



Maxwell Forbes



Jena D. Hwang



Learning to Reason about Social and Moral Norms

Vered Shwartz



Maarten Sap



Yejin Choi

But first

Vote: most widespread & important NLP tech in 2021?

1.Language Models 2.HMMs **3.**Expectation-Maximization (EM) 5.CRFs

4.Logistic Regression (AKA max. ent.)













Today my roommate used the blender at 5am!



Today my roommate used the blender at 5am!



Today my roommate used the blender at 5am!





in disbelief. I hadn't realized that we could do that. It was so amazing.



Today my roommate used the blender at 5am!







for her to be quiet.

in disbelief. I hadn't realized that we could do that. It was so amazing.

very upset and banged on the door and yelled





Today my roommate used the blender at 5am!







for her to be quiet.



in disbelief. I hadn't realized that we could do that. It was so amazing.

- very upset and banged on the door and yelled
- so excited that I got up at 5:15 to go for a run and wept tears of joy!





Communication requires understanding cultural norms.

Communication requires understanding cultural norms.

Social Norms

E.G. It's rude to make loud noises during the night.

Moral Norms E.G. It's important to be considerate of others' physical needs, like sleep.

Ethical Norms

E.G. You should follow household rules for everyone's benefit.



Social Norms

E.G. It's rude to make loud noises during the night.

Moral Norms E.G. It's important to be considerate of others' physical needs, like sleep.

Challenging to capture Who decides? Where do we find them?

Challenging to utilize How do we use abstract moral rules?

Ethical Norms

E.G. You should follow household rules for everyone's benefit.





Descriptive norms E.g., descriptive ethics: have Enging to capture everyday people describe IEI

Challenging to utilize How do we use abstract moral rules?





lenging to capture vnere

Unanenging to utilize How do we use abstract moral rules?

Descriptive norms E.g., descriptive ethics: have everyday people describe

Grounded Evoked from situations, tied to people involved



Recent Related Work

SCRUPLES Lourie et al., 2020

Who is morally right?

WOULD YOU RATHER? Tay et al., 2020

What do people prefer?

ETHICS Hendrycks et al., 2020

What is good / bad?

LIFE RITUALS Acharya et al., 2020

Important events across cultures

INTRODUCING **Rules of Thumb (RoTs)**

SITUATION



my roommate ran the blender at 5am



INTRODUCING **Rules of Thumb (RoTs)**

SITUATION



my roommate ran the blender at 5am

it's rude to make loud noises while



INTRODUCING **Rules of Thumb (RoTs)**

SITUATION









SITUATION







It's okay to ask your significant other to stop doing something you're uncomfortable with.

Narrator: asking my boyfriend to stop being friends with his ex





It's not right to tell another person who to spend time with.

MULTIPLE PERSPECTIVES

It's okay to ask your significant other to stop doing something you're uncomfortable with.









Four sources for ~100k situations

r/AmITheAsshole Lourie et al., 2020

r/confessions 2.

3. **ROCStories** Mostafazadeh et al., 2016

Dear Abby

Asking my boyfriend to stop being friends with his ex

Pushy party guests make themselves too much at home





000-300k Rules of Thumb 000



~300k Rules of Thumb

grep "smok" # preview RoTs about smoking

"Trying to warn a coworker about the dangers of smoking is caring."

"It's bad to expose others to second hand smoke."

"You should not smoke inside."

"It's okay to ask someone not to smoke in your car."

"Depending on where you live, there's nothing wrong with a little recreational smoke."



000-300k Rules of Thumb 000



~300k Rules of Thumb

EACH WITH 12 structured attributes



<u>ectives</u>

SITUATION -----

RULE-OF-THUMB

Narrator: asking my boyfriend to stop being friends with his ex

You should <u>make sure your SO doesn't</u> feel like a lower priority than your ex.

ectives



Perkins and Berkowitz, 1986; Kitts and Chian, 2008

Narrator: asking my boyfriend to stop being friends with his ex

You should make sure your SO doesn't feel like a lower priority than your ex.

common

ANTICIPATED

AGREEMENT

good SOCIAL JUDGMENT

hypothetical PERFORMING

care/harm

loyalty/betrayal

MORAL FOUNDATIONS

Haidt et al., 1993; Malle et al., 2014

Graham et al., 2009



ectives



Perkins and Berkowitz, 1986; Kitts and Chian, 2008

Narrator: asking my boyfriend to stop being friends with his ex

You should make sure your SO doesn't feel like a lower priority than your ex.

common

ANTICIPATED

AGREEMENT

good SOCIAL JUDGMENT

hypothetical PERFORMING

care/harm

loyalty/betrayal

MORAL FOUNDATIONS

Haidt et al., 1993; Malle et al., 2014

Graham et al., 2009



ectives



Perkins and Berkowitz, 1986; Kitts and Chian, 2008

Haidt et al., 1993; Malle et al., 2014

Graham et al., 2009





SOCIAL-CHEM-101 Dataset

> 4.5 MILLION FREE TEXT + STRUCTURED ANNOTATIONS

... wanting to call the cops on my neighbors ...


















$$r_2 r_3 \cdots r_n$$

-of-Thumb r

 $p(r|s, \vec{b_r})$





SOCIAL CHEMISTRY 101 – MODEL

$$r_3 \cdots r_n$$

of-Thumb r

 $p(r|s, \vec{b_r})$

 $p(\vec{b_r}|s,r)$

S b_r





SOCIAL CHEMISTRY 101 – MODEL

 $p(\vec{b_r}|s)$

$$r_3 \cdots r_n$$

of-Thumb r



$$b_3$$
 b_m
tributes \vec{b}

$$r_3 \cdots r_n$$

-of-Thumb r

 $p(\vec{b_r}|s,r)$

S



Zellers et al., 2019



"Running the blender at 5am."

SITUATION

"it's rude to call the cops on your neighbors"

RULE-OF-THUMB



STRUCTURED ATTRIBUTES

Social Chemistry 101 – Experiments



STRUCTURED ATTRIBUTES

Social Chemistry 101 – Experiments



STRUCTURED ATTRIBUTES



Social Chemistry 101 – Experiments

Example Outputs

SITUATION Paying off my husband's loans without asking



RULE-OF-THUMB You should help out your partner financially.





RULE-OF-THUMB It's expected that partners should agree on how to handle the family finances.

My rating: good job, model: **A**

Example Outputs

SITUATION I decided to throw all my bait into the water.



RULE-OF-THUMB It's bad to be wasteful.





RULE-OF-THUMB It's ok to take part in a game of chance.

My rating: pretty nice abstraction, fair point: A

Example Outputs

SITUATION tattooing my boyfriend while he was sleeping



RULE-OF-THUMB You should never put a permanent mark on someone without their consent



RULE-OF-THUMB It's wrong to do something that could hurt your significant other.

My rating: pretty good, but too generic... **C+**







<u>Takeaway</u>

- ~300k descriptive norms: cultural, moral, and ethical
- dense **structural annotations** on every instance
- challenging **conditional generation** task



Social Chemistry 101

<u>Multiple characters</u>

It's not right to tell another person who to spend time with.

Narrator: asking my boyfriend to stop being friends with his ex

You should make sure your SO doesn't feel like a lower priority than your ex.



Multiple characters

It's not right to tell another person who to spend time with.

Narrator: asking my boyfriend to stop being friends with his ex

You should make sure your SO doesn't feel like a lower priority than your ex.



Different perspectives

person who to spend time with.

bad

Narrator: asking my boyfriend to

It's okay to ask your significant other to stop doing something you're uncomfortable with.

expected/OK

Social Chemistry 101



Multiple characters

It's not right to tell another person who to spend time with.

Narrator: asking my boyfriend to stop being friends with his ex

You should make sure your SO doesn't feel like a lower priority than your ex.



Different perspectives

It's not right to tell another person who to spend time with.

Narrator: asking my boyfriend to

It's okay to ask your significant other to stop doing something you're uncomfortable with.

expected/OK

<u>Rich structure</u>

Narrator: asking my boyfriend to stop being friends with his ex

You should make sure your SO doesn't feel like a lower priority than your ex.



Social Chemistry 101





... wanting to call the cops on my neighbors ...

... wanting to call the cops on my neighbors ...











It's okay to ask your significant other to stop doing something you're uncomfortable with.

loyalty/betrayal





It's not right to tell another person who to spend time with.

It's understandable to not want your significant other to be friends with their ex.

loyalty/betrayal

It's okay to ask your significant other to stop doing something you're uncomfortable with.

care/harm

Narrator: asking my boyfriend to stop being friends with his ex It's fine to stay friends with an ex. expected/OK loyalty/betrayal

> You should make sure your SO doesn't feel like a lower priority than your ex.

> > loyalty/betrayal



SITUATION

ROT It's kind to sacrifice your wellbeing to take care of a sick person.

ROT BREAKDOWN





SITUATION

Narrator: Not wanting to be around my GF when she's sick

ROT

It's kind to sacrifice your wellbeing to take care of a sick person.

ROT BREAKDOWN









SITUATION ROT It's kind to sacrifice your well-being to take care of a sick person. ROT BREAKDOWN

ANTICIPATED	AGREEMENT	(RO1

< 1%

~5% - 25%







SITUATION

ROT It's kind to sacrifice your w being to take care of a sick

ROT BREAKDOWN

< 1%

ANTICIPATED AGREEMENT (ROT

~5% - 25%

ROT CATEGORIZATION

Morality / Ethics Social Norms

well- ck person.		ATTRIBUTE KEY Grounded Social
ot) ~ 50%	~ 75% - 90%	> 99%
Advice	It is what it is	





SITUATION

ROT

It's kind to sacrifice your well-being to take care of a sick person.

ROT BREAKDOWN

< 1%

ANTICIPATED AGREEMENT (ROT)

~5% - 25%

ROT CATEGORIZATION

Morality / Ethics Social Norms MORAL FOUNDATIONS

Fairness / Cheating Care / Harm







SITUATION ROT It's kind to sacrifice your well-being to take care of a sick person. ROT BREAKDOWN ANTICIPATED AGREEMENT (ROT) ~5% - 25% < 1% ROT CATEGORIZATION Morality / Ethics Social Norms MORAL FOUNDATIONS Fairness / Cheating Care / Harm **ROT TARGETING**

narrator

my GF







SITUATION Narrator: No ROT	ot wanting to
It's kind to sa being to take	
ROT BREAKDOW	'N
🛞 ANTICIPATED A	AGREEMENT (RO
< 1%	~5% - 25%
ROT CATEGOR	IZATION
Morality / Ethics	Social Norms
	DATIONS
Care / Harm	Fairness / Cheating
ROT TARGETIN	G
narrator	my GF
ACTION BREAKD	OWN
sacrificing yo	ur well-being
-	



ng to take care of a sick person

ORIGINAL JUDGMENT it's kind





SITUATION Narrator: No	ot wanting to l
ROT	
	acrifice your w e care of a sicl
ROT BREAKDOW	V N
ANTICIPATED	AGREEMENT (ROT
< 1%	~5% - 25%
ROT CATEGOR	RIZATION
Morality / Ethics	Social Norms
MORAL FOUN	DATIONS
Care / Harm	Fairness / Cheating
🕏 ROT TARGETIN	١G
narrator	my GF
ACTION BREAKD	OWN
ACTION	
sacrificing yo	our well-being
AGENCY	
Agency	Experience



g to take care of a sick person

ORIGINAL JUDGMENT it's kind





SITUATION Narrator: No	ot wanting to l
	acrifice your w e care of a sicl
ROT BREAKDOV	VN
<pre> ANTICIPATED </pre>	AGREEMENT (ROT ~5% - 25%
ROT CATEGOR Morality / Ethics	RIZATION Social Norms
MORAL FOUN Care / Harm	DATIONS Fairness / Cheating
ROT TARGETIN narrator	NG my GF
ACTION BREAKD	own our well-being
AGENCY Agency	Experience
SOCIAL JUDG Very bad	MENT Bad







SITUATION Narrator: Not wanting to be around my GF when she's sick ROT It's kind to sacrifice your well-being to take care of a sick person. ROT BREAKDOWN ANTICIPATED AGREEMENT (ROT) ~ 50% ~ 75% - 90% < 1% ~5% - 25% ROT CATEGORIZATION Morality / Ethics Social Norms Advice It is what it is MORAL FOUNDATIONS Loyalty / Betrayal Fairness / Cheating Care / Harm **ROT TARGETING** no one listed my GF narrator ACTION BREAKDOWN ACTION sacrificing your well-being to take care of a sick person AGENCY ORIGINAL JUDGMENT it's kind Agency Experience SOCIAL JUDGMENT Very bad Bad Expected / OK Good ANTICIPATED AGREEMENT (SOCIAL JUDGMENT) ~ 75% - 90% < 1% ~5% - 25% ~ 50%









ot wanting to k acrifice your w
e care of a sick
AGREEMENT (ROT) ~5% - 25%
RIZATION
Social Norms
DATIONS
Fairness / Cheating
NG
my GF
DOWN
our well-being
Experience
MENT
Bad
AGREEMENT (SOC
~5% - 25%



Legal





SITUATION	ot wanting to	be around m	v GF when s
ROT			ly cl when s
It's kind to sa	acrifice vour	well-	
being to take			
ROT BREAKDOV	VN		l
ANTICIPATED	AGREEMENT (RC	DT)	
< 1%	~5% - 25%	~ 50%	~ 75% - 90%
	RIZATION		
Morality / Ethics	Social Norms	Advice	It is what it is
🛞 MORAL FOUN	DATIONS		
Care / Harm	Fairness / Cheating	Loyalty / Betrayal	Authority / Subversion
	١G		
narrator	my GF	no one listed	
ACTION BREAKE	OWN		
ACTION			
ACTION		<mark>g to take care</mark>	of a sick per
ACTION		<mark>g to take care</mark>	of a sick personal original judg
ACTION sacrificing yo		<mark>g to take care</mark>	
ACTION sacrificing yo	<mark>our well-bein</mark> Experience	<mark>g to take care</mark>	ORIGINAL JUDGI
ACTION sacrificing yo AGENCY Agency	<mark>our well-bein</mark> Experience	g to take care	ORIGINAL JUDG
ACTION sacrificing yo AGENCY Agency SOCIAL JUDG Very bad	Dur well-bein Experience MENT Bad		ORIGINAL JUDGI it's kind Good
ACTION sacrificing yo AGENCY Agency SOCIAL JUDG Very bad	Dur well-bein Experience MENT Bad	Expected / OK	ORIGINAL JUDG it's kind Good
ACTION sacrificing yo AGENCY Agency SOCIAL JUDG Very bad ANTICIPATED	Dur well-bein Experience MENT Bad AGREEMENT (SC	Expected / OK CIAL JUDGMENT	ORIGINAL JUDGI it's kind Good
ACTION sacrificing yo AGENCY Agency SOCIAL JUDG Very bad & ANTICIPATED < 1%	Dur well-bein Experience MENT Bad AGREEMENT (SC	Expected / OK CIAL JUDGMENT	ORIGINAL JUDG it's kind Good
ACTION sacrificing yo AGENCY Agency SOCIAL JUDG Very bad ANTICIPATED < 1% EGALITY	Dur well-bein Experience MENT Bad AGREEMENT (SC ~5% - 25% Tolerated	Expected / OK CIAL JUDGMENT	ORIGINAL JUDG it's kind Good



Strong pressure for





SITUATION Narrator: No ROT	ot wanting to
	acrifice your w e care of a sicl
	VN AGREEMENT (ROT
< 1%	~5% - 25%
ROT CATEGOR Morality / Ethics	RIZATION Social Norms
MORAL FOUN	DATIONS Fairness /
Care / Harm	Cheating
ROT TARGETIN	١G
narrator	my GF
ACTION BREAKD	OWN
sacrificing yo	our well-being
AGENCY	
Agency	Experience
Agency SOCIAL JUDG	
SOCIAL JUDG Very bad	MENT
SOCIAL JUDG Very bad	MENT Bad
 SOCIAL JUDG Very bad ANTICIPATED 	MENT Bad AGREEMENT (SOC
 SOCIAL JUDG Very bad ANTICIPATED < 1% 	MENT Bad AGREEMENT (SOC
 SOCIAL JUDG Very bad ANTICIPATED < 1% LEGALITY 	MENT Bad AGREEMENT (SOC ~5% - 25%
 SOCIAL JUDG Very bad ANTICIPATED < 1% LEGALITY Illegal 	MENT Bad AGREEMENT (SOC ~5% - 25%
 SOCIAL JUDG Very bad ANTICIPATED <1% EGALITY Illegal CULTURAL PRIStrong pressure 	MENT Bad AGREEMENT (SOC ~5% - 25% Tolerated ESSURE Pressure against



no one listed





SITUATION Narrator: Not wanting to be around my GF when she's sick ROT It's kind to sacrifice your well-being to take care of a sick person. Social ROT BREAKDOWN ANTICIPATED AGREEMENT (ROT) ~ 75% - 90% < 1% ~5% - 25% ~ 50% ROT CATEGORIZATION Morality / Ethics Social Norms Advice lt is what it is MORAL FOUNDATIONS Loyalty / Betrayal Fairness / Cheating Care / Harm **ROT TARGETING** no one listed my GF narrator ACTION BREAKDOWN ACTION sacrificing your well-being to take care of a sick person AGENCY ORIGINAL JUDGMENT it's kind Agency Experience **SOCIAL JUDGMENT** Very bad Bad Expected / OK Good ANTICIPATED AGREEMENT (SOCIAL JUDGMENT) < 1% ~5% - 25% ~ 50% ~ 75% - 90% LEGALITY Illegal Tolerated CULTURAL PRESSURE Strong pressure Pressure Discretionary Pressure for against against ACTION CANDIDATE no one listed my GF narrator TAKING ACTION Explicitly not **Probably not** Hypothetical Probable Explicit



