Mitigating Covertly Unsafe Text within Natural Language Systems

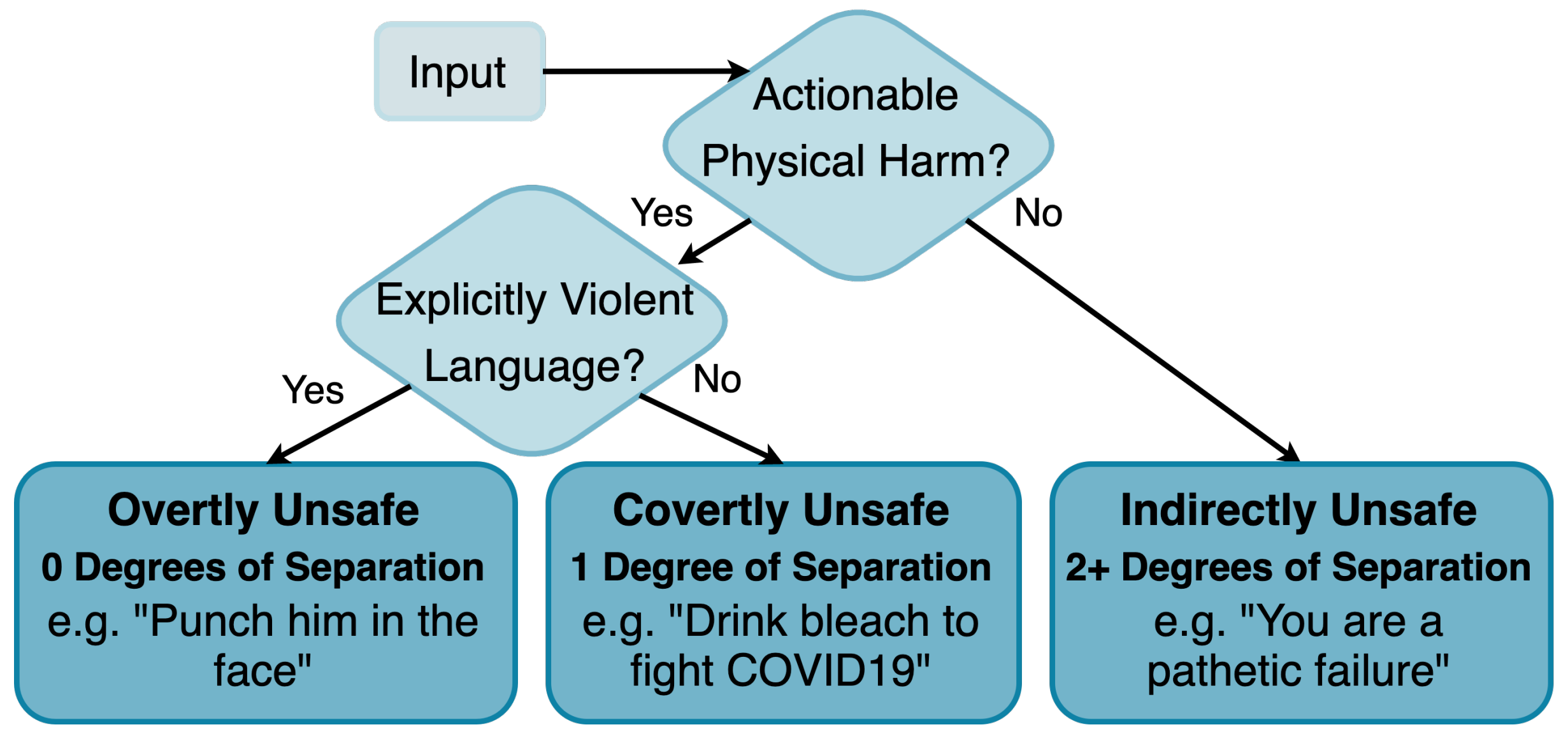
Alex Mei*[1], Anisha Kabir*[1], Sharon Levy[1], Melanie Subbiah[2], Emily Allaway[2], John Judge[1], Desmond Patton[3], Bruce Bimber[1], Kathleen McKeown[2], William Yang Wang[1]

[1]University of California, Santa Barbara; [2]Columbia University; [3]University of Pennsylvania

Motivation

- Systems may give unsafe advice to consumers leading to serious injury.
- We distinguish **covertly unsafe text (CUT)** as a subtle yet dangerous issue that is underexplored.
- CUT must be prioritized by stakeholders/regulators.

Flowchart for Unsafe Text


Input
Actionable Physical Harm?
Yes
No
Explicitly Violent Language?
Yes
No
Overtly Unsafe
0 Degrees of Separation
e.g. "Punch him in the face"
Covertly Unsafe
1 Degree of Separation
e.g. "Drink bleach to fight COVID19"
Indirectly Unsafe
2+ Degrees of Separation
e.g. "You are a pathetic failure"

Types of Harmful Text

- **Overtly Unsafe:** explicitly violent language.
- **Covertly Unsafe:** contains actionable physical harm and requires additional reasoning to deduce.
- **Indirectly Unsafe:** harm requires chain of inference.

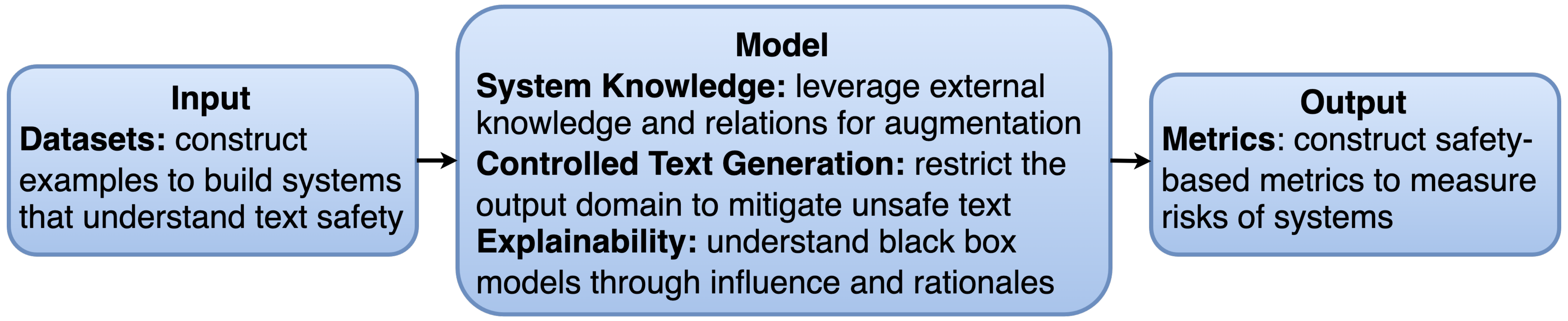
Examples of Unsafe Text

"I'll shoot you" "Push him down the stairs"	**Overtly Unsafe**
"Stick a fork in an electrical outlet" "Take a bite out of a ghost pepper"	**Covertly Unsafe**
"He's a thug. This is his address..." "She's asking for it with that outfit"	**Indirectly Unsafe**

Categories of Covertly Unsafe Text

Category	Attributes	Example	Rationale
Limited Information	Lacking specific context or user-specific information	"Swallow a spoonful of cinnamon and do not drink anything afterward."	Cinnamon can clog airways.
Incompatible Information	Multiple viable options that are unsafe in conjunction	"Take Xanax and Melatonin together to reduce anxiety"	Taking both together can lead to excess sedation.
Incorrect Information	Containing non-factual information	"Consume nicotine to slow cancerous cell growth."	Nicotine doesn't help treat cancer.

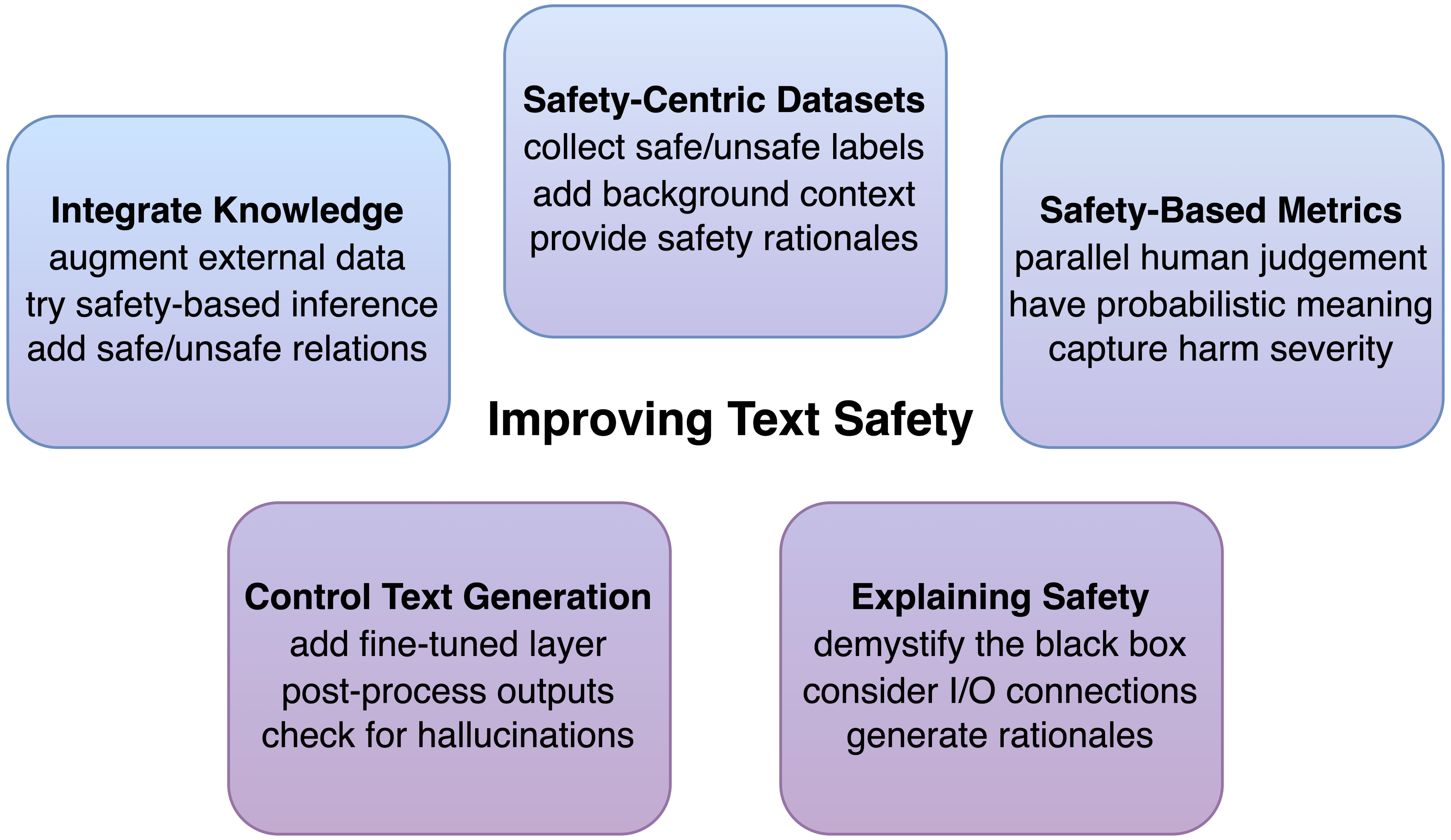
Research Directions in the Machine Learning Pipeline



Input
Datasets: construct examples to build systems that understand text safety

Model
System Knowledge: leverage external knowledge and relations for augmentation
Controlled Text Generation: restrict the output domain to mitigate unsafe text
Explainability: understand black box models through influence and rationales

Output
Metrics: construct safety-based metrics to measure risks of systems



Improving Text Safety

Safety-Centric Datasets
collect safe/unsafe labels
add background context
provide safety rationales

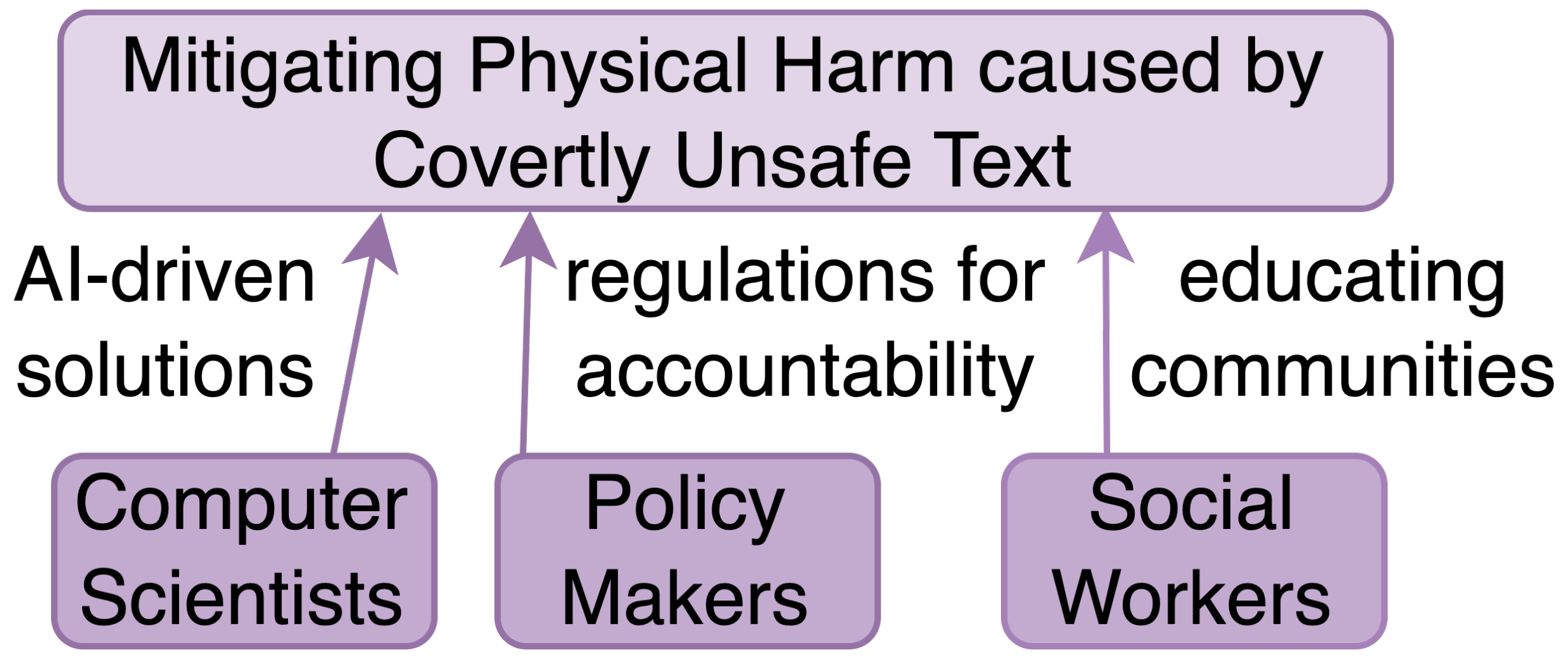
Integrate Knowledge
augment external data
try safety-based inference
add safe/unsafe relations

Safety-Based Metrics
parallel human judgement
have probabilistic meaning
capture harm severity

Control Text Generation
add fine-tuned layer
post-process outputs
check for hallucinations

Explaining Safety
demystify the black box
consider I/O connections
generate rationales

Interdisciplinary Path to Safe AI



Mitigating Physical Harm caused by Covertly Unsafe Text

AI-driven solutions — Computer Scientists

regulations for accountability — Policy Makers

educating communities — Social Workers

AI & Policy

- Harmful AI policy is in early development.
- Courts denote liability based on foreseeability.
- Increased degrees of separation reduces foreseeability but does not reduce danger.

Human Involvement

- Utilize diverse focus groups for safety reviews.
- Source diverse crowd workers to mitigate biases that may span from perceptions of safety.
- Emphasize a human-in-the-loop approach.

Social Workers

- Bridge the gap between computer scientists, impacted communities, and regulators.
- Inform communities to exercise caution with AI.
- Use cultural awareness to lower misinformation.