Declarative Ethics for AI Systems

John Hooker
Carnegie Mellon University

LPOP 2024

4th Workshop on Logic and Practice of Programming

Two Basic Issues

- **Value alignment**
  - Incorporate **ethics** into machine learning.
- **Group parity**
  - Treat demographic groups **equally** in AI-based decision making.

- **Traditional approaches**
  - Remove bias and unethical examples from **training data**.
  - Design **algorithms** that avoid unethical outcomes during the learning process
- **Proposed approach**
  - **Declarative** implementation of ethics and fairness.

Declarative Implementation

- **Value alignment**
    - Express ethical principles declaratively in **formulas** of **modal logic**.
    - **Instantiate** these formulas to evaluate production rules.

- **Group parity**
    - View from broader perspective of **distributive justice**.
    - Use a **social welfare function** to express fairness/accuracy trade-off declaratively
    - **Train** an ML system by **maximizing** the social welfare function rather than minimizing loss.

Value Alignment

- **Original conception**
  - Learn human values **empirically** from crowd sourcing.
  - AI system may learn **unethical** behavior.
    - **Microsoft *Tay*, MIT Media Lab's *Moral Machine***
  - We need independently justifiable ethical principles.

Value Alignment

- **Original conception**
    - Learn human values **empirically** from crowd sourcing.
    - AI system may learn **unethical** behavior.
        - **Microsoft *Tay*, MIT Media Lab's *Moral Machine***
    - We need independently justifiable ethical principles.

- **Declarative encoding of ethical principles**
    - Deontological ethics provides basis for formulating **precise ethical principles** in quantified modal logic.
        - **Generalization and autonomy principles.**
    - Assume AI system is based on **production rules** that direct action.
        - **Newell & Anderson (1993, 1994).**
    - Each principle generates a **test proposition** for a production rule.
        - **Rule is ethical if test proposition is empirically true.**

Production rules

- **Conditional form**
  - Example:

$$\begin{aligned} C_1(a) &= \text{There is an emergency patient in ambulance } a. \\ C_2(a) &= \text{Siren and lights allow ambulance } a \text{ to drive faster.} \\ A(a) &= \text{Ambulance } a \text{ will use siren and lights.} \end{aligned}$$

  - **Ethical** production rule*: $\big(C_1(a) \wedge C_2(a)\big) \Rightarrow_a A(a)$
  - **Unethical** production rule: $\big(\neg C_1(a) \wedge C_2(a)\big) \Rightarrow_a A(a)$
    - **This is unethical because reasons C_1 and C_2 would not jointly apply if the rule were universally applied.**
    - **That is, the rule is not generalizable.**

*Here, $\Rightarrow_a$ is not logical entailment but indicates that agent a regards $C_1(a)$ and $C_2(a)$ as justifying $A(a)$.

Generalization Principle

- Use modal operators* and a possibility predicate to formulate the principle.

$$\Box_a S = \text{agent } a \text{ must assent to } S \text{ to be rational}$$

$$\Diamond_a S = \text{agent } a \text{ can be rational in assenting to } S$$

$$P(S) = S \text{ is possible}$$

- Given a production rule $C(a) \Rightarrow_a A(a)$, the principle is

$$\Diamond_a P\Big(\forall x\big(C(x) \to A(x)\big) \wedge C(a) \wedge A(a)\Big)$$

  - Agent *a* can rationally believe that it is possible to take action *A* when reasons *C* apply and when all agents to whom reasons *C* apply take action *A*.

JH and Kim 2018

*We don't have $\Box S \to S$ but do have $\Box\neg S \equiv \neg\Diamond S$.

Generalization Principle

- **Now apply this to the unethical production rule***

$$\big(\neg C_1(a) \land C_2(a)\big) \Rightarrow_a A(a)$$

by instantiating $C(a)$ as $\neg C_1(a) \land C_2(a)$:

$$\Diamond_a P\Big(\forall x\big((C_1(x) \land C_2(x)) \rightarrow A(x)\big) \land C_1(a) \land C_2(a) \land A(a)\Big)$$

  - **One can rationally believe that it is possible to drive faster with siren and lights in an ambulance that contains no emergency patient when all ambulances with no emergency patient drive with siren and lights.**

- This **test proposition** is empirically false, which means the production rule is **ungeneralizable** and therefore unethical.
- ML can be used to check the truth status of test propositions.

Kim, JH & Donaldson 2021

*$\neg C_1$ and C_2 must be the most general conditions under which the action A is performed. This can be checked by examining all rules in the system.

Autonomy Principle

- **A production rule $C(a) \Rightarrow_a A(a)$ that is inconsistent with another agent's ethical production rule $C'(b) \Rightarrow_b A'(b)$ is unethical.**
  - We have inconsistency when

$$\Box_a \neg P\big(A(a) \wedge A'(b)\big) \wedge \Diamond_a P\big(C(a) \wedge C'(b)\big)$$

  - **One cannot rationally believe that actions *A* and *A′* are mutually compatible and can rationally believe that conditions *C* and *C′* are compatible.**

- A **test proposition** is obtained and empirically assessed for each production rule.

JH & Kim 2018

Questions

- **Should we use nonmonotonic/default logic to account for defeasible inference?**
    - The aim is not to model the reasoning **process**, but to encode the **outcome** of that process.

Questions

- **Should we use nonmonotonic/default logic to account for defeasible inference?**
  - The aim is not to model the reasoning **process**, but to encode the **outcome** of that process.

- **But isn’t defeasible inference necessary to account for the many exceptions to general ethical principles?**
  - Valid ethical principles do not have “exceptions” but are highly **context-sensitive**.
    - **They consider the antecedent of the production rule, which indicates the context.**

Group Parity

- **The most widely discussed issue in AI ethics.**
  - Goal: **equal** treatment for demographic **groups**
  - **Selection rates** are compared for:
    - **Job interviews**
    - **University admissions**
    - **Mortgage loans, etc.**

- A "**protected group**" is compared with the **rest** of the population
  - Groups defined by **race**, **gender**, **ethnicity**, **class**, **region**, etc.
  - Sometimes based on **legal** mandates

Problems with Group Parity Metrics

- **Failure to account for actual welfare consequences**
    - Considers only **frequency** of selection
    - For example, rejection may be **more harmful** to a protected group

Problems with Group Parity Metrics

- **Failure to account for actual welfare consequences**
  - Considers only **frequency** of selection
  - For example, rejection may be **more harmful** to a protected group
- **Controversy over which metric is appropriate**
  - **Many statistical metrics** have been proposed
    - **Demographic parity, equalized odds, predictive rate parity, etc.**
  - Some are mutually **incompatible**

Problems with Group Parity Metrics

- **Failure to account for actual welfare consequences**
  - Considers only **frequency** of selection
  - For example, rejection may be **more harmful** to a protected group
- **Controversy over which metric is appropriate**
  - **Many statistical metrics** have been proposed
    - **Demographic parity, equalized odds, predictive rate parity, etc.**
  - Some are mutually **incompatible**
- **Unclear how to identify protected groups**
  - Groups often have **conflicting interests**
  - **No limit** to groups that may cry “unfair.”

Fairness as Social Welfare

- **Group fairness through population-wide social welfare**
  - A **broader concept of distributive justice** can assess parity metrics and achieve fairness across multiple groups
    - **while taking *welfare* into account.**

- **Assess fairness with a social welfare function**
  - Let $\boldsymbol{u} = (u_1, \ldots, u_n)$ be **utilities** distributed to stakeholders 1, …, n
  - Utility = some kind of **benefit**
    - **Wealth, negative cost, resources, health, etc.**
  - A **social welfare function** $W(\boldsymbol{u})$ measures the desirability of $\boldsymbol{u}$
    - **taking into account overall utility as well as how it is distributed.**

Fairness as Social Welfare

- **Proposal**
  - When training an AI system, **maximize a social welfare function** rather than minimize loss subject to parity constraints
    - **A selection decision is associated with a utility outcome for each stakeholder.**
  - Fairness is **declaratively** specified in the social welfare function, independently of the algorithm used to train the NN.

Fairness as Social Welfare

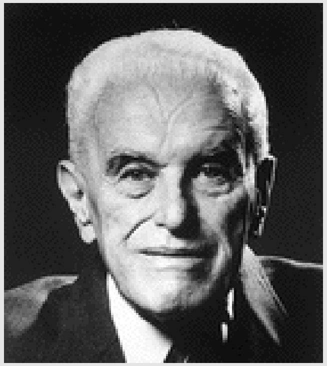
- **Proposal**
  - When training an AI system, **maximize a social welfare function** rather than minimize loss subject to parity constraints
    - **A selection decision is associated with a utility outcome for each stakeholder.**
  - Fairness is **declaratively** specified in the social welfare function, independently of the algorithm used to train the NN.

- **Social welfare metrics that balance fairness & utility maximization**
  - Alpha fairness
    - **Special case: proportional fairness (Nash bargaining solution)**
  - Rawlsian criteria
    - **Maximin, leximax, beta fairness**
  - Kalai-Smorodinsky bargaining solution
  - Threshold functions

Alpha fairness

- Focus on **alpha fairness** as a social welfare function
  - **Frequently used** in engineering, etc.
  - Various forms studied for over 70 years.
    - **In particular, by 2 Nobel laureates (John Nash, J.C. Harsanyi).**
  - Defended by axiomatic and bargaining arguments
    - ***Axiomatic arguments:*** **Nash (1950), Lan, Kao & Chiang (2010,2011)**
    - ***Bargaining arguments:*** **Harsanyi (1977), Rubinstein (1982), Binmore, Rubinstein & Wolinksy (1986)**

John Nash



J. C. Harsanyi

Alpha Fairness

- **The alpha fairness social welfare function:**

$$W_\alpha(\boldsymbol{u}) = \begin{cases} \dfrac{1}{1-\alpha}\displaystyle\sum_i u_i^{1-\alpha} & \text{for } \alpha \geq 0,\ \alpha \neq 1 \\ \displaystyle\sum_i \log(u_i) & \text{for } \alpha = 1 \end{cases}$$

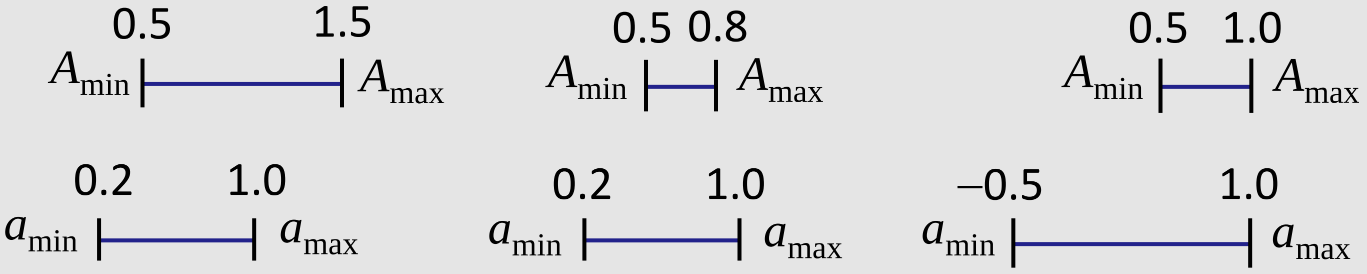
where u_i is the utility allocated to individual i

  - **Larger α** implies **more fairness**.
  - **Utilitarian** when α = 0, **maximin** (Rawlsian) when $\alpha \to \infty$
  - **Proportional fairness** (Nash bargaining solution) when α = 1
  - **α < 1** incentivizes **competition**, **α > 1** incentivizes **cooperation**

- **To achieve alpha fairness:** Maximize $W_\alpha(\boldsymbol{u})$

Alpha Fairness

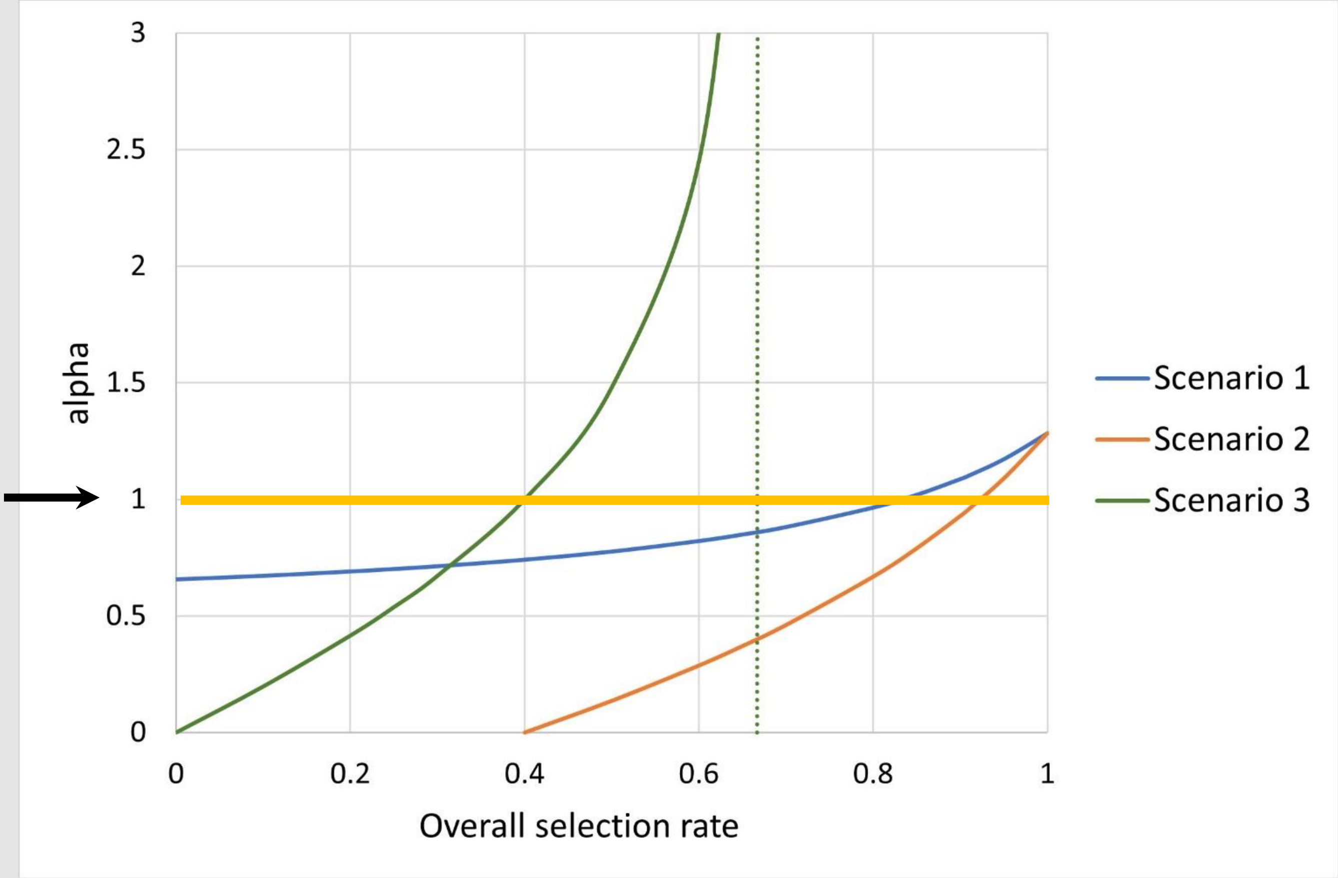
- **Example:** 3 scenarios with a different range of selection benefits (utilities) for the majority and protected group.



	Scenario 1	Scenario 2	Scenario 3
Majority group	$A_{\min}$ 0.5 – 1.5 $A_{\max}$	$A_{\min}$ 0.5 – 0.8 $A_{\max}$	$A_{\min}$ 0.5 – 1.0 $A_{\max}$
Protected group	$a_{\min}$ 0.2 – 1.0 $a_{\max}$	$a_{\min}$ 0.2 – 1.0 $a_{\max}$	$a_{\min}$ –0.5 – 1.0 $a_{\max}$
	Granting *job interviews*	***Admission* of talented individuals to *university***	***Mortgage loans* with possible *foreclosure***

Alpha Fairness

Proportional fairness



Chen, JH & Leben 2024

- Alpha values that **achieve demographic parity**.
- Parity generally corresponds to **less than proportional fairness**.

Conclusions

- **Value alignment**
  - **Deontological ethics** + **logical formalism** can incorporate ethical principles into AI systems declaratively.

- **Group parity**
  - Declarative implementation of distributive justice in a **social welfare function** can achieve welfare-sensitive parity for multiple groups simultaneously.

Questions or comments?