Maarten Sap
1314 Spring street, Apt 309
Seattle, WA 98104

http://maartensap.com
(+1) 443 248-6215
msap@cs.washington.edu

Research interests

Social commonsense reasoning, measuring and rewriting social biases in text, ethics and fairness in AI systems

Employment Experience

Carnegie Mellon University, Language Technologies Institute   starting 08/2022
Assistant Professor

Allen Institute for AI   07/2021 – present
Postdoc / Young Investigator,

University of Washington, Computer Science and Engineering   03/2016 – 07/2021
Research and Teaching Assistant with Noah Smith and Yejin Choi

Microsoft Research AI (MSR AI)   06/ 2019 – 09/2019
Research Intern with Eric Horvitz

Allen Institute for Artificial Intelligence (AI2)   06/2018 – 06/2019
Research Intern, MOSAIC project with Yejin Choi

University of Pennsylvania, Positive Psychology Center/Penn Medicine   06/2013 – 08/2015
Research Programmer at the World Well Being Project and Social Media and Health Innovation Lab

Education

University of Washington, Seattle, WA, USA   09/2015 – 07/2021
PhD in Computer Science & Engineering, research focus on Natural Language Processing
MS of Computer Science & Engineering   03/2018
Advised by Yejin Choi and Noah A. Smith

Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland   06/2014
BS in Communications and Information Systems

Awards & Nominations

WeCNLP Best paper   10/2020
"Social Bias Frames: Reasoning about Social and Power Implications of Language" Sap et al. (2020)

ACL Best short paper nomination (*top 5%*)   07/2019
"The Risk of Racial Bias in Hate Speech Detection" Sap et al. (2019)

Amazon Alexa Prize   11/2017
First place in the inaugural social chatbot development competition

Talks

Positive AI with Social Commonsense Models
MilaNLP seminar at Università Bocconi   10/2021
PAN workshop at CLEF 2021   09/2021

Positive AI with Social Commonsense Models
AKBC Workshop on Commonsense Reasoning   10/2021
University of Toronto Computer Science   04/2021
MIT EECS   03/2021
CMU LTI/MLD   03/2021

UChicago CS	03/2021
TTIC	02/2021
Emory CS	02/2021
Vanderbilt CS	02/2021
EPFL I&C	01/2021
Yale Data Science & Statistics seminar	01/2021

PowerTransformer: Unsupervised Controllable Revision for Biased Language Correction 11/2020
EMNLP conference

Social Bias Frames: Reasoning About Social and Power Dynamics
WeCNLP Summit   10/2020
ACL Conference   07/2020

Reasoning about Social Dynamics and Social Bias in Language
SRI seminar   01/2021
Georgia Tech NLP seminar   10/2020
Berkeley NLP seminar   02/2020
Stanford NLP seminar   02/2020

Social and Ethical Considerations in English Toxic Language Detection 08/2020
NLP with Friends

Recollection versus Imagination: Exploring Human Memory and Cognition via Neural Language Models
ACL Conference   07/2020

COMET: Commonsense Transformers for Automatic Knowledge Graph Construction
DARPA Communicating with Computers grant   11/2019

The Risk of Racial Bias in Hate Speech Detection
ACL Conference   07/2019
ICML Queer in AI workshop   06/2019

ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning
AAAI conference   01/2019
AI2 seminar   01/2019

Event2Mind: Commonsense Inference on Events, Intents, and Reactions
DARPA Communicating with Computers grant   07/2018

Detecting Implicit Bias in Text through Connotative Language   04/2018
UW Social Psychology seminar

Teaching

EMNLP 2021   11/2021
"Crowdsourcing Beyond Annotation" Tutorial – *co-presenter*

AI2 academy   08/2020
Tutorial on Commonsense Reasoning in Natural Language Processing – *co-presenter*

ACL 2020   07/2020
Tutorial on Commonsense Reasoning in Natural Language Processing – *co-presenter*

University of Washington
CSE 473 Artificial Intelligence – *Guest Lecture on Natural Language Processing*   Spring 2019

CSE 481 Natural Language Processing Capstone – *Teaching Assistant*	Spring 2017
CSE 490U Natural Language Processing – *Teaching Assistant*	Spring 2016

Mentoring

Mentored 14 students (11 bachelor's, 3 master's); 5 submitted papers to top-tier conferences (3 accepted to date)

• Sravani Nanduri – UW CSE BS Student	09/2021 – present
• Zhilin Wang – UW CLMS Student	01/2021 – present
• Skyler Hallinan – UW CSE BS Student	01/2021 – present
• Xuhui Zhou – UW CLMS Student	09/2019 – 08/2021
• Michelle Ma – UW CSE BS Student	09/2019 – 12/2020
• Sam Gehman – UW CSE MS student	09/2019 – 07/2020
• Aishwarya Nirmal – UW CSE MS student	01/2018 – 06/2019
• Kenta Takatsu – Cornell BS Student	07/2018 – 03/2019
• Zachary Horvitz – AI2 intern	07/2018 – 03/2019
• Sarah Yu – UW CSE BS Student	03/2018 – 06/2018
• Lanhao Wu – UW CSE BS Student	03/2018 – 06/2018
• Boyan Li – UW CSE BS Student	01/2018 – 06/2018
• Amy Shah – UW CSE BS Student	09/2017 – 06/2018
• Emily Allaway – UW CSE BS Student	07/2017 – 06/2018
• Marcela Cindy Prasetio – UW CSE BS Student	01/2016 – 06/2017

Service

Community Service

University of Washington, Seattle, WA

Diversity Committee	09/2016 – 12/2020
Graduate student advisory council (G5PAC)	01/2018 – 12/2020
Social Chair	09/2016 – 06/2017

ACL 2020	07/2020
Socio-cultural diversity and inclusion committee	

Queer in AI	07/2020
ACL social event organizer	

iPraxis Philadelphia	11/2013 – 03/2014
Scienteer – volunteer tutor for middle school science projects	

Johns Hopkins University	01/2013 – 06/2013
Secretary of Diverse Sexuality and Gender Alliance (DSAGA)	

Program Committee & Reviewing

Senior program committee

• ACL rolling review – action editor	2021
• AAAI – meta reviewer	2021

Conferences

• EMNLP	2018 – 2021
• ACL	2019 – 2020
• AAAI	2020
• NAACL	2019

Maarten Sap

1314 Spring street, Apt 309

Seattle, WA 98104

http://maartensap.com

(+1) 443 248-6215

msap@cs.washington.edu

- ICWSM   2021

Journals
- Transactions of ACL   2020
- Dementia and Geriatric Cognitive Disorders   2020
- Computational Linguistics   2019 – 2020
- Humanities and Social Sciences Communications   2019
- Journal of Artificial Intelligence Research   2019
- IEEE Transactions on Cognitive and Developmental Systems   2019
- Social Psychological and Personality Science   2018

Workshops
- ACL workshop on causal inference   2021
- NAACL Student Research Workshop   2019
- CLPsych workshop at ACL and NAACL   2016 – 2018
- Stylistic Variation workshop at NAACL   2018

Publications

Peer-reviewed conference articles

Jesse Dodge, **Maarten Sap**, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell & Matt Gardner (2021) *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus*. EMNLP

Ashutosh Baheti, **Maarten Sap**, Alan Ritter & Mark Riedl (2021) *Just Say No: Analyzing the Stance of Neural Dialogue Generation in Offensive Contexts*. EMNLP

Alisa Liu, **Maarten Sap**, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith & Yejin Choi (2021) *DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts*. ACL

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, **Maarten Sap** & Dan Klein (2021) *Detoxifying Language Models Risks Marginalizing Minority Voices*. NAACL

Xuhui Zhou, **Maarten Sap**, Swabha Swayamdipta, Yejin Choi & Noah A. Smith (2021) *Challenges in Automated Debiasing for Toxic Language Detection*. EACL

Xinyao Ma[*], **Maarten Sap**[*], Hannah Rashkin & Yejin Choi (2020) *PowerTransformer: Unsupervised Controllable Revision for Biased Language Correction*. EMNLP

Sam Gehman, Suchin Gururangan, **Maarten Sap**, Yejin Choi & Noah A Smith (2020) *RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models*. Findings of EMNLP

Maxwell Forbes, Jena D Hwang, Vered Shwartz, **Maarten Sap** & Yejin Choi (2020) *Social Chemistry 101: Learning to Reason about Social and Moral Norms*. EMNLP

Maarten Sap, Eric Horvitz, Yejin Choi, Noah A Smith & James W Pennebaker (2020) *Recollection versus Imagination: Exploring Human Memory and Cognition via Neural Language Models*. ACL

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith & Yejin Choi (2020) *Social Bias Frames: Reasoning about Social and Power Implications of Language*. ACL

Maarten Sap[*], Hannah Rashkin[*], Derek Chen, Ronan LeBras & Yejin Choi (2019) *Social IQa: Commonsense Reasoning about Social Interactions*. EMNLP

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi & Noah A Smith (2019) *The Risk of Racial Bias in Hate Speech Detection*. ACL

Maarten Sap
1314 Spring street, Apt 309
Seattle, WA 98104

http://maartensap.com
(+1) 443 248-6215
msap@cs.washington.edu

Antoine Bosselut, Hannah Rashkin, **Maarten Sap**, Chaitanya Malaviya, Asli Celikyilmaz & Yejin Choi (2019) *COMET: Commonsense Transformers for Automatic Knowledge Graph Construction*. ACL

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith & Yejin Choi (2019) *ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning*. AAAI

Hannah Rashkin, Antoine Bosselut, **Maarten Sap**, Kevin Knight & Yejin Choi (2018) *Modeling Naive Psychology of Characters in Simple Commonsense Stories*. ACL

Hannah Rashkin[*], **Maarten Sap**[*], Emily Allaway, Noah A. Smith & Yejin Choi (2018) *Event2Mind: Commonsense Inference on Events, Intents, and Reactions*. ACL

Maarten Sap, Marcella Cindy Prasetio, Ari Holtzman, Hannah Rashkin & Yejin Choi (2017) *Connotation Frames of Power and Agency in Modern Films*. EMNLP

Roy Schwartz, **Maarten Sap**, Ioannis Konstas, Li Zilles, Yejin Choi & Noah A Smith (2017) *The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task*. CoNLL

H. Andrew Schwartz, Gregory Park, **Maarten Sap**, Evan Weingarten, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Jonah Berger, Martin Seligman & Lyle Ungar (2015) *Extracting Human Temporal Orientation from Facebook Language*. NAACL

Maarten Sap, Gregory Park, Johannes C. Eichstaedt, Margaret L. Kern, David J. Stillwell, Michal Kosinski, Lyle H. Ungar & Hansen Andrew Schwartz (2014) *Developing Age and Gender Predictive Lexica over Social Media*. EMNLP

Peer-reviewed journal articles

Gregory Park, H Andrew Schwartz, **Maarten Sap**, Margaret L Kern, Evan Weingarten, Johannes C Eichstaedt, Jonah Berger, David J Stillwell, Michal Kosinski, Lyle H Ungar & Martin E P Seligman (2017) *Living in the Past, Present, and Future: Measuring Temporal Orientation with Language*. Journal of Personality

Margaret L Kern, Gregory Park, Johannes C Eichstaedt, H Andrew Schwartz, **Maarten Sap**, Laura K Smith & Lyle H Ungar (2016) *Gaining Insights From Social Media Language: Methodologies and Challenges*. Psychological Methods

Johannes C Eichstaedt, H Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, **Maarten Sap**, Christopher Weeg, Emily Larson, Lyle H Ungar & Martin E P Seligman (2015) *Psychological Language on Twitter Predicts County-level Heart Disease Mortality*. Psychological Science

Charlene A Wong, **Maarten Sap**, Hansen Andrew Schwartz, Robert Town, Tom Baker, Lyle Ungar & Raina M Merchant (2015) *Twitter Sentiment Predicts Affordable Care Act Marketplace Enrollment*. Journal of Medical Internet Research

Raina M. Merchant, Yoonhee P. Ha, Charlene A. Wong, H. Andrew Schwartz, **Maarten Sap**, Lyle H. Ungar & David A. Asch (2014) *The 2013 US Government Shutdown (#Shutdown) and Health: An Emerging Role for Social Media*. American Journal of Public Health

Peer-reviewed workshop articles

Tal August, **Maarten Sap**, Elizabeth Clark, Katharina Reinecke & Noah A. Smith (2020) *Exploring the Effect of Author and Reader Identity in Online Story Writing: the StoriesInTheWild Corpus*. Workshop on Narrative Understanding, Storylines, and Events (NUSE) @ ACL

Roy Schwartz, **Maarten Sap**, Ioannis Konstas, Li Zilles, Yejin Choi & Noah A Smith (2017) *Story Cloze task: UW NLP System*. EACL Workshop LSD Sem

Daniel Preotiuc-Pietro, **Maarten Sap**, H Andrew Schwartz & Lyle Ungar (2015) *Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task*. NAACL Workshop on CLPsych

Maarten Sap

http://maartensap.com

1314 Spring street, Apt 309

(+1) 443 248-6215

Seattle, WA 98104

msap@cs.washington.edu

Daniel Preotiuc-Pietro, Johannes Eichstaedt, Gregory Park, **Maarten Sap**, Laura Smith, Victoria Tobolsky, H Andrew Schwartz & Lyle Ungar (2015) *The Role of Personality, Age and Gender in Tweeting about Mental Illnesses*. NAACL Workshop on CLPsych

H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, **Maarten Sap**, David Stillwell, Michal Kosinski & Lyle Ungar (2014) *Towards Assessing Changes in Degree of Depression through Facebook*. ACL Workshop on CLPsych

Other peer-reviewed articles (demos, etc.)

Hao Fang, Hao Cheng, **Maarten Sap**, Elizabeth Clark, Ariel Holtzman, Yejin Choi, Noah A Smith & Mari Ostendorf (2018) *Sounding Board: A User-Centric and Content-Driven Social Chatbot*. NAACL System Demonstrations

H Andrew Schwartz, Salvatore Giorgi, **Maarten Sap**, Patrick Crutchley, Lyle Ungar & Johannes Eichstaedt (2017) *DLATK: Differential Language Analysis ToolKit*. EMNLP System Demonstrations

Hao Fang, Hao Cheng, Elizabeth Clark, Ariel Holtzman, **Maarten Sap**, Mari Ostendorf, Yejin Choi & Noah A Smith (2017) *Sounding Board - University of Washington's Alexa Prize Submission*. Alexa Prize Proceedings

H Andrew Schwartz, **Maarten Sap**, Margaret L Kern, Johannes C Eichstaedt, Adam Kapelner, Megha Agrawal, Eduardo Blanco, Lukasz Dziurzynski, Gregory Park, David Stillwell, Michal Kosinski, Martin E P Seligman & Lyle H Ungar (2016) *Predicting individual well-being through the language of social media*. Biocomputing 2016: Proceedings of the Pacific Symposium

Thesis

Maarten Sap (2021) *Positive AI with Social Commonsense Models*.

References

Yejin Choi (PhD advisor)

Brett Helsel Associate Professor of Computer Science & Engineering

University of Washington

yejin@cs.washington.edu

Noah A. Smith (PhD advisor)

Professor of Computer Science & Engineering

University of Washington

nasmith@cs.washington.edu

Eric Horvitz

Technical Fellow and Chief Scientific Officer

Microsoft Research

horvitz@microsoft.com

Dan Jurafsky

Jackson Eli Reynolds Professor in Humanities
Professor of Linguistics
Professor of Computer Science

Stanford University

jurafsky@stanford.edu