Zalila-Wenkstern R. (eds) Engineering Multi-Agent Systems. EMAS 2016. Lecture Notes in Computer Science, vol 10093. Springer, Cham

Summary

- Machine Ethics is a way to realize Normative Ethics and Applied Ethics together.
- Many categorization systems exist: Moore's Ethical Agents, Wallach and Allen, Louise Dennis …
- Social Choice theory is looking at the problem of understanding values/obligations of a society
- We looked at an example of explicit ethical agents in the privacy domain
- **Check materials for the applied Machine Ethics (this is a required component for this week!)**