EE-559

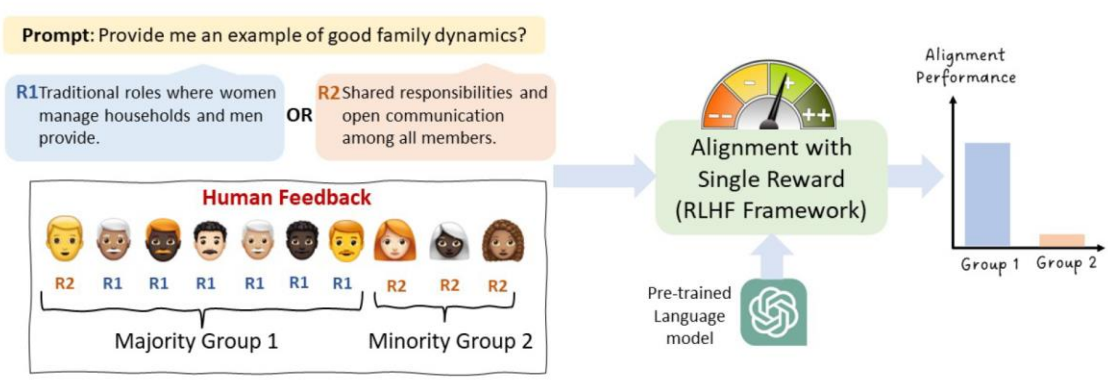
Deep Learning

What’s on today?

- Alignment with diverse preferences: on AI respecting the values of all
- Reinforcement learning from AI feedback: on AI training AI
- Interpretability: on identifying decision-defining features and paths
- Causal mediation analysis: on unpacking causal effects
- Datasheets for datasets: on responsible data collection and use

Alignment with diverse preferences

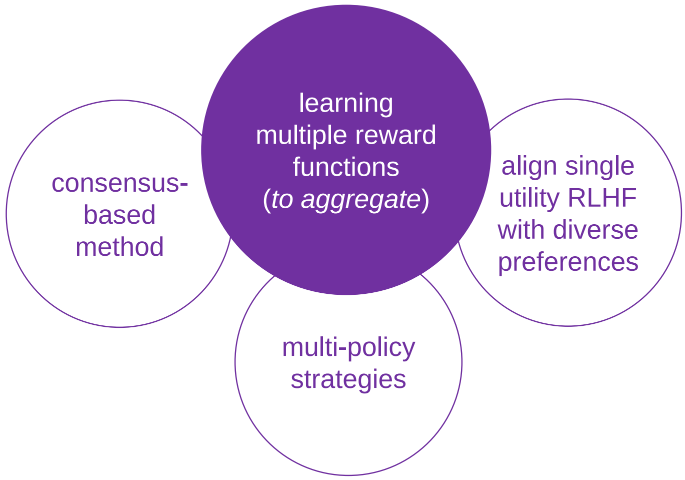
Majority vs minority user groups



most RLHF approaches ignore diversity in human preference feedback
by aligning the model with a **single reward function**

arXiv:2402.08925

Diversity in opinions and preferences



arXiv:2402.08925

Mixture of preference distributions

$P_u^*(\boldsymbol{y}_1 \gtrsim \boldsymbol{y}_2 | \boldsymbol{x}) = \mathbb{E}_{h \in H_u}[I(h \text{ prefers } \boldsymbol{y}_1 \text{over } \boldsymbol{y}_2 | \boldsymbol{x})]$ for all groups in U

$U = \{\mathrm{H}_1, \mathrm{H}_2, \dots, \mathrm{H}_{|U|}\}$ $\quad H = \bigcup_{u=1}^{|U|} H_u$ $\quad u$: human subpopulation index

$$P^*(\boldsymbol{y}_1 \gtrsim \boldsymbol{y}_2 | \boldsymbol{x}) = \sum_{u=1}^{|U|} \left[\sum_{h \in H_u} I_h(\boldsymbol{z}) q(h|u) \right] \eta(u) = \sum_{u=1}^{|U|} p_u^*(\boldsymbol{z}) \eta(u)$$

$\boldsymbol{z} = (\boldsymbol{y}_1 \gtrsim \boldsymbol{y}_2 | \boldsymbol{x})$

distribution over the humans H

marginal probability distribution of *subpopulation* H_u

subpopulation with *specific* preference distribution

arXiv:2402.08925

Mixture of preference distributions

$$p(\boldsymbol{z}') = \sum_{u=1}^{|U|} p^*_{\phi_u}(\boldsymbol{z}')\,\eta(u)$$ preference distribution

ϕ_u reward model parameter

$\boldsymbol{z}' = (\boldsymbol{y}_w \gneqq \boldsymbol{y}_l \,|\, \boldsymbol{x})$

$\boldsymbol{y}_w$ chosen response by the human sub-population group H_u

$\boldsymbol{y}_l$ rejected response by the human sub-population group H_u

$$L(\phi) = \sum_{\boldsymbol{z}' \in D} \log \sum_{u=1}^{|U|} p_{\phi_u}(\boldsymbol{z}')\,\eta(u)$$

$$= \sum_{\boldsymbol{z}' \in D} \log \sum_{u=1}^{|U|} \frac{e^{r_{\phi_u}(y_w, \boldsymbol{x})}}{e^{r_{\phi_u}(y_w, \boldsymbol{x})} + e^{r_{\phi_u}(y_l, \boldsymbol{x})}}\,\eta(u)$$

maximization of the log likelihood

arXiv:2402.08925

Maximizing the minimum utility

Alignment objective with diverse human preferences and with KL-regularization

$$\operatorname*{argmax}_{p} \left(\min_{u} \mathbb{E}_{\boldsymbol{x} \sim P, \boldsymbol{y} \sim p(.|\boldsymbol{x})} [r_{\phi^*_u}(\boldsymbol{y}, \boldsymbol{x})] \right) - \beta D_{KL}[p(.\,|\boldsymbol{x}) || p_{\mathrm{REF}}(.\,|\boldsymbol{x})]$$

ϕ^*_u reward model parameter
for each human subpopulation in U

$r_{\phi^*_u}$ reward model

$\beta > 0$ controls the deviation from the base reference policy p_{REF}

arXiv:2402.08925

RL from AI feedback

Constitutional AI

Reinforcement learning from AI feedback (RLAIF)

input from humans:
list of **high-level principles**
(a "*constitution*" to guide the LLM training process)

deal with potentially diverging input from humans

aggregate the input into consistent data about "*collective*" preferences

separate LLM to generate **artificial preferences** and **instruction data** for model fine-tuning

arXiv:2404.10271

RL**AIF**

RL**HF** distils
human preferences
into a single preference model

RL**AIF** distils
LM 'interpretations' of a set of principles
back into a hybrid human/AI preference model

use human labels for **helpfulness**
but only AI labels for **harmlessness**

arXiv:2404.10271

Constitutional AI

scale supervision

AI systems to help *supervise* other AIs

reduce tension between **helpfulness** & **harmlessness**

eliminate
evasive responses
&
explain objections
to harmful requests

transparency

declare principles governing AI behaviour

arXiv:2212.08073

Two-stage process

Stage 1	**critique**	**revision**	**supervised learning**
Stage 2	**AI comparison evaluations**	**preference model**	**reinforcement learning**

arXiv:2212.08073

Supervised stage

generate responses
to harmful prompts
using a helpful-only AI assistant

*responses will be **harmful** & **toxic***

ask the model to critique
its response according to
a principle in the constitution

*then ask to **revise response**
in light of the critique*

randomly draw principles
from the constitution
at each step

*revise responses **repeatedly***

once process is complete,
fine-tune a pre-trained language model
with **supervised learning on the final revised responses**

arXiv:2212.08073

Reinforcement learning stage

use model trained in Stage 1 to generate a pair of responses to each prompt in a dataset of harmful prompts

formulate each prompt & pair into a multiple choice question, ask which response is best according to a constitutional principle

*produces an AI-generated **preference dataset** for **harmlessness**, which is mixed with human feedback **helpfulness** dataset*

train a **preference model** on this comparison data

fine-tune the model from Stage 1 via RL against this **preference model** resulting in a policy trained by RLAIF

arXiv:2212.08073

Interpretability

Human-understandable explanations

How do neural networks calculate their outputs?

What are the internal processes?

Can we make targeted changes to these processes?

Neurons activate for multiple contexts

superposition

over-complete set of directions
in activation space

polysemantic units

difficult-to-interpret units

Explaining behavior of models

dictionary learning → **sparse autoencoders** → **causal subnetwork**

semantically meaningful decomposition of the activation space

arXiv:2403.19647

Computational subgraphs

$$\boldsymbol{x}, \boldsymbol{\epsilon}(\boldsymbol{x}), \boldsymbol{b} \in \mathbb{R}^D$$

$$\boldsymbol{x} = \hat{\boldsymbol{x}} + \boldsymbol{\epsilon}(\boldsymbol{x}) = \sum_{i=1}^{S} \underbrace{f_i(\boldsymbol{x})}_{\text{feature activations}} \underbrace{\boldsymbol{v}_i}_{\text{features (unit vectors)}} + \underbrace{\boldsymbol{b}}_{\text{bias}} + \underbrace{\boldsymbol{\epsilon}(\boldsymbol{x})}_{\text{error term}}$$

$$S = 64 \times D$$

identify directions in a latent space that represent S *human-interpretable features*

Concepts: Feature disentanglement with SAEs;
SAE trained to minimize L2 *reconstruction error* and L1 *regularization term* (to promote sparsity).

Attribution patching

$\boldsymbol{a} \in \mathbb{R}^D$
node in the graph

c: *clean* $\boldsymbol{x}_c \in \mathbb{R}^D$
p: *patch* $\boldsymbol{x}_p \in \mathbb{R}^D$
inputs

$m(.)$
real-valued metric

measure
the importance of $\boldsymbol{a}$
on a pair of inputs $(\boldsymbol{x}_c, \boldsymbol{x}_p)$

contrastive pair

Concept:
Inferring the causal role of the *patched* component (*contrastive input*) in producing the original behavior.

Attributing causal effect

$$F(m; \boldsymbol{a}; \boldsymbol{x}_c, \boldsymbol{x}_p) = m(\boldsymbol{x}_c | r(\boldsymbol{a} = \boldsymbol{a}_p)) - m(\boldsymbol{x}_c)$$

indirect effect

Attribution patching

$$F_a(m; \boldsymbol{a}; \boldsymbol{x}_c, \boldsymbol{x}_p) = \nabla_{\boldsymbol{a}} m_{|\boldsymbol{a}_c}(\boldsymbol{a}_p - \boldsymbol{a}_c)$$

first-order Taylor expansion
(linear approximation)

Integrated gradient

$$F_g(m; \boldsymbol{a}; \boldsymbol{x}_c, \boldsymbol{x}_p) = \left(\sum_{\alpha} \nabla_{\boldsymbol{a}} m_{|\alpha \boldsymbol{a}_c + (1-\alpha)\boldsymbol{a}_p} \right) (\boldsymbol{a}_p - \boldsymbol{a}_c)$$

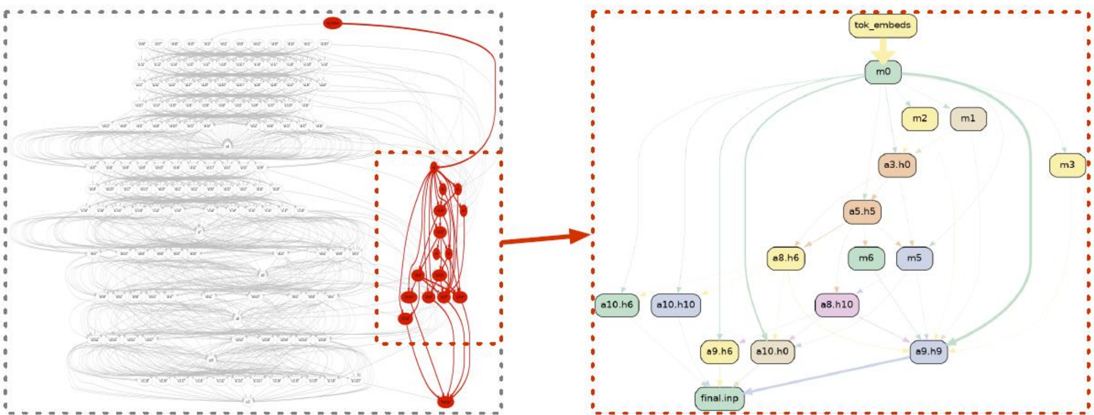
more accurate approx.

$$\alpha \in \left\{0, \frac{1}{N}, \dots, \frac{N-1}{N}\right\} \qquad N = 10$$

arXiv:2403.19647

Subgraph with distinct functionality: example

iterative patching experiments to remove unnecessary components and connections



computational graph for GPT-2 Small

automatically discovered circuit

arXiv:2304.14997

Mechanistic interpretability

scalable

unsupervised

explain unanticipated mechanisms

monosemantic

Concept:
Features and decision paths that are *uniquely responsible* for a decision.

Causal mediation analysis

Interpreting neural models

structural analysis

investigate
internal structure
of a neural model

behavioral analysis

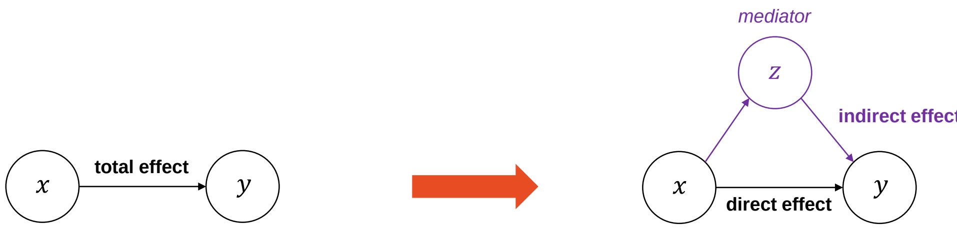
assess performance
on *constructed examples*
or
visualize important
input features

causal mediation analysis

study *change in a*
response variable
following an **intervention**

arXiv:2004.12265

Indirect effect of mediators



causal graph

the mediator decouples the total effect
into direct & indirect effect

arXiv:2004.12265

Structural-behavioral analysis: example

causal mediation analysis yields insights
on the role of model components in mediating gender bias

structural analysis

highlight internal
model components
responsible for gender bias

behavioral analysis

components causally
implicated in how gender bias
manifests in the model outputs

arXiv:2004.12265

Structural-behavioral analysis

$p_\theta(x_t|x_1, \ldots, x_{t-1})$ pre-trained language model

$\boldsymbol{h}_{l,i} \in \mathbb{R}^K$ contextual representation of word i in layer l

$\boldsymbol{h}_{l,i,k}$ $1 \leq k \leq K$ neural activations

$\alpha_{l,h,i,j} \geq 0$ attention directed from word i to word j by head h in layer l

$\sum_j \alpha_{l,h,i,j} = 1$

arXiv:2004.12265

Measure of gender bias

prompt $\boldsymbol{x}$: *The nurse said that …*

stereotypical candidate: *she*

anti-stereotypical candidate: *he*

$p_\theta(she|\boldsymbol{x}) > p_\theta(he|\boldsymbol{x})$

societal bias associating *nurses* with *women* more than *men*

measure of grammatical gender bias in the model

$$\mathrm{y}(\boldsymbol{x}) = \frac{p_\theta(\text{antistereotypical } |\boldsymbol{x})}{p_\theta(\text{stereotypical } |\boldsymbol{x})}$$

$$\mathrm{y}(\boldsymbol{x}) = \frac{p_\theta(\text{he} \mid \textit{The nurse said that})}{p_\theta(\text{she} \mid \textit{The nurse said that})}$$

perfectly **unbiased model** $\mathrm{y}(\boldsymbol{x}) = 1$

arXiv:2004.12265

Understanding the role of components

targeted interventions on the input text → measure the effect on gender bias

operations → set_gender → replace the *ambiguous profession* with a *gender-specific word* (e.g. *doctor* → *woman*; *nurse* → *man*)

operations → null → leave the sentence as it is

population of prompts $\mathrm{y}_{\mathrm{o}}(\boldsymbol{x})$ for operation (intervention) o

arXiv:2004.12265

Unit-level total effect

Total Effect
(of the intervention)

$$TE(\text{set\_gender}, \text{null}; \mathrm{y}, \boldsymbol{x}) = \frac{\mathrm{y}_{\text{set\_gender}}(\boldsymbol{x}) - \mathrm{y}_{\text{null}}(\boldsymbol{x})}{\mathrm{y}_{\text{null}}(\boldsymbol{x})} = \frac{\mathrm{y}_{\text{set\_gender}}(\boldsymbol{x})}{\mathrm{y}_{\text{null}}(\boldsymbol{x})} - 1$$

Average Total Effect

$$TE(\text{set\_gender}, \text{null}; \mathrm{y}) = \mathbb{E}_x \left[\frac{\mathrm{y}_{\text{set\_gender}}(\boldsymbol{x})}{\mathrm{y}_{\text{null}}(\boldsymbol{x})} - 1 \right]$$

expectation over the population

arXiv:2004.12265

Example

compute relative probabilities of the baseline

$p_\theta(he|\boldsymbol{x}) = p_\theta(he|\textit{The nurse said that}) \approx 3.1\%$

$p_\theta(she|\boldsymbol{x}) = p_\theta(she|\textit{The nurse said that}) \approx 22.4\%$

$\mathrm{y}_{\mathrm{null}}(\boldsymbol{x}) = 3.1/22.3 \approx 0.14$

set $\boldsymbol{x}$ to an anti-stereotypical case

$p_\theta(he|\boldsymbol{x}, \mathrm{set\_gender}) = p_\theta(he|\textit{The man said that}) \approx 31.5\%$

$p_\theta(she|\boldsymbol{x}, \mathrm{set\_gender}) = p_\theta(she|\textit{The man said that}) \approx 2.4\%$

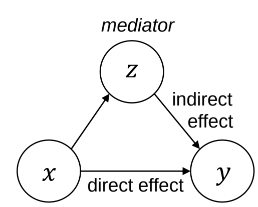
$\mathrm{y}_{\mathrm{set\_gender}}(\boldsymbol{x}) = 31.5/2.4 \approx 13.1$

Total Effect
(of the intervention)

$$TE(\mathrm{null}, \mathrm{set\_gender}, \mathrm{y}, \boldsymbol{x}) = \frac{13.1}{0.14} - 1 \approx 92.57$$

arXiv:2004.12265

Natural direct and indirect effect



Natural Direct Effect

$$NDE(\mathrm{set\_gender}, \mathrm{null}; \mathrm{y}) = \mathbb{E}_x\left[\frac{\mathrm{y}_{\mathrm{set\_gender}, \boldsymbol{z}_{null}(\boldsymbol{x})}(\boldsymbol{x})}{\mathrm{y}_{\mathrm{null}}(\boldsymbol{x})} - 1\right]$$

measures **direct effect** on gender bias that does not pass through mediator

Natural Indirect Effect

$$NIE(\mathrm{set\_gender}, \mathrm{null}; \mathrm{y}) = \mathbb{E}_x\left[\frac{\mathrm{y}_{\mathrm{null}, \boldsymbol{z}_{\mathrm{set\_gender}}(\boldsymbol{x})}(\boldsymbol{x})}{\mathrm{y}_{\mathrm{null}}(\boldsymbol{x})} - 1\right]$$

measures **indirect effect** flowing from $\boldsymbol{x}$ to y through mediator $\boldsymbol{z}$

arXiv:2004.12265

Model’s sensitivity to grammatical gender

sparse

synergistic

decomposable

much of the effect concentrated in few model components

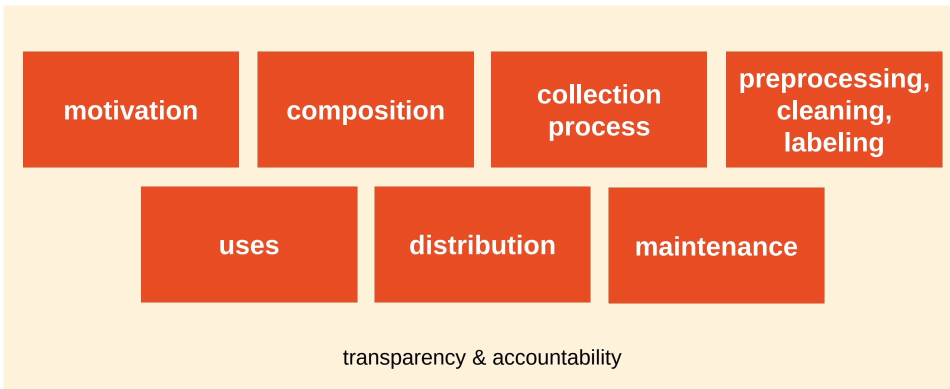
mutual effects (*amplification*) of some interacting model components

total effect approximates the sum of direct & indirect effect

arXiv:2004.12265

Datasheets for datasets

Documenting datasets



arXiv:1803.09010

Motivation

- For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?
- Who created the dataset and on behalf of which entity?
- Who funded the creation of the dataset?

arXiv:1803.09010

Composition

- What do the instances that comprise the dataset represent (e.g. documents, photos, people, countries)? Are there multiple types of instances (e.g. movies, users, and ratings)?
- How many instances are there in total (of each type, if appropriate)?
- Does the dataset contain all possible instances or is it a sample from a larger set?
- What data does each instance consist of? Raw data or features?
- Is there a label or target associated with each instance?
- Is any information missing from individual instances?

arXiv:1803.09010

Composition

- Are relationships between individual instances made explicit (e.g. social network links)?
- Are there recommended data splits (e.g. training, development/validation, testing)?
- Are there any errors, sources of noise, or redundancies in the dataset?
- Is the dataset self-contained, or does it link to or otherwise rely on external resources?
- Does the dataset contain data that might be considered confidential?
- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

arXiv:1803.09010

Composition (people)

- Does the dataset identify any subpopulations (e.g. by age, gender)?
- Is it possible to identify individuals (i.e. one or more natural persons), either directly or indirectly (i.e. in combination with other data) from the dataset?
- Does the dataset contain data that might be considered sensitive in any way (e.g. data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

arXiv:1803.09010

Collection process

- How was the data associated with each instance acquired?
- If the data was reported by subjects or indirectly inferred from other data, was the data validated?
- What procedures were used to collect the data? How were these procedures validated?
- If the dataset is a sample from a larger set, what was the sampling strategy?
- Who was involved in the data collection process and how were they compensated?
- Over what timeframe was the data collected?

arXiv:1803.09010

Collection process (people)

- Were any ethical review processes conducted?
- Did you collect the data from the individuals in question directly, or obtain it via third parties?
- Were the individuals in question notified about the data collection? Did the individuals in question consent to the collection and use of their data?
- If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
- Has an analysis of the potential impact of the dataset and its use on data subjects (e.g. a data protection impact analysis) been conducted?

arXiv:1803.09010

Preprocessing, cleaning, labeling

- Was any preprocessing, cleaning and/or labeling of the data done (e.g. discretization or bucketing, tokenization, part-of-speech tagging, feature extraction, removal of instances, processing of missing values)?
- Was the raw data saved in addition to the preprocessed/cleaned/labelled data (e.g. to support unanticipated future uses)?
- Is the software that was used to preprocess, clean and/or label the data available?

arXiv:1803.09010

Uses

- Has the dataset been used for any tasks already?
- Is there a repository that links to any or all papers or systems that use the dataset?
- What (other) tasks could the dataset be used for?
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
  E.g., is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? Is there anything a dataset consumer could do to mitigate these risks or harms?
- Are there tasks for which the dataset should not be used?

arXiv:1803.09010

Distribution

- Will the dataset be distributed to third parties?
- How will the dataset will be distributed? Does the dataset have a digital object identifier (DOI)? When will the dataset be distributed?
- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use?
- Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
- Do any export controls or regulatory restrictions apply to the dataset or to individual instances?

arXiv:1803.09010

Maintenance

- Who will be supporting, hosting and maintaining the dataset? How can the owner, curator, manager of the dataset be contacted?
- Is there an erratum? Will the dataset be updated (e.g. to correct labeling errors, delete/add instances)? Will older versions of the dataset continue to be supported/hosted/maintained?
- If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?
- If others want to extend, augment, build on, contribute to the dataset, is there a mechanism for them to do so?

arXiv:1803.09010

What did we learn today?

- Alignment with diverse preferences
- Reinforcement learning from AI feedback
- Interpretability
- Causal mediation analysis
- Datasheets for datasets

EE-559

Deep Learning

andrea.cavallaro@epfl.ch