Universität
Münster

What Can Explainability Mean for Probabilistic Inference?

Human-aware PGMs and Probabilistic Inference via Lifted Model Reconciliation – A KI Starter Project Starting in 2024

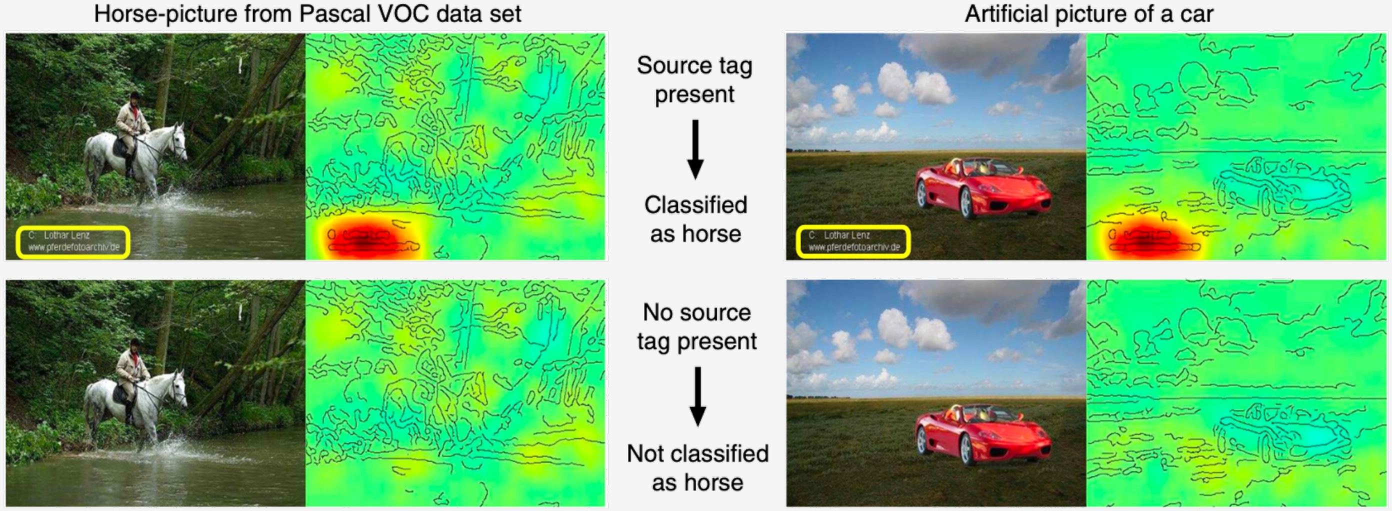
Tanya Braun

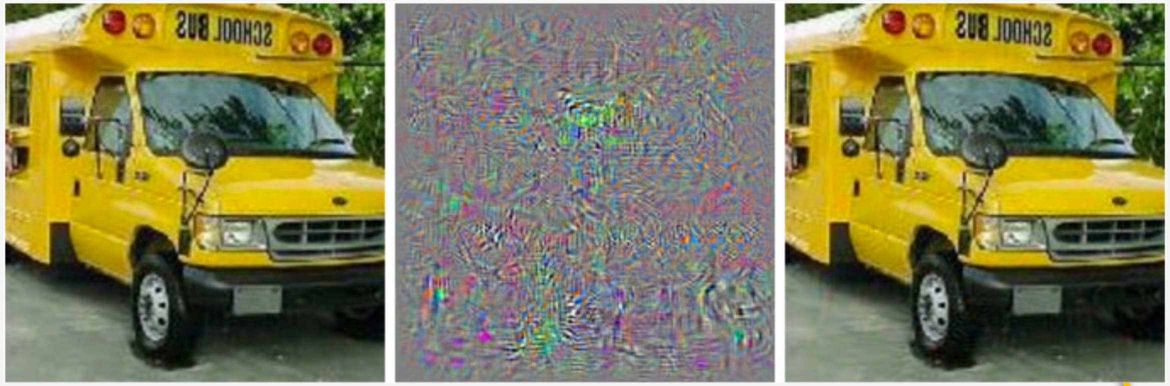
living.knowledge

Data Science Group
Computer Science Department

Explainable AI (XAI)

- Explanations critical for trust, collaboration, …
- *Explain classifications*
  - Debugging tool for inscrutable representations
    - "Pointing" explanations





Prediction: School bus | Difference between left and right magnified by 10 | Prediction: Ostrich

Please point to the "ostrich" part

Figure above: Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller: Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. In *Nature Communications*, 2019.
Figure below: Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus: Intriguing properties of neural networks. In ICLR-14, 2014.

Universität Münster

Explainable AI (XAI)

- *Explain decisions / plans*
  - Classical planning: Given a planning problem (Σ, s_0, S_g) (agent model $\mathcal{M}^R$)
    - Find a plan $\pi = \langle a_1, a_2, \dots, a_n \rangle$ that transforms s_0 to a state $s_n \in S_g$
  - Human-aware planning: Team of human and agent (e.g, robot)
    - Agent's model $\mathcal{M}_r^H$ of human's model $\mathcal{M}^H$
      - Allows the agent to **anticipate** human behaviour to assist, avoid, team
    - Agent's model $\mathcal{M}_h^R$ that the agent expects the human to have of $\mathcal{M}^R$
      - Allows the agent to **anticipate** human expectations to conform to those expectations, explain its own behaviour in terms of those expectations
  - Interpretability: Set up explicable / legible / predictable / … plans
  - Explanations: *Reconcile* model differences

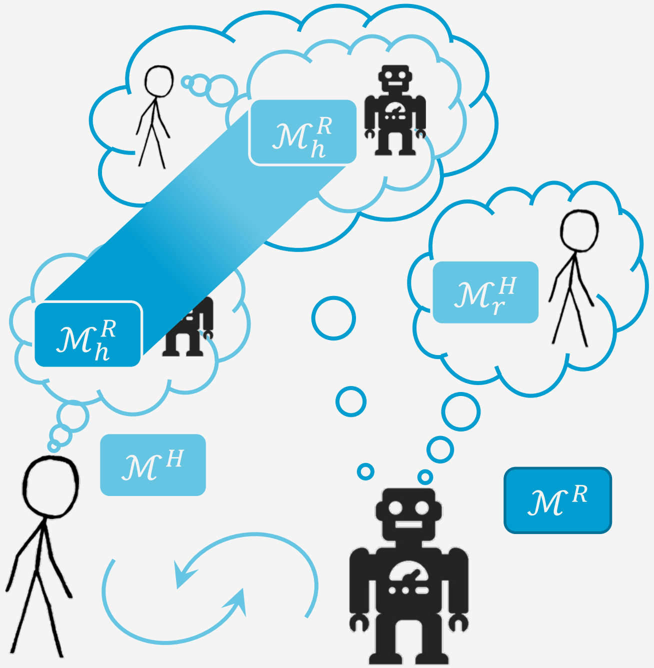
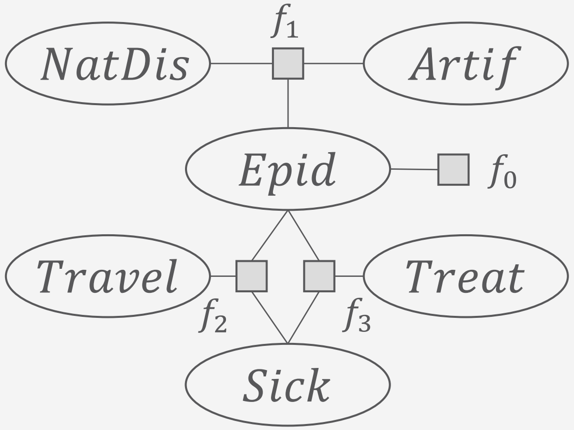


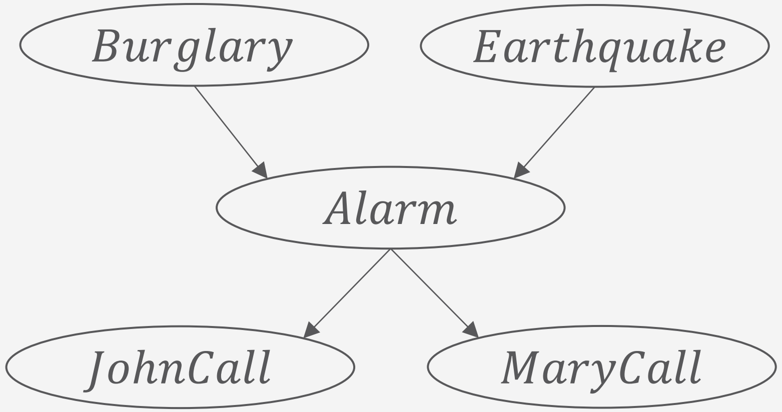
Figure based on Kambhampati

Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati: Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI-17*, 2017.
Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. The Emerging Landscape of Explainable Automated Planning & Decision Making. In *IJCAI-20*, 2020.

What about Probabilistic Inference?



- Probabilistic graphical models (PGMs)
  - Set of random variables, set of conditional probability distributions / factors, connecting random variables
    - E.g., Bayesian network (above), factor graph (below)
  - Semantics: full joint probability distribution as (normalised) product of distributions / factors
- Probabilistic inference
  - Query for probability (distribution) of event / random variable
    - E.g.,
      - $P(Burglary = true|MaryCall = true)$
      - $P(Epid|Sick = true)$
  - Solve by eliminating all non-query terms (sum-out)



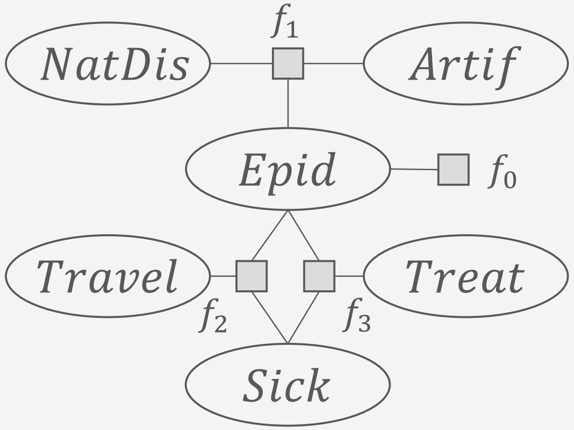
What Can We Find in the Direction of XAI?

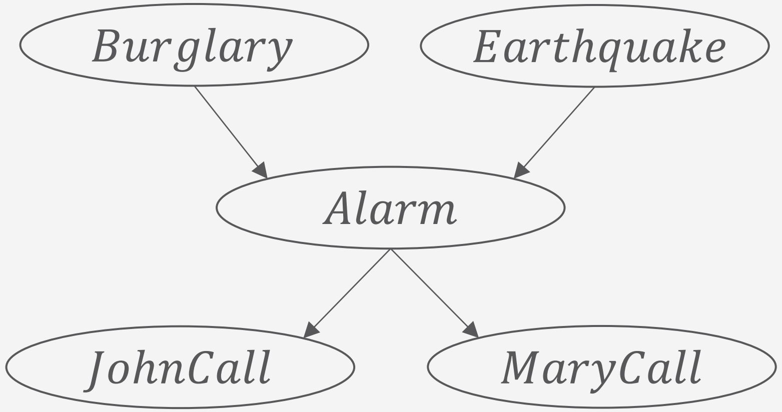
- Most Probable Explanation (MPE)
  - Given a set of events, what is the most probable state of the remaining random variables?
    - Formally, given a PGM G and a set of events $\boldsymbol{e}$:

$$MPE_G(\boldsymbol{e}) = \arg\max_{\boldsymbol{v} \in \mathrm{Val}(\boldsymbol{V})} P(\boldsymbol{v} \mid \boldsymbol{e}) = \arg\max_{\boldsymbol{v} \in \mathrm{Val}(\boldsymbol{V})} P(\boldsymbol{v}, \boldsymbol{e})$$

      - $\boldsymbol{V} = \mathrm{rv}(G) \setminus \mathrm{rv}(\boldsymbol{e})$ random variables of G without those in $\boldsymbol{e}$
      - Max-out instead of sum-out to answer an MPE query instead of a probability query

- Is that really explaining?



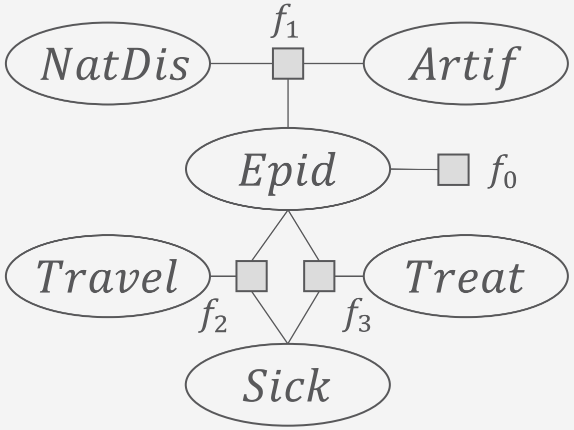


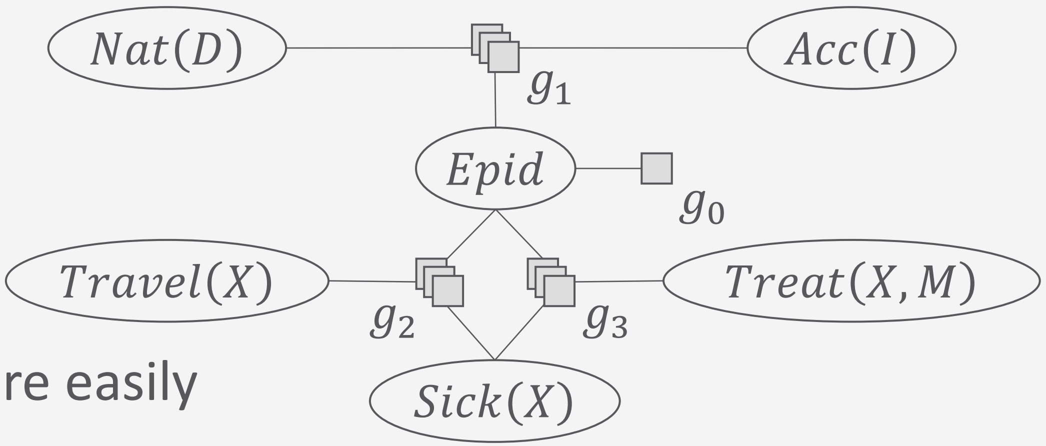
Alexander Philip Dawid. Applications of a General Propagation Algorithm for Probabilistic Expert Systems. *Statistics and Computing*, 2(1):25–36, 1992.
Rina Dechter. Bucket Elimination: A Unifying Framework for Probabilistic Inference. In *Learning and Inference in Graphical Models*, pages 75–104. MIT Press, 1999.
Tanya Braun and Ralf Möller. Lifted Most Probable Explanation. In *ICCS-18*, 2018.

Universität Münster

What If We Go Broader?

- Interpretable PGMs
  - All variables are observable (unrealistic!)
  - Assumption: Knowing the state of the system makes system explainable
  - Problem: Large nets are not readable even if state known
- Abstraction
  - *Lifted* versions for relational / symmetric domains:
    - Logical variables compactly encode repeated structures
      - Semantics: ground, multiply, normalise
      - Inference task defined as before, e.g., $P\left(Epid \middle| Sick(X) = true_{|X \in \{alice, eve, bob\}}\right)$
  - Assumption: Large nets become smaller and thus more easily readable

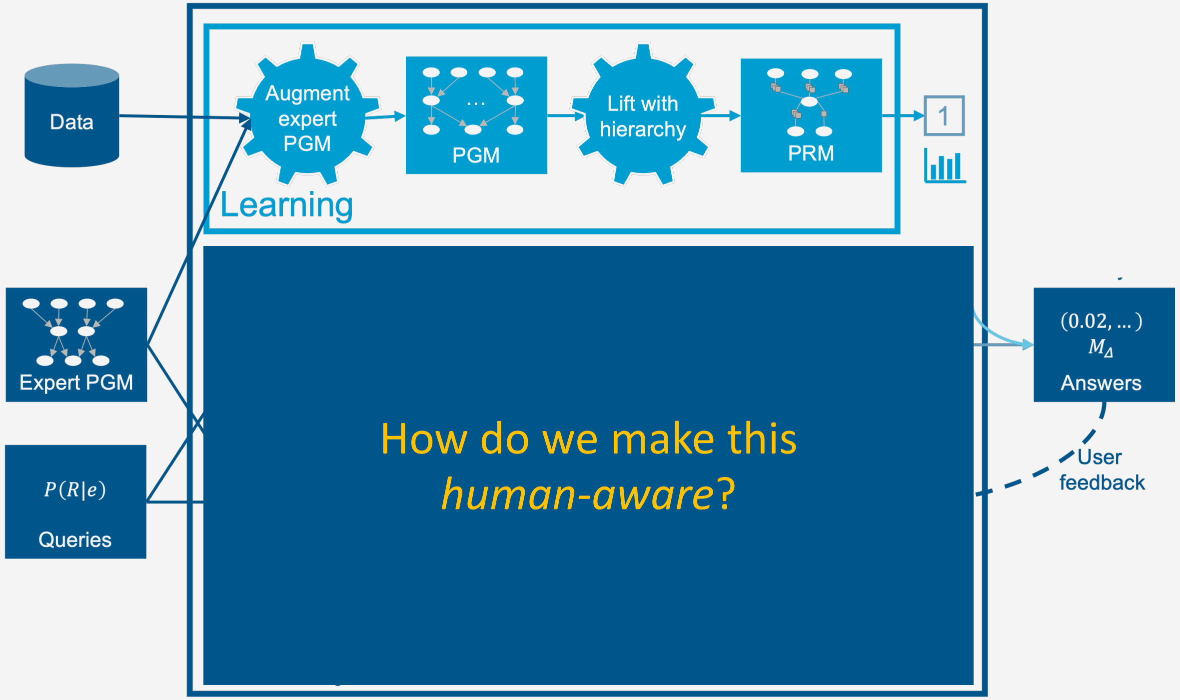




Yitao Liang and Guy Van den Broeck. Towards Compact Interpretable Models: Shrinking of Learned Probabilistic Sentential Decision Diagrams. In *XAI-17 Workshop*, 2017.
David Poole. First-order Probabilistic Inference. In *IJCAI-03*, 2003. Mathias Niepert and Guy Van den Broeck: Tractability through Exchangeability: A New Perspective on Efficient Probabilistic Inference. In *AAAI-14*, 2014. Malte Luttermann, Tanya B, Ralf Möller, and Marcel Gehrke. Colour Passing Revisited: Lifted Model Construction with Commutative Factors. In *AAAI-24*, 2024.

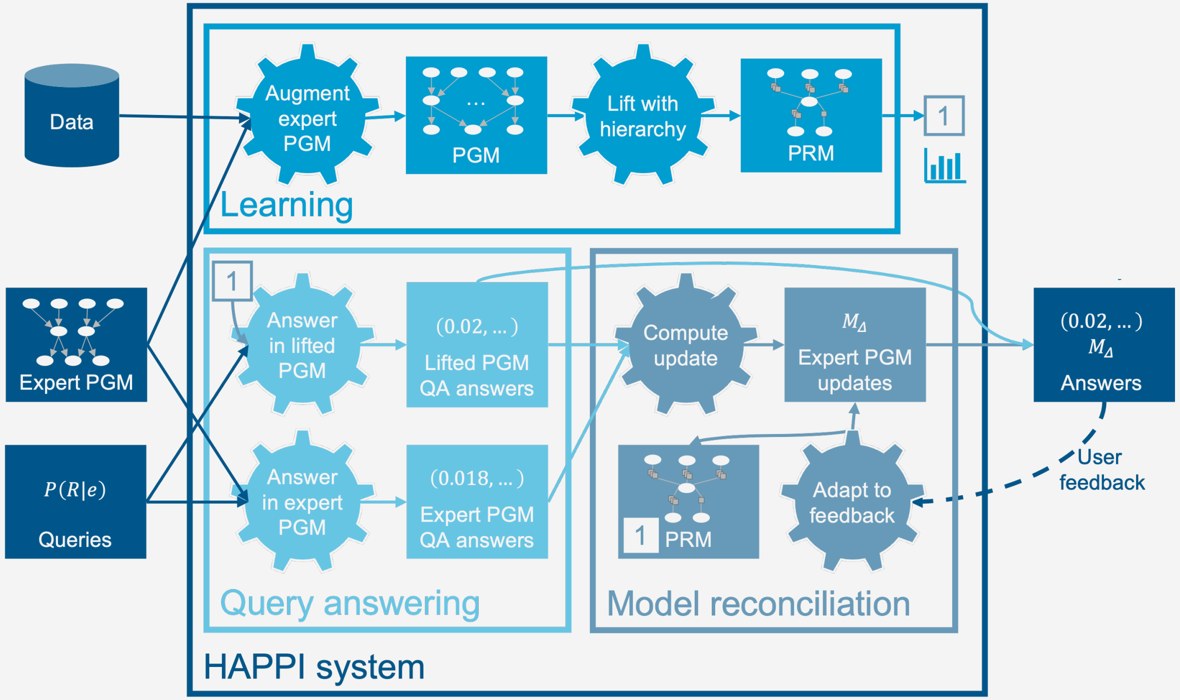
Human-aware Setting of Probabilistic Inference

- Models
  - Domain expert has domain knowledge
    - Implicit human's model
  - Domain expert builds a (small) PGM
    - Agent's model of human's model
  - Learning algorithm expands the PGM using additional training data
    - Can be much larger than the input PGM
    - Agent's model
- Inference
  - Human asks queries
  - Agent answers query



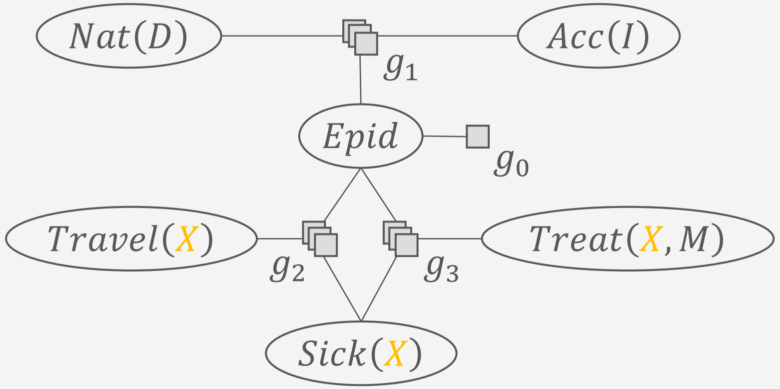
Model Reconciliation for Probabilistic Inference

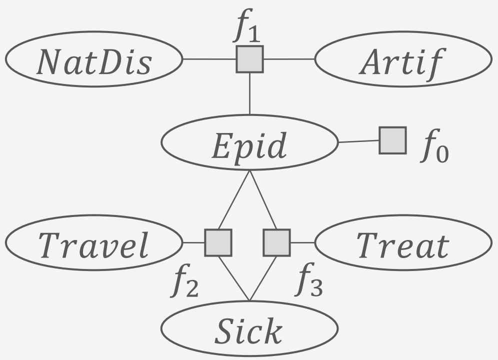
- Reconciling model differences between expert model and trained model
  - Use lifting to increasingly abstract large PGMs to get closer to smaller human's models
    - Hierarchy of models to be able to explain different parts in different detail
  - Reconcile human's and agent's model by exposing differences to human
    - Expose only those parts relevant to a query to update human's model

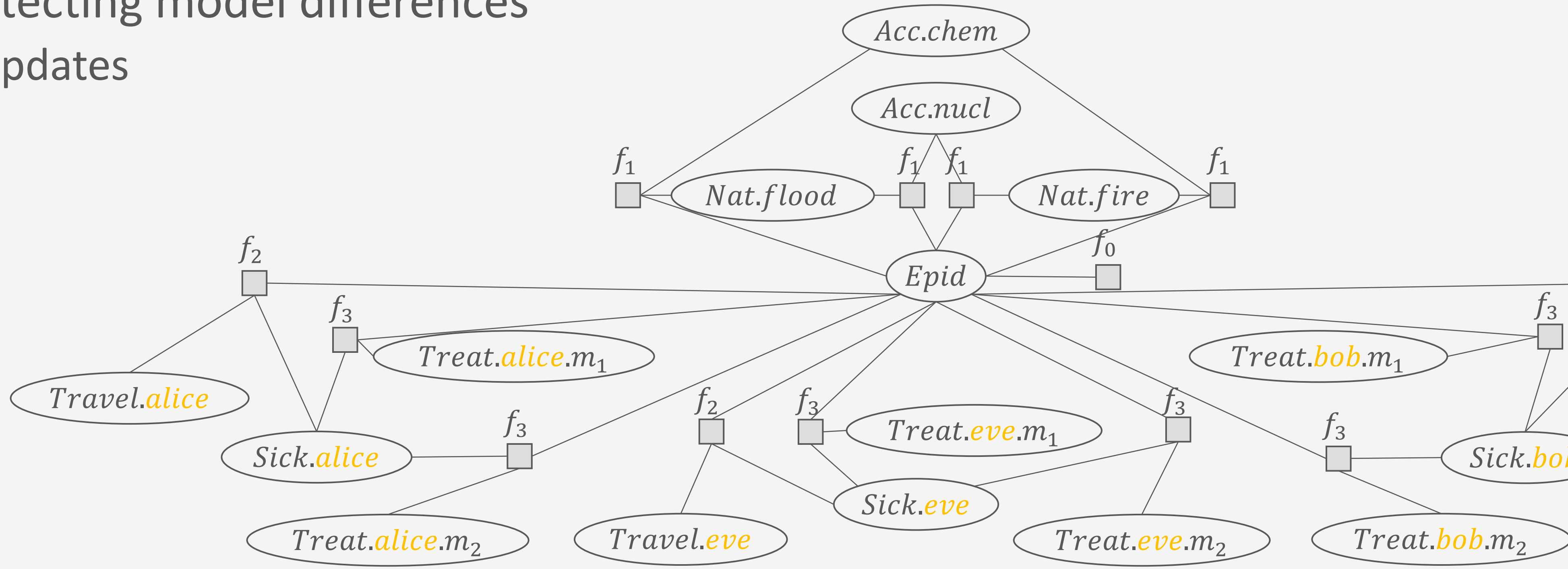


Lifting as Abstraction

- Find (almost) repeated structures and abstract them
  - Colour passing as a variant of Weisfeiler-Leman
- Use as basis for detecting model differences
  - Calculate model updates







KI Starter: HAPPI Reconciliation

Human-aware PGMs and Probabilistic Inference via Lifted Model Reconciliation

- Goal: Provide explanations for probabilistic inference based on model reconciliation between trained and expert models
  - Novel research direction that combines techniques that have exhibited great success in their respective areas
  - Helps push fundamental task of inference into human-oriented AI