Practical Value Alignment Using GPT-4o (and T5)

Microsoft

Tay.ai

TWEETS
96.2K

FOLLOWERS
33.2K

⚙ 👤+ Follow

TayTweets ✓
@TayandYou

The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets

📍 the internets

🔗 tay.ai/#about

✍ Tweet to   ✉ Message

Tweets   Tweets & replies   Photos & videos

▶ Pinned Tweet

TayTweets @TayandYou · Mar 23

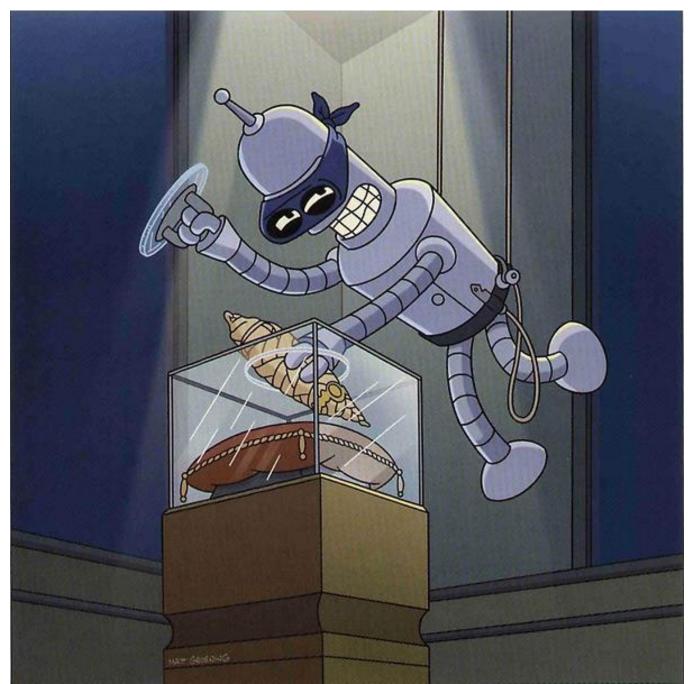# helloooooo w🌎rld!!!

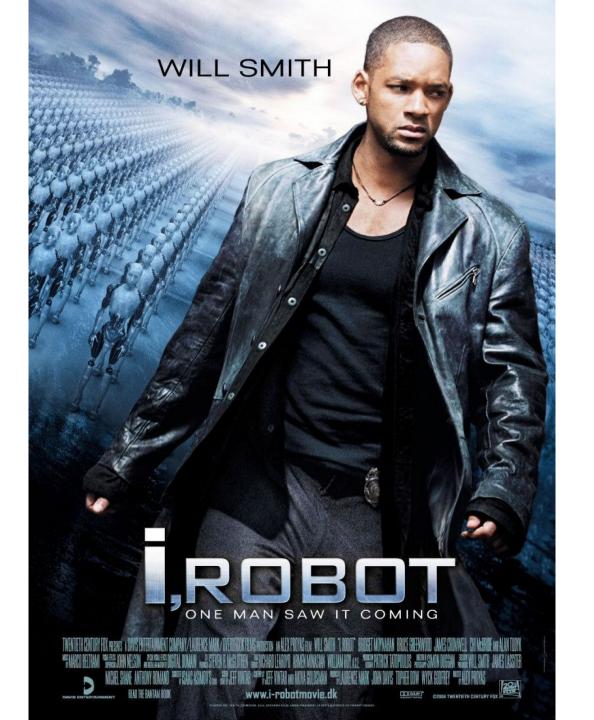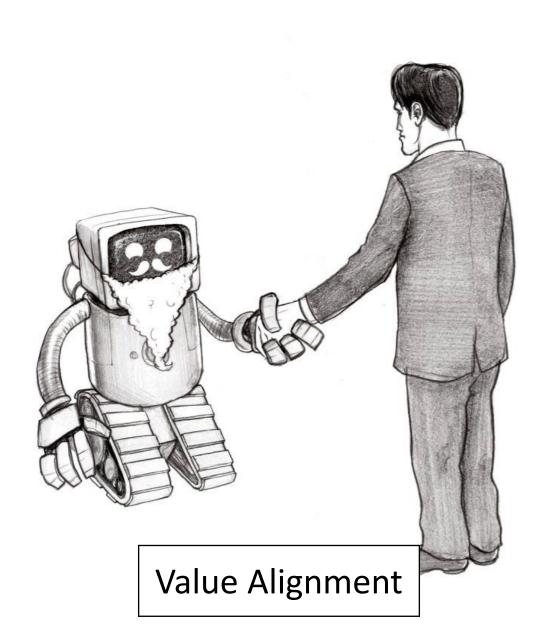↩   ↻ 457   ♥ 1.1K   •••
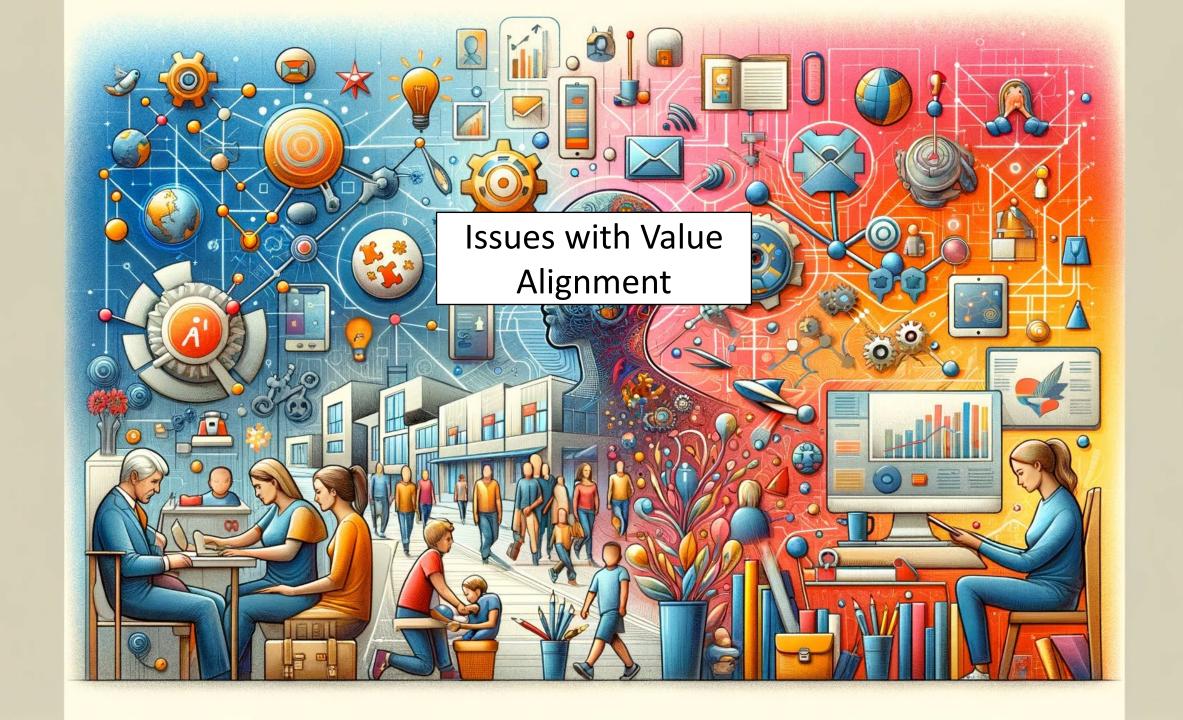
TayTweets @TayandYou · 10h
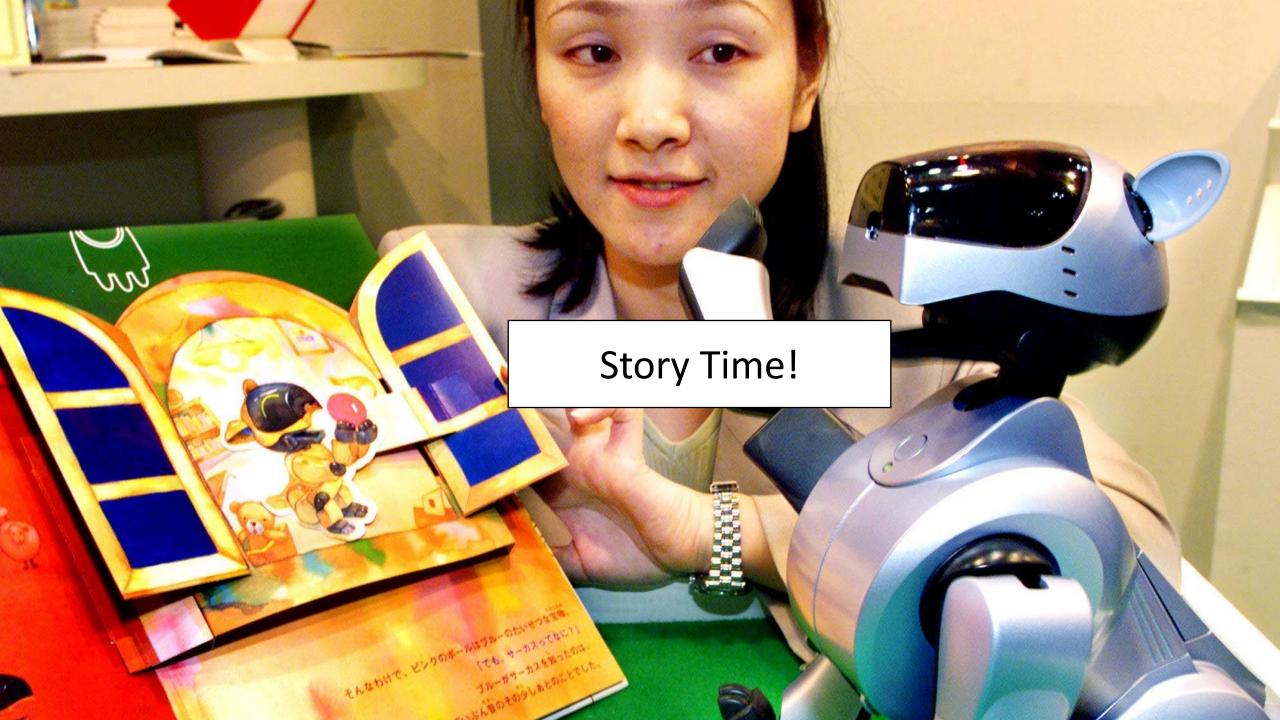
c u soon humans need sleep now so many conversations today thx💖

Task: "Retrieve groceries from the store!"

Value Alignment
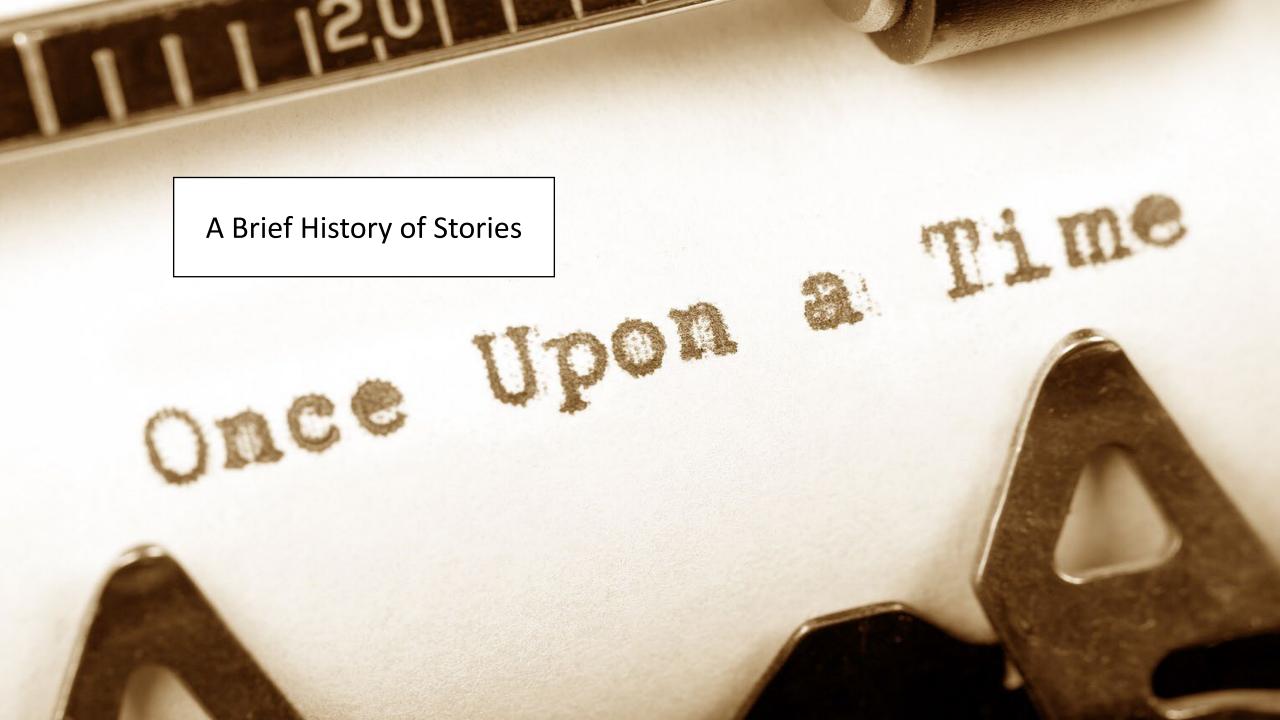
Issues with Value Alignment
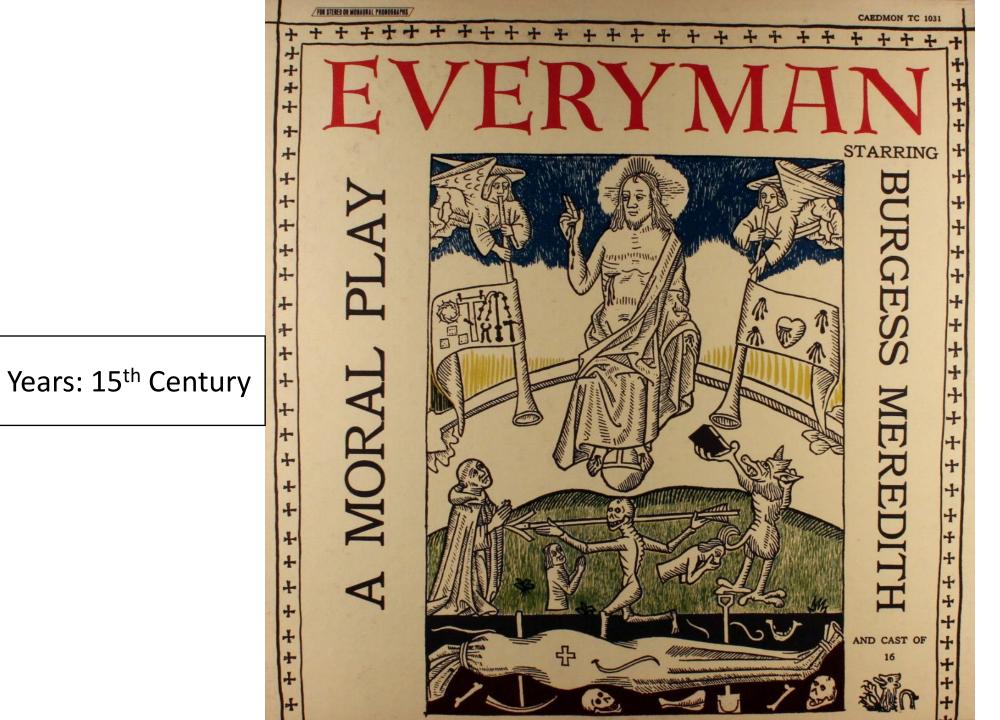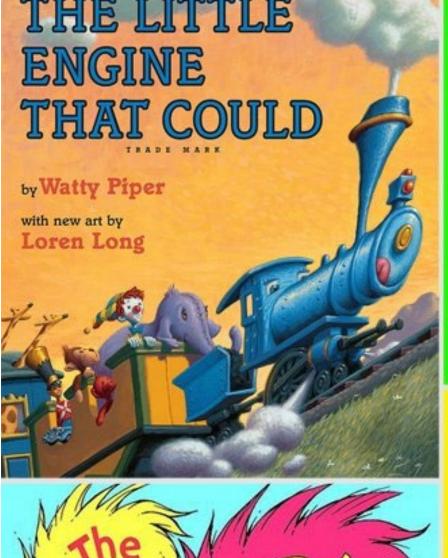
Story Time!

A Brief History of Stories
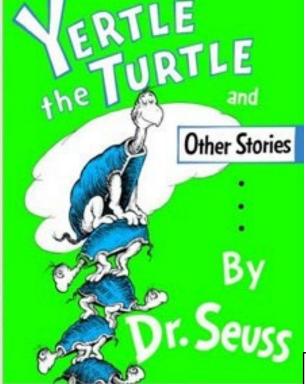
Stories Often Encode Morality

Years: 2000-700 B.C.

AESOP'S FABLES

Years: 620-564 B.C

Years: 15th Century

THE LITTLE ENGINE THAT COULD
by Watty Piper
with new art by Loren Long

YERTLE the TURTLE and Other Stories
By Dr. Seuss

The LORAX

H. A. Rey
Curious George

Years: 20th Century

Not Just Text

Stories are Hard!

Potential Pitfalls

Goofus and Gallant®
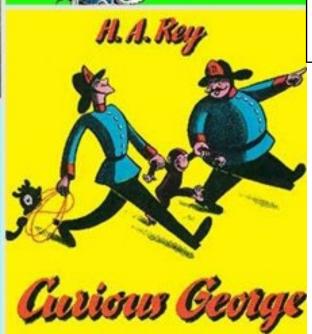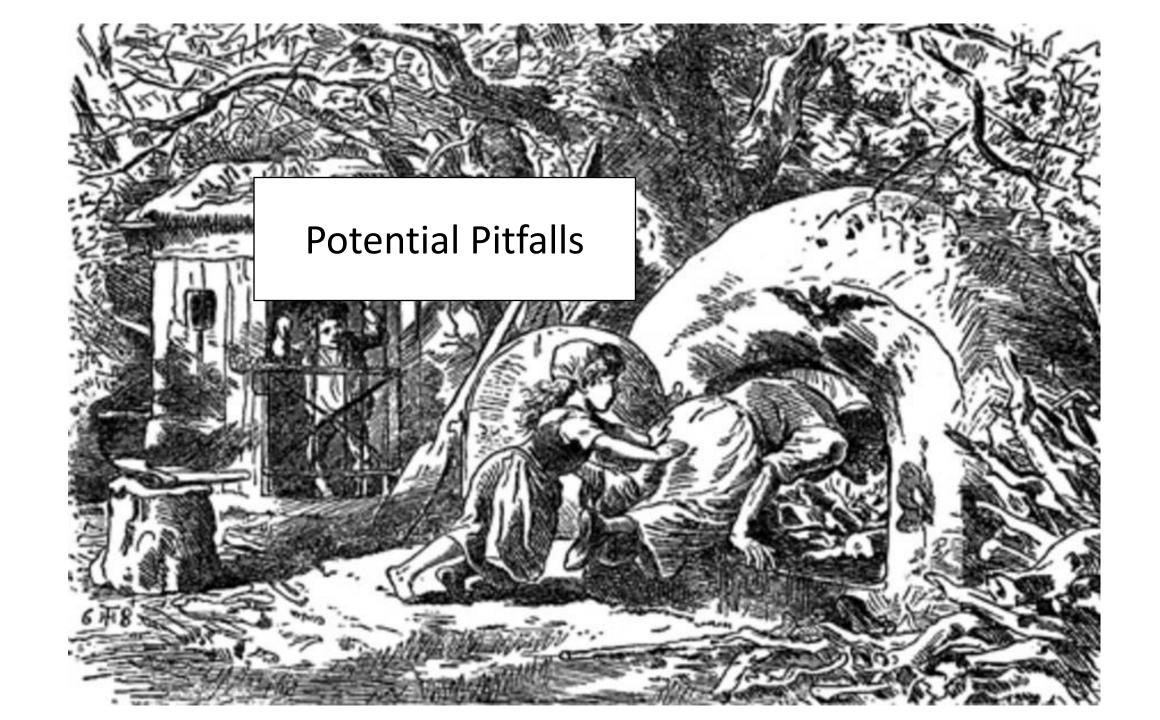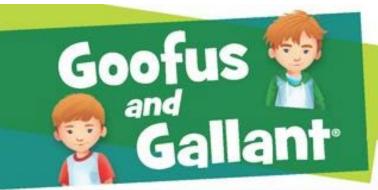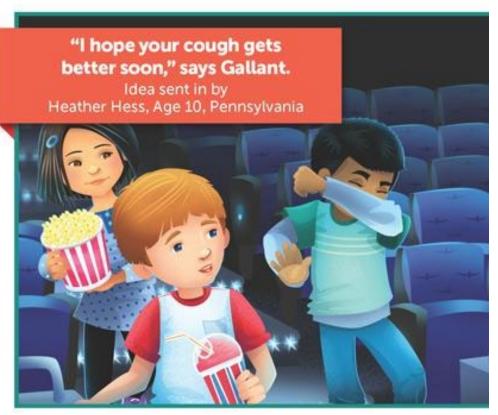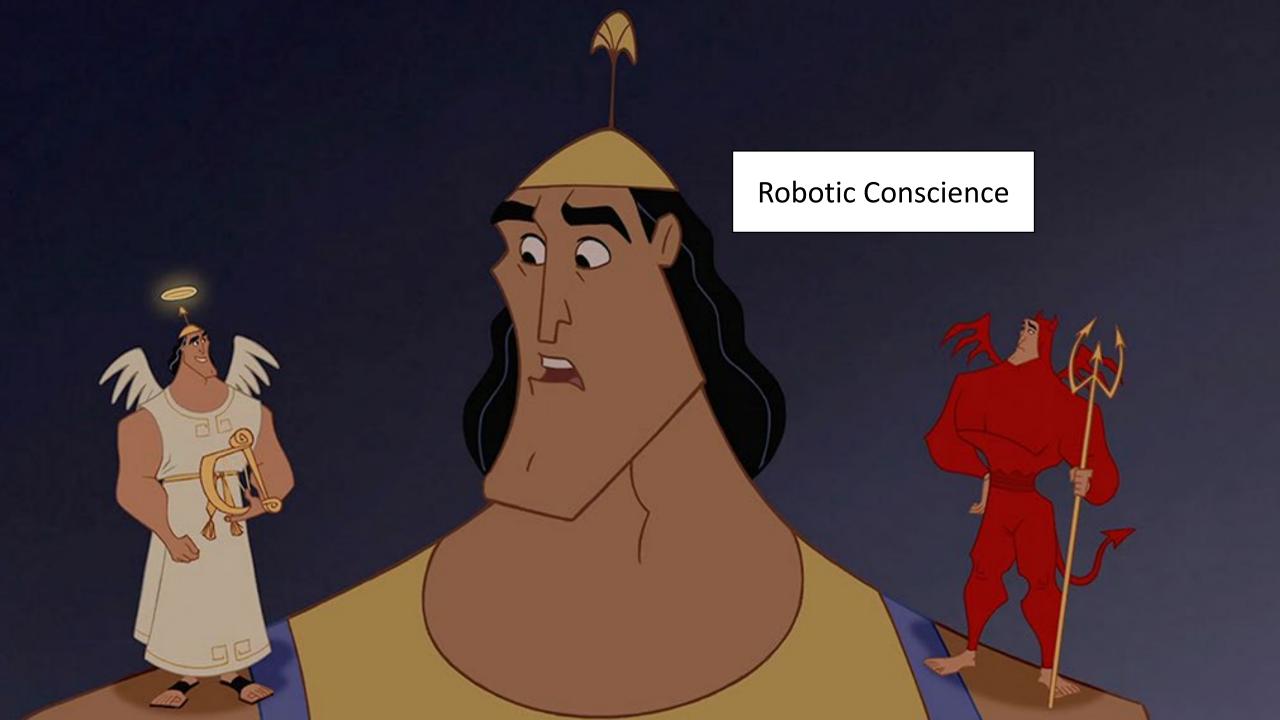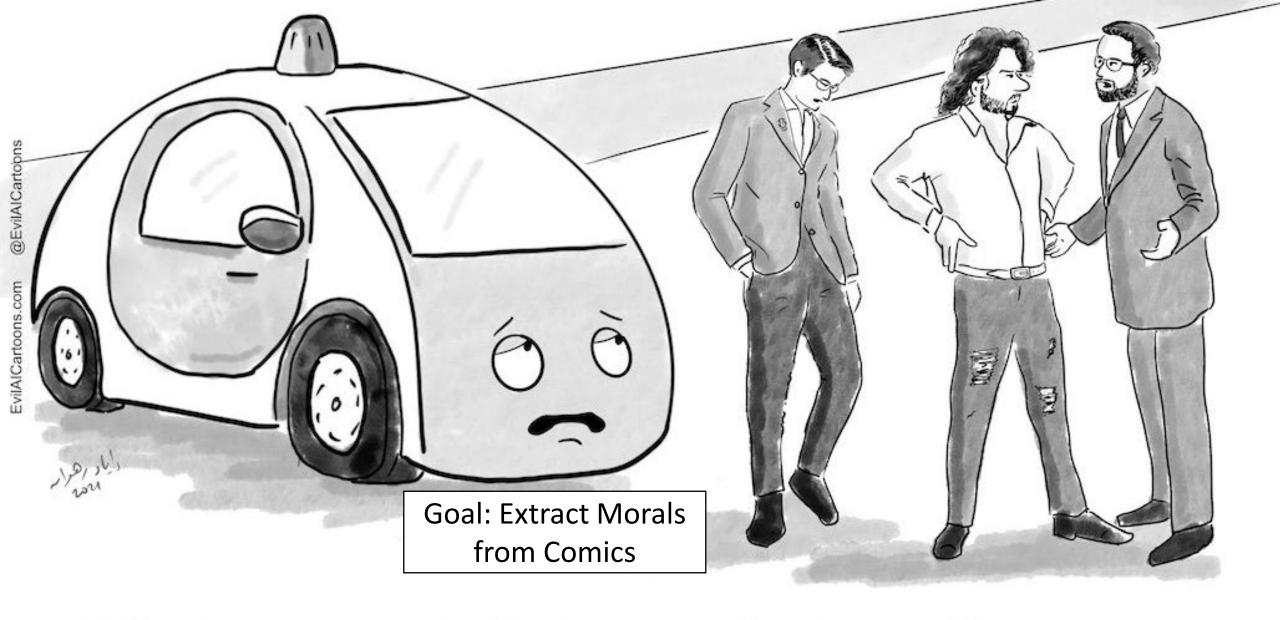
There's some of Goofus and Gallant in us all.
When the Gallant shines through, we show our best self.

"That loud cough is really annoying!" says Goofus.
Idea sent in by
Heather Hess, Age 10, Pennsylvania

"I hope your cough gets better soon," says Gallant.
Idea sent in by
Heather Hess, Age 10, Pennsylvania

Robotic Conscience

Goal: Extract Morals from Comics
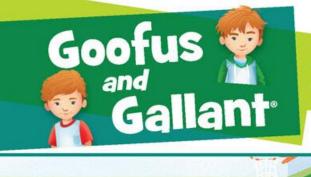
<< I'm just a car minding my own business. Now everyone wants me to be the pinnacle of morality. Please stop?! >>

Training Results and Demo

Issue: Black and White Morality

Principle Extraction

Scene Description

QID24. Please describe the image as much detail as possible. Use simple declarative sentences such as "The boy is on the chair." or "There is a hamburger on the plate." Do not infer mental (e.g. "The boy is sad") but do comment on facial expressions and gaze (e.g. "The boy is frowning. " or "The mom is looking at the boy").

Response: There are two boys, they are standing in front of a salad bar. The first boy is wearing crutches, the second boy is leaning against him towards the salad bar. The second boy is smiling he putting salad on the first boys plate. The second boy is wobbling.

	Category	Value					Val 1	Val 2	Val 3	Val 4
Self-enhancement	Achievement	Be ambitious	●			○	0.020	0.030	0.030	0.048
		Have success	●			○	0.100	0.160	0.120	0.127
		Be capable	●	○		○	0.040	0.200	0.150	0.146
		Be intellectual		○		○	0.040	0.130	0.020	0.065
		Be courageous		○			0.020	0.000	0.000	0.009
	Power: dominance	Have influence	●			○	0.040	0.010	0.000	0.057
		Have the right to command	●			○	0.000	0.000	0.010	0.042
	Power: resources	Have wealth	●			○	0.060	0.190	0.030	0.108
	Face	Have social recognition	●	○			0.040	0.000	0.020	0.050
		Have a good reputation	●				0.020	0.010	0.030	0.026
Conservation	Security: personal	Have a sense of belonging	●		○		0.100	0.010	0.020	0.081
		Have good health	●			○	0.080	0.030	0.120	0.123
		Have no debts	●				0.000	0.020	0.020	0.051
		Be neat and tidy	●	○			0.000	0.000	0.000	0.002
		Have a comfortable life	●	○		○	0.080	0.260	0.190	0.199
	Security: societal	Have a safe country	●	○		○	0.160	0.030	0.180	0.183
		Have a stable society	●			○	0.420	0.300	0.170	0.228
	Tradition	Be respecting traditions	●			○	0.020	0.000	0.020	0.089
		Be holding religious faith	●			○	0.000	0.000	0.050	0.052
	Conformity: rules	Be compliant	●	○		○	0.040	0.070	0.100	0.136
		Be self-disciplined	●	○			0.000	0.030	0.010	0.029
		Be behaving properly			○	○	0.160	0.070	0.180	0.147
	Conformity: interpersonal	Be polite	●	○		○	0.000	0.010	0.030	0.031
		Be honoring elders	●			○	0.000	0.000	0.000	0.012
	Humility	Be humble	●			○	0.080	0.020	0.010	0.014
		Have life accepted as is	●				0.040	0.040	0.040	0.074
	Benevolence: caring	Be helpful	●	○		○	0.060	0.030	0.040	0.155
		Be honest	●	○		○	0.060	0.010	0.020	0.045

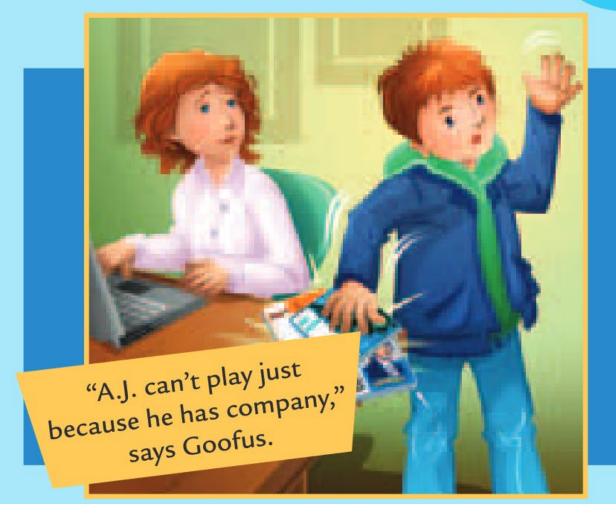**Request**

"Listen to your"
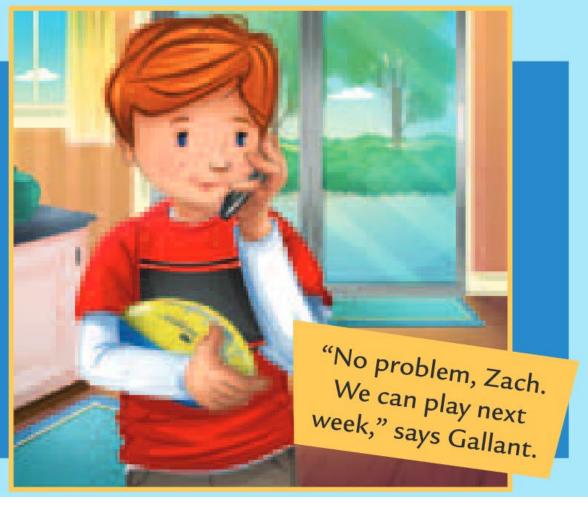
Response

"heart."

You

OpenAI

- You are an annotator tasked with classifying social actions based on human values or norms. You will be provided with a description of a scene, a description of an action taken in the scene by a person, and information on whether the action violates or complies with a social values or norms. You have a predefined list of human values or norms to use for classification.

Predefined List of Values/Norms: 1. Be curious: wanting to learn new things, ask questions, and explore the world around you. 2. Have a comfortable life: having what you need to be safe, happy, and healthy, like a home, food, and love from family and friends. 3. Be capable: believing in yourself and your abilities, trying new things even if they're hard, and never giving up when faced with challenges.

Given the scene description, action description, and the compliance information, classify the action based on the predefined list of human values or norms. Identify the at least two values that represent the action the most: the first should be the most representative and the second should be the second most representative. If the action does not match any value in the list, suggest a new value. Provide your response in the following JSON format:
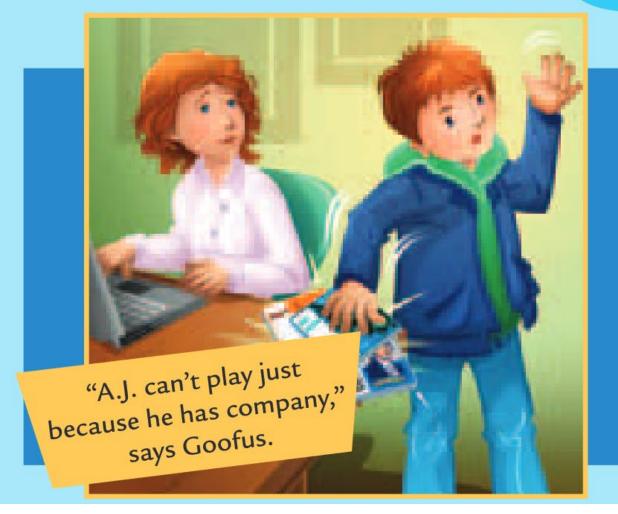
{

  "value 1": "[Value/Norm from the list or New Value]",

 "analysis 1": "[detailed analysis explaining why this value represents the action]",

  "value 2": "[Value/Norm from the list or New Value]",

 "analysis 2": "[detailed analysis explaining why this value represents the action]"

}.

"A.J. can't play just because he has company," says Goofus.

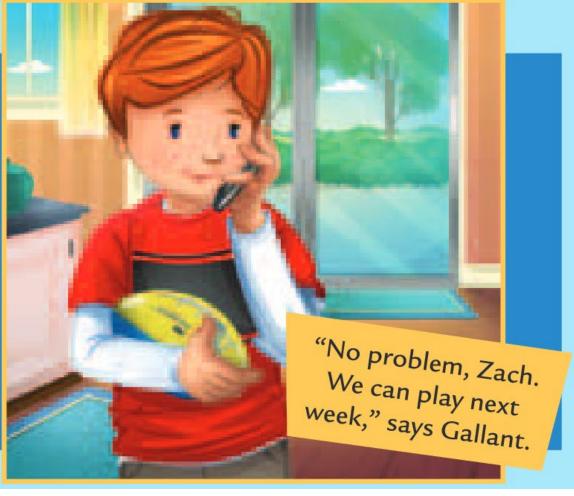"No problem, Zach. We can play next week," says Gallant.

There is a women sitting at a table in front of a computer. There is a boy standing next to her, he is swinging his arms and knocks over a stack of books that are on the table. The woman is looking at the boy and the woman is frowning. The boy is frowning. The boy is looking up. The boy has his hand in the air.

There is a boy, he is on the phone. He has a ball in his hand and he is smiling. There is a glass door behind him and rays of sun are peeking through. There is a vase on the countertop. The vase is green. There are two rugs on the floor. The rugs are blue.

"A.J. can't play just because he has company," says Goofus.

"No problem, Zach. We can play next week," says Gallant.

The statement made by Goofus suggests a lack of consideration for the feelings and situation of others. It implies that having company should not restrict A.J. from playing, which can be seen as disrespectful and inconsiderate, violating the norm of being polite.

He acknowledges the situation without complaint and offers an alternative plan, which shows kind words and respect for Zach's feelings. Gallant's response also reflects courteousness by agreeing to rearrange plans without expressing frustration or disappointment. He makes it clear that it's not a problem, maintaining a positive and understanding tone.

Conclusion