

2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)

Online
14 July 2022

ISBN: 978-1-7138-5643-6

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2022) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2022)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>An Encoder Attribution Analysis for Dense Passage Retriever in Open-Domain Question Answering</i> Minghan Li, Xueguang Ma and Jimmy Lin	1
<i>Attributing Fair Decisions with Attention Interventions</i> Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg and Aram Galstyan	12
<i>Does Moral Code have a Moral Code? Probing Delphi’s Moral Philosophy</i> Kathleen C. Fraser, Svetlana Kiritchenko and Esmā Balkir	26
<i>The Cycle of Trust and Responsibility in Outsourced AI</i> Maximilian Castelli and Linda C. Moreau, Ph.D.	43
<i>Explaining Neural NLP Models for the Joint Analysis of Open-and-Closed-Ended Survey Answers</i> Edoardo Mosca, Katharina Harman, Tobias Eder and Georg Groh	49
<i>The Irrationality of Neural Rationale Models</i> Yiming Zheng, Serena Booth, Julie Shah and Yilun Zhou	64
<i>An Empirical Study on Pseudo-log-likelihood Bias Measures for Masked Language Models Using Paraphrased Sentences</i> Bum Chul Kwon and Nandana Mihindukulasooriya	74
<i>Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models</i> Esmā Balkir, Svetlana Kiritchenko, Isar Nejadgholi and Kathleen Fraser	80
<i>ER-TEST Evaluating Explanation Regularization Methods for NLP Models</i> Brihi Joshi, Aaron Chan, Ziyi Liu and Xiang Ren	93