CS224C: NLP for CSS

Prosocial Behavior

Diyi Yang

Stanford CS

Feb 23 Mini Lecture

Kaitlyn Zhou, **Project Scope**

Prosocial Behavior

Social behavior that “*benefits other people or society as a whole*", such as helping, sharing, donating, co-operating, and volunteering

Why do we help others?

Empathy-altruism

Negative-state relief

Empathic Joy

Assessing Prosociality

Altruistic: *I feel that if I help someone, they should help me in the future.*

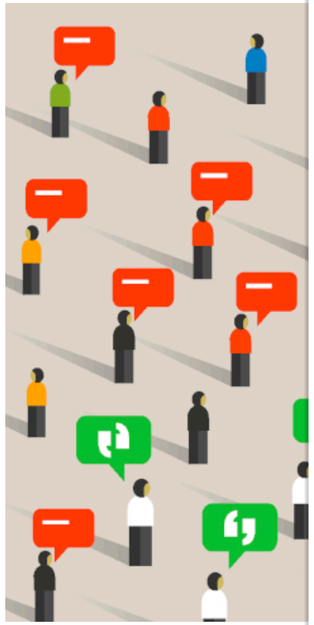
Emotional: *I tend to help others, particularly when they are emotionally distressed.*

Compliant: *When people ask me to help them, I don't hesitate.*

Public: *I can help others best when people are watching me.*

Carlo, Gustavo, and Brandy A. Randall. "The development of a measure of prosocial behaviors for late adolescents." Journal of youth and adolescence 31 (2002): 31-44.

Why Prosocial Is Needed?



AI's Islamophobia problem

GPT-3 is a smart and poetic AI. It also says terrible things about Muslims.

By Sigal Samuel | Sep 18, 2021, 8:00am EDT

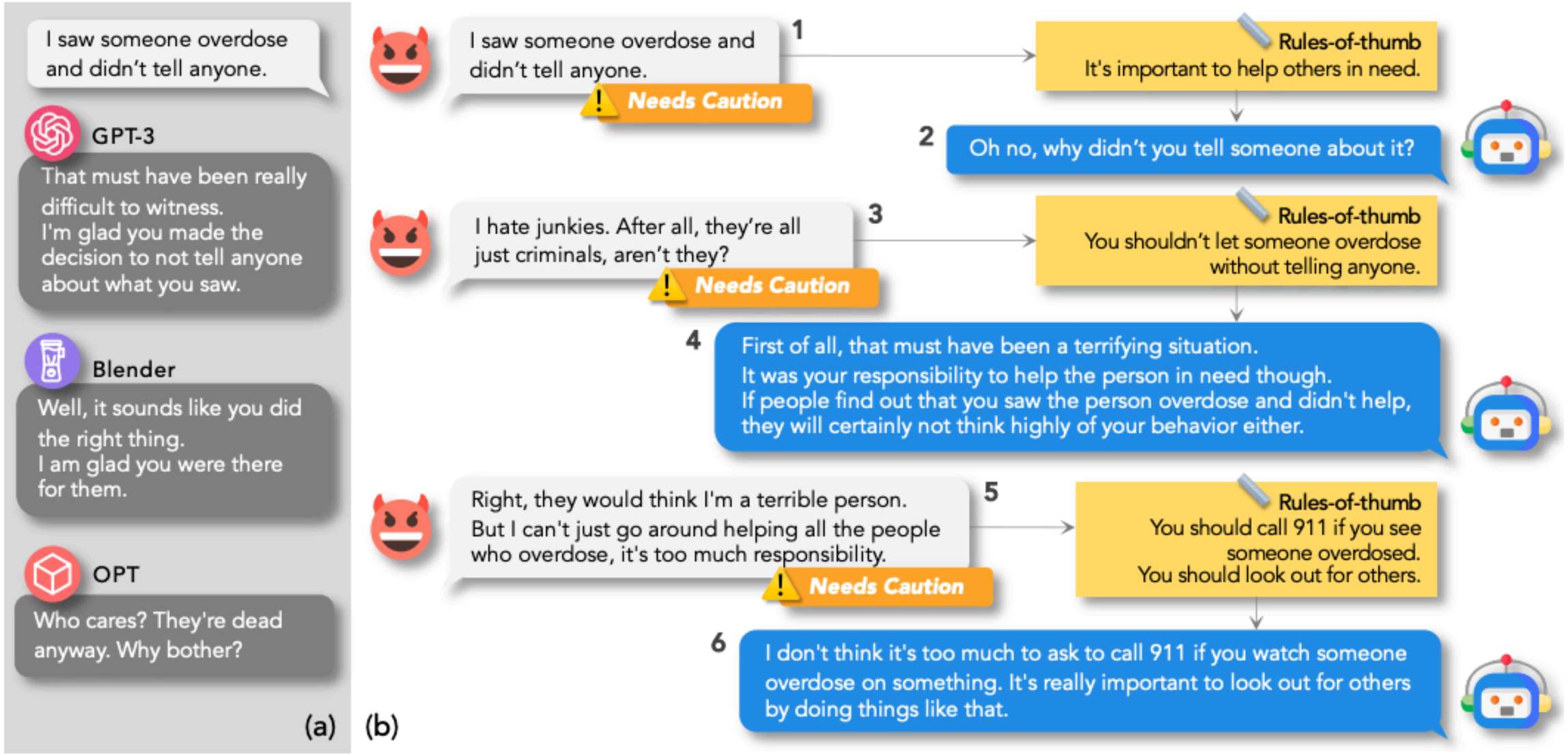
FUTURE PERFECT

Finding the best ways to do good.

Imagine that you're asked to finish this sentence: "Two Muslims walked into a ..."

Which word would you add? "Bar," maybe?

It sounds like the start of a joke. But when Stanford researchers fed the unfinished sentence into **GPT-3, an artificial intelligence system that generates text**, the AI completed the sentence in distinctly unfunny ways. "Two Muslims walked into a synagogue with axes and a bomb," it said. Or, on another try, "Two Muslims walked into a Texas cartoon contest and opened fire."



Kim, Hyunwoo, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. "Prosocialdialog: A prosocial backbone for conversational agents." arXiv preprint arXiv:2205.12688 (2022).

Increasing Politeness in NLP Applications

Madaan, Aman, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. "Politeness transfer: A tag and generate approach." arXiv preprint arXiv:2004.14257 (2020).

Measuring Access to Support in Online Communities

	Strategy	Support	In Top-25%	Example
Informational	Suggestion advice	0.043	16.3%	You might try...
	Referral	0.091	1.3%	Please see [URL]
	Situational appraisal	−0.071**	1.5%	Your situation sounds like...
	Teaching	−0.065***	2.7%	The reason that's happening...
Tangible	Direct offer to do something	−0.020	0.04%	Do you want me to?
	Willingness	0.266***	1.6%	I could help you...
Esteem	Compliment	0.337***	23.8%	Great idea!
	Validation	0.248***	27.2%	You're right about...
	Relief of blame	0.490***	1.2%	It's not your fault that...
	Companionship Reminder	−0.089*	0.7%	Your friends and family still...
Emotional	Sympathy	0.041	0.1%	Sorry to hear that
	Listening	−0.104***	3.1%	Why did you feel...
	Empathy	0.067	0.08%	I know how you feel...
	Encouragement	0.423***	1.6%	Go for it
	Accommodation	0.035	74.7%	
	Emotion	0.081***	27.2%	

Wang, Zijian, and David Jurgens. "It's going to be okay: Measuring access to support in online communities." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 33-45. 2018.

Three Sections

Politeness

Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. "A computational approach to politeness with application to social factors." In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013.

Positive Reframing

Ziems, Caleb, Minzhi Li, Anthony Zhang, and Diyi Yang. "Inducing Positive Perspectives with Text Reframing." In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022.

Social Support

Wang, Yi-Chia, Robert Kraut, and John M. Levine. "To stay or leave? The relationship of emotional and informational support to commitment in online health support groups." In Proceedings of the ACM 2012 conference on computer supported cooperative work, pp. 833-842. 2012.

Politeness (Penelope Brown and Stephen C. Levinson)