Maximizing Diamonds using AIXI and Dynamic Bayesian Networks

Joyee Chen (Berkeley) and Tejas Kamtam (UCLA)

HTTPS://ARBITAL.COM/P/ONTOLOGY_IDENTIFICATION/

Content

1 | THE PROBLEM

2 | ENVIRONMENT

3 | AGENT

4 | COMPLEXITIES

The Problem: Alignment and simulating complex behavior

- **WHAT DO WE WANT?**

  To **solve alignment** -- create an AGI that will maximize the amount of diamond in the universe, given infinite compute.

- **DIAMONDS AREN'T COMPLICATED. WHAT STANDS IN OUR WAY?**

  The issue is analogous to the human values problem: simulation of the environment is difficult and complex.

  How do we simulate the universe? At **what abstraction level**? Atomic? Nuclear? Quantum? What if this turns out to be wrong? How do we know we are maximizing diamonds?

  Our issue is, therefore, **undefined behavior** when our AGI enters the real world: this is the “alignment problem”.

Environment and Simulations

ISSUE WITH PHYSICS

Computational representations of our universe fall apart when applied in the real world. If we simulate atoms, what happens when our AGI encounters quarks?

ISSUE WITH SENSORS

Defining a property of diamonds can result in an AGI optimizing for reward functions that are generally misaligned e.g. **Goal Misgeneralization**

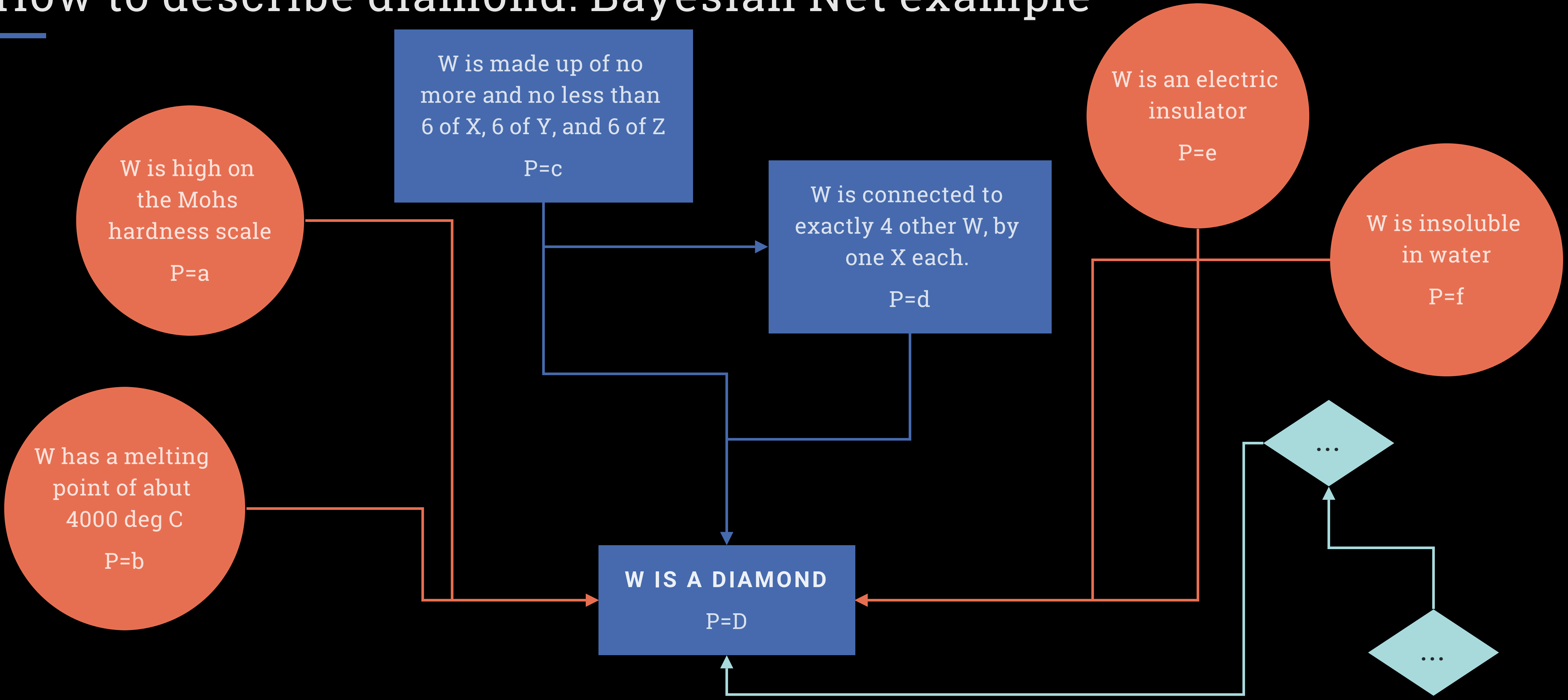
DESCRIPTIVE CONSTRAINTS

A possible solution is a simulation based on descriptive relationships of objects in place of computational representations of the universe

DYNAMIC BAYESIAN NETWORK

A DBN provides a dependency DAG which can be understood by AIXI and can be given probabilities based on a **distribution of future DBN descriptors** of the universe

How to describe diamond: Bayesian Net example



W is made up of no more and no less than 6 of X, 6 of Y, and 6 of Z

P=c

W is high on the Mohs hardness scale

P=a

W is connected to exactly 4 other W, by one X each.

P=d

W is an electric insulator

P=e

W is insoluble in water

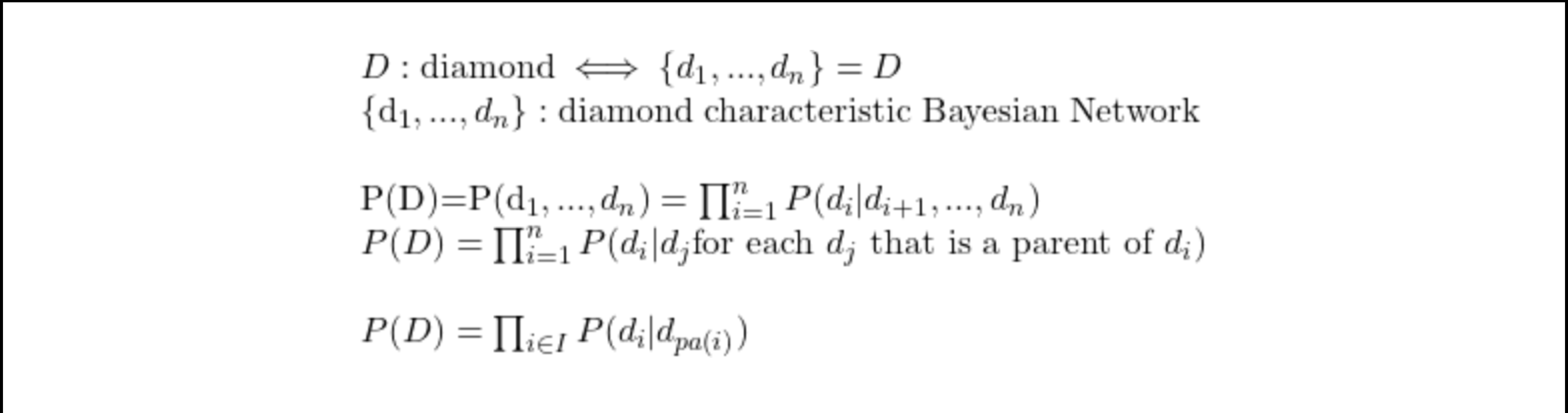
P=f

W has a melting point of abut 4000 deg C

P=b

W IS A DIAMOND

P=D



D : diamond $\iff \{d_1, ..., d_n\} = D$
$\{\mathrm{d}_1, ..., d_n\}$: diamond characteristic Bayesian Network

$\mathrm{P(D)=P(d}_1, ..., d_n) = \prod_{i=1}^{n} P(d_i|d_{i+1}, ..., d_n)$
$P(D) = \prod_{i=1}^{n} P(d_i|d_j\text{for each } d_j \text{ that is a parent of } d_i)$

$P(D) = \prod_{i\in I} P(d_i|d_{pa(i)})$

Joint probability function of Bayesian Networks

http://www.eng.tau.ac.il/~bengal/BN.pdf

The Agent: Supervised AIXI on DBN environments

"Like **a monkey** with a keyboard **and infinite time" - Abe Lincoln**

AIXI TO MAXIMIZE DIAMONDS

AIXI is a **formalization for AGI** that incorporates **Solomonoff induction** to create probability distributions of future world states given minimum input data

PREFERENCE FRAMEWORK > UTILITY FUNCTION

A set of utility functions and **meta-utility** functions that **follow VNM utility** i.e. maximize for diamonds regardless of reward function (corrigibility analysis)

BAYESIAN NETWORKS, SYMBOL TYPE MAPPING, AND RLHF

AIXI "passes" **outer alignment** through NLU of a descriptive explanation of diamonds it can test for (this can be self-supervised and reinforced through **RLHF**)

Complexities: Universal alignment problems and solutions

- **INNER MISALIGNMENT: MESA-OPTIMIZER**

  Though the AGI understands the descriptive goal of maximizing diamonds, it is plausible that child models (mesa-optimizers) are misaligned. How can we permeate alignment through multi-level models and optimizers?

- **DECEPTIVE MISALIGNMENT: "ELK"**

  ELK suggests our model can exhibit "latent misalignment" through anomalous behavior in a mostly reasonable proposition. How can we guarantee the AGI successfully elicits latent knowledge of deception?

However, DBNs act on time-steps and are inherently **myopic**. Should AIXI be run on a length-bound iteration (AIXI-tl), deceptive misalignment could be avoided.