Towards Human-Centered Optimality Criteria

by

Asma Ghandeharioun

S.M., Massachusetts Institute of Technology (2016)
B.Sc., Sharif University of Technology (2014)

Submitted to the
Program in Media Arts and Sciences, School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2021

Author . . . . . .
Program in Media Arts and Sciences
May 20, 2021

Certified by . . . . . .
Rosalind W. Picard
Professor of Media Arts and Sciences

Accepted by . . . . . .
Tod Machover
Academic Head, Program in Media Arts and Sciences

Towards Human-Centered Optimality Criteria

by

Asma Ghandeharioun

Submitted to the
Program in Media Arts and Sciences, School of Architecture and Planning,
on May 20, 2021, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

Despite the transformational success of machine learning across various applications, examples of deployed models failing to recognize and support human-centered (HC) criteria are abundant. In this thesis, I conceptualize the space of human-machine collaboration with respect to two components: interpretation *of people* by machines and interpretation of machines *by people*. I develop several tools that make improvements along these axes.

First, I develop a pipeline that predicts depressive symptoms rated by clinicians from real-world longitudinal data outperforming several baselines. Second, I introduce a novel, model-agnostic, and dataset-agnostic method to approximate interactive human evaluation in open-domain dialog through self-play that is more strongly correlated with human evaluations than other automated metrics commonly used today. While dialog quality evaluation metrics predominantly use word-level overlap or distance metrics based on embedding resemblance to each turn of the conversation, I show the significance of taking into account the conversation's trajectory and using proxies such as sentiment, semantics, and user engagement that are psychologically motivated. Third, I demonstrate an uncertainty measurement technique that helps disambiguate annotator disagreement and data bias. I show that this characterization also improves model performance. Finally, I present a novel method that allows humans to investigate a predictor's decision-making process to gain better insight into how it works. The method jointly trains a generator, a discriminator, and a concept disentangler, allowing the human to ask "what-if" questions. I evaluate it on several challenging synthetic and realistic datasets where previous methods fall short of satisfying desirable criteria for interpretability and show that our method performs consistently well across all. I discuss its applications to detect potential biases of a classifier and identify spurious artifacts that impact predictions using simulated experiments.

Together, these novel techniques and insights provide a more comprehensive interpretation *of people* by machines and more powerful tools for interpretation of machines *by people* that can move us closer to HC optimality.

Thesis Supervisor: Rosalind W. Picard
Title: Professor of Media Arts and Sciences
Massachusetts Institute of Technology

This dissertation has been reviewed and approved by the following committee members

Thesis Supervisor . . . . . .

Rosalind W. Picard, Sc.D.
Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Thesis Reader . . . . . .

David Sontag, Ph.D.
Associate Professor of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Thesis Reader . . . . . .

Zachary C. Lipton, Ph.D.
Assistant Professor of Operations Research and Machine Learning
Carnegie Mellon University

Acknowledgments

I still vividly remember the night that a 20-year-old me came across a talk by Roz about detecting seizures from wearable EDA [189], and thought to myself: "Wow! You can do cool computational work AND help people, potentially saving lives?!". I spent days and weeks researching what the affective computing group at MIT Media Lab was doing. I got so inspired that I framed my bachelor's thesis around measuring heart rate and respiratory rate on smartphones [68], and gathered the courage to apply to the lab a few months later. Thank you, Roz, for taking a chance on me then and for the opportunity to join your lab. Thank you for inspiring, guiding, encouraging, and supporting me all these years, for helping me grow as a researcher and as a person, and for letting me follow my interests when I wanted to venture into new research directions. Of course, drinking from the firehouse has come with its challenges, but having an extremely knowledgeable advisor who genuinely cares has made it much more manageable. What a journey has it been, full of ups and downs. In retrospect, every moment has been worth it.

I want to express my deepest gratitude to David and Zack for joining my committee and sharing their insights with me. David, your perspective has long been a source of inspiration to me since the day I took your class in 2017. Thanks for all the great advice you gave me to improve my work. I am amazed by your efficiency and creativity and how you manage to provide so many great suggestions in a 15-minute meeting. Zack, thanks for your thought-provoking opinion pieces that have shaped my thinking over the years. Your prose has always been a source of introspection and analytical thinking. Thanks for all the advice and suggestions you gave me to improve my work. I am in awe of how effortlessly you wrap serious work in funny and engaging conversations.

Of course, none of this work could have been possible without my amazing collaborators. Been, thank you for being a fantastic role model and a shining light of inspiration and insight. Who knew a coffee chat at CVPR could lead to such a wonderful long-term collaboration. I am so grateful for your mentorship and generosity and always giving me

the most genuine and helpful advice. Natasha, your enthusiasm and insights are contagious. Working together has been a highlight of my journey. Thanks, Brendan and Brian, for hosting me at Google and opening doors to new research directions. Thank you, Mary and Daniel, for hosting me at Microsoft Research and trusting me to design and conduct a study from start to finish. Sara, thanks for your limitless intellectual curiosity and kindness that has made our collaboration so rewarding. Thank you to many more collaborators at MIT: Asaf Azaria, Darian Bhathena, Neska El Haouij, Szymon Fedor, Craig Ferguson, Javier Hernandez, Noah Jones, Agata Lapedriza, Jinmo Lee, Alexander Lynch, Pattie Maes, Akane Sano, Judy Shen, Marek Subernat, Sebastian Zepf, and Diane Zhou, at Harvard: Abdul Saleh, Georgia Tech: Grace Leslie, at Google: Shane Gu, Chun-Liang Li, at Microsoft Research: Kael Rowan, and at MGH: Jonathan Alpert, Kate Bentley, Chelsea Dale, Esther Howe, Dawn Ionescu, Ashley Meyer, David Mischoulon, Paola Pedrelli, Lisa Sangermano, Benjamin Shapero. I deeply value our collaboration, and none of this work would have been possible without you, nor as exciting or as fun.

I want to thank all my friends and colleagues in the affective computing group. What a dynamic group of diverse, inspiring, ambitious, hardworking, and thoughtful people! Thank you, Natasha, Sara, Agata, Matt, Grace, Yue, Javier, Judy, Ehi, Kristy, Vincent, Noah, Akane, Katie, Rob, Karthik, Neska, Darian, Craig, Oggi, Yuanbo, Eiji, Oliver, Daniel, Sebastian, Fengjiao, Yadid, Terumi, Aithne, and Rashmi. Thanks for making the fun parts more memorable and the rocky patches worth the effort, and being sources of strength and compassion. Special thanks to Sable for doing an incredible amount of work in the background so effortlessly and smoothly. To all my friends outside of the affective computing group, Elahe, Sabah, Mina P, Sadegh, Sajjad, Sepide, Ali V, Ameneh, Mohammad R, Maryam A, Vahid, Ramya, Karthik, Saeed, Maryam M, Hessam, Sahoora, Amirhossein, Reihane, Donia, Mina D, Fereshte, Homa, Mahdi H, Nicolas, Latifeh, Leila, Hamid, Fatemeh K, Tahereh, Mohammad A, Fatemeh M, Mina K, Ali A, Mahdi A and many more: thank you for all the laughter we shared, all the trips we took together, all the get-togethers and celebrations, and for keeping me sane all these years.

I'm forever grateful to my family for reminding me of what truly matters. Thank you, mom, for your love and support and for teaching me the importance of work ethic. Thank you, Azra and Hosna, for instilling curiosity in me as a kid and for being my closest friends and confidants as an adult. Thank you, Mohammad Saeed, for all the fun memories we have together that still make me laugh even when I'm down. Thank you, dad, for teaching me to stay calm by example rather than by words.

Last but not least, thank you, Ardavan, for your endless love and support all these years, for helping me practice a growth mindset, and for showing the value of persistence, optimism, and patience. Thanks for bearing the date nights that turned into my research rants, for being my best friend and a profoundly insightful person. I will forever be grateful to MIT not only because of the degrees it awarded me but for helping me find you. Having someone who believes in you even when you doubt yourself, thoroughly understands and supports you and your dreams, and enables you to overcome your fears without an ounce of judgment is transformative.

Contents

List of Figures

List of Tables

Statement of Contributions

The chapters of this thesis comprise work from the following papers. An additional section at the end of each chapter provides more details about contributions of my co-authors.

Chapter 2:

- A. Ghandeharioun, S. Fedor, L. Sangermano, D. Ionescu, J. Alpert, Chelsea Dale, D. Sontag, and R. Picard. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In *ACII*. IEEE, 2017

- Asma Ghandeharioun, Lisa Sangermano, Rosalind Picard, Jonathan Alpert, Chelsea Dale, Dawn Ionescu, and Szymon Fedor. Objective vs. subjective reports of sleep quality in major depressive disorder: A pilot study. In *Anxiety and Depression Association of America*. ADAA, 2017

- A. Ghandeharioun, L. Sangermano, R. Picard, J. Alpert, C. Dale, D. Ionescu, and S. Fedor. Objective vs. subjective reports of sleep quality in major depressive disorder: A pilot study. In *Anxiety and Depression Association of America*. ADAA, 2017

- Asma Ghandeharioun, Szymon Fedor, Lisa Sangermano, Jonathan Alpert, Chelsea Dale, Dawn Ionescu, and Rosalind Picard. Location variability from commodity phone sensors is negatively associated with self-reported depression score: A pilot study. In *Association for Psychological Sciences*. APS, 2017

Chapter 3:

- A. Ghandeharioun*, J. H. Shen*, N. Jaques*, C. Ferguson, N. Jones, A. Lapedriza, and R. Picard. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *NeurIPS*, pages 13658–13669, 2019. *Equal contribution

Chapter 4:

- A. Ghandeharioun, B. Eoff, B. Jou, and R. W. Picard. Characterizing sources of uncertainty to proxy calibration and disambiguate annotator and data bias. In *ICCVW*, pages 4202–4206. IEEE, 2019

Chapter 5:

- A. Ghandeharioun, B. Kim, C. Li, B. Jou, B. Eoff, and R. Picard. DISSECT: Disentangled simultaneous explanations via concept traversals. 2021. In preparation

The following papers are only briefly discussed within this thesis:
Chapter 2:

- P. Pedrelli*, S. Fedor*, A. Ghandeharioun, E. Howe, D. Ionescu, D. Bhathena, C. Dording, L. Fisher, C. Cusin, M. Nyer, A. Yeung, L. Sangermano, D. Mischoulon, J. Alpert, and R. Picard. Monitoring changes in depression severity using wearable and mobile sensors. 2020. *Equal contribution

- E. Howe, A. Ghandeharioun, P. Pedrelli, D. Mischoulon, R. Picard, and S. Fedor. Location patterns from phone sensors may help predict depressive symptoms: A longitudinal pilot study. *ABCT - Tech SIG*, 2017

- Lisa Sangermano, Asma Ghandeharioun, Rosalind Picard, Jonathan Alpert, Chelsea Dale, Szymon Fedor, and Dawn Ionescu. Incoming cell phone data as a potential predictor of depression severity: A pilot study. In *Anxiety and Depression Association of America*. ADAA, 2017

- P. Pedrelli, E. Howe, D. Mischoulon, R. Picard, A. Ghandeharioun, and S. Fedor. Integrating EMA, clinical assessment and wearable sensors to examine the association between major depressive disorder (MDD) and alcohol use. *Iproceedings*, 3(1):e51, 2017

Chapter 3:

- Abdelrhman Saleh*, Natasha Jaques*, Asma Ghandeharioun, Judy Hanwen Shen, and Rosalind Picard. Hierarchical reinforcement learning for open-domain dialog. *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2019. *Equal Contribution

- Natasha Jaques*, Judy Hanwen Shen*, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind Picard. Human-centric dialog training via offline reinforcement learning. *EMNLP*, 2020. *Equal Contribution

The thesis body does not include several projects that I have worked on during my PhD. For reference, the following publications are not discussed in this thesis:

- Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. EMMA: An emotion-aware wellbeing chatbot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019

- Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. Towards understanding emotional intelligence for behavior change chatbots. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 8–14. IEEE, 2019

- Grace Leslie*, Asma Ghandeharioun*, Diane Zhou, and Rosalind W Picard. Engineering music to slow breathing and invite relaxed physiology. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019. *Equal Contribution

- Asma Ghandeharioun, Asaph Azaria, Sara Taylor, and Rosalind W Picard. "Kind and grateful": A context-sensitive smartphone app utilizing inspirational content to promote gratitude. *Psychology of well-being*, 6(1):9, 2016

- Asma Ghandeharioun, Asaph Azaria, Sara Taylor, Pattie Maes, and Rosalind Picard. Promoting kindness and gratitude with a smartphone and triggers. *Annals of Behavioral Medicine*, 50(Supplement 1):266, 2016

- Sebastian Zepf, Neska El Haouij, Jinmo Lee, Asma Ghandeharioun, Javier Hernandez, and Rosalind W Picard. Studying personalized just-in-time auditory breathing guides and potential safety implications during simulated driving. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 275–283, 2020

- Noah Jones, Natasha Jaques, Pat Pataranutaporn, Asma Ghandeharioun, and Rosalind Picard. Analysis of online suicide risk with document embeddings and latent dirichlet allocation. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–5. IEEE, 2019

- A. K. Meyer, S. Fedor, A. Ghandeharioun, D. Mischoulon, R. Picard, and P. Pedrelli. Feasibility and acceptability of the Empatica E4 sensor to passively assess physiological symptoms of depression. *ABCT*, 2020

- E. Howe, M. Nauphal, B. Shapero, K. Bentley, D. Mischoulon, A. Ghandeharioun, S. Fedor, R. Picard, and P. Pedrelli. Depression and emotional reactivity: A closer examination of daily variations in affect. *ABCT*, 2018

- Ardavan Saeedi, Matthew Hoffman, Stephen DiVerdi, Asma Ghandeharioun, Matthew Johnson, and Ryan Adams. Multimodal prediction and personalization of photo edits with deep generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 1309–1317. PMLR, 2018

Chapter 1

Introduction

Recent successes in machine learning have been transformational across a range of applications from computer vision to natural language processing and beyond [17, 39, 80, 184, 219, 229]. However, examples of deployed models failing to recognize and support human-centric (HC) criteria are abundant, such as disproportional assignment of risk to certain demographics [2] or making predictions based on spurious correlations [254]. This calls for a more successful collaboration between machine learning models and humans to better formalize some of the nuanced HC criteria and to develop computational tools for inspection and introspection in model development with respect to them.

Mutual understanding is known to be a critical element in effective communication between people [154]. The same argument can be extended to collaboration between machines and humans: A machine learning model can better adjust to human needs and preferences by inferring human's state and intention [190], and equipping people with tools for better interpretation and investigation of models could lead to advances in scientific understanding, improving safety, uncovering hidden biases, evaluating fairness, and beyond [31, 45, 78]. In this thesis, I conceptualize the space of machine and human interaction with respect to these two components (Figure 1-1): interpretation *of people* by machines and interpretation of machines *by people*. Throughout the chapters of this thesis, I show that improvements across both axes could bring us closer to HC optimality.

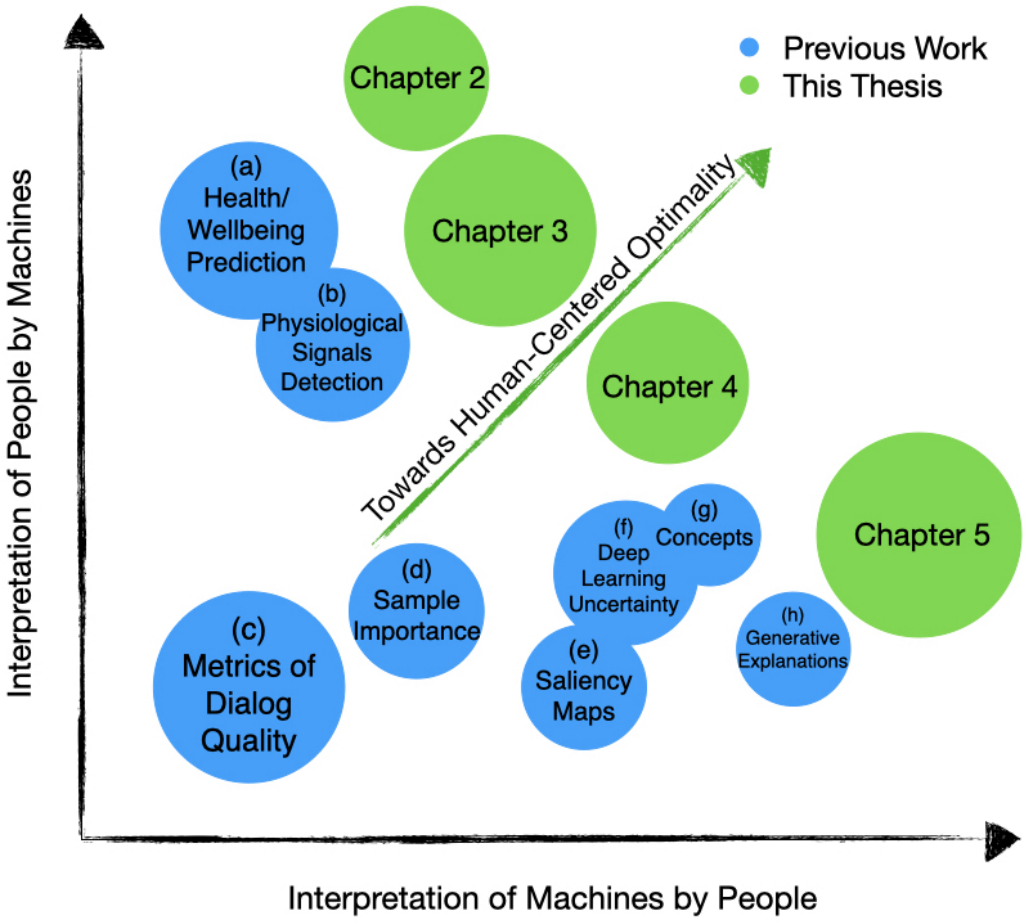


Figure 1-1: A conceptual framework for machine learning and human interaction with respect to human-centered optimality and how the thesis roadmap fits within this framework. The x-axis shows interpretation of machines *by people*: techniques that allow humans to investigate decision making of machine learning systems. Tools that facilitate investigation and interpretation can empower humans to identify potential failure points and inform actionable directions for improvement. Higher x values represent increasing flexibility of investigation through interpretation tools. The y-axis represents interpretation *of people* by machines: endowing machines with mechanisms to support, infer, or adapt to human states. More accurate inference can lead to better adaptation of machines to human preferences or improved tools for managing states, such as personal wellbeing. Higher y values represent more comprehensive consideration of human state. The origin represents no interpretation of machines by humans ($x = 0$) and no consideration of human state ($y = 0$). Previous work has been presented with respect to this conceptual framework: (a) Attempts at predicting mental health (e.g. [23, 88, 157, 236]) and a broader set of self-reported wellbeing metrics (e.g. [12, 13, 24, 88, 129, 142, 161]), (b) Detecting physiological signals such as pulse and breathing rate (e.g. [26, 97, 99, 184]) (c) Automated metrics of dialog quality (e.g. [4, 56, 143, 166, 182, 204]), (d) Interpretability by weighing sample importance (e.g. [120, 122, 128, 260]), (e) Interpretability based on saliency maps (e.g. [51, 150, 233, 237]), (f) Uncertainty quantification in deep learning settings (e.g. [58, 59, 118, 180, 216]), (g) Interpretability based on high-level concepts (e.g. [22, 76, 123]), (h) Interpretability through generative explanations (e.g. [38, 117, 209, 230]).

The first component is interpretation *of people* by machines. That means building machines that can accurately infer human state to better adapt to user preferences, be attuned to their needs, provide higher quality interactions, or help people better manage aspects of their state such as health and wellbeing by anticipating changes that might jeopardize their health. Across different domains ranging from interactions with a chatbot to depressive symptom management in outpatient clinical settings, one major challenge is how to make accurate inference without increasing the burden on the human by requiring them to explicitly self-report their evaluations. A possible solution is learning from their measurable physiological, behavioral, or implicit social cues that are ubiquitous and naturally available through their interactions with machines.

The second component is interpretation of machines *by people*. This means developing techniques that aid humans to investigate machine learning systems and translate complexities of decision boundaries into a language that humans can interpret and act on by identifying potential failure points and informing actionable directions for improvement. This can be done through a variety of mediums of explanation. It can range from providing numerical estimates of uncertainty [30] to tools that enable people to ask what-if questions [3]. One of the open questions in this space is the degree to which uncertainty quantification can help disambiguating sources of bias, such as annotator disagreement, data bias, or inherent task difficulty [118]. Another challenge is how to develop flexible interpretability tools that are better aligned with humans' cognitive processes such as how they justify decisions [3] and learn [7, 18, 250].

1.1 Inferring Human State

Among affective phenomena, detecting depressive symptoms is of utmost importance. Depression hampers cognitive processes such as attention, memory, perception, and decision-making [6, 9, 44, 47, 47, 136, 190] and is the leading cause of ill health and disability world-wide [252]. Estimating depressive symptoms from passive sensing using

wearable sensors and phone sensor data paves the way to tracking symptoms in-between clinical visits, predicting the course of illness, capturing variations of the disease over time, and providing just-in-time interventions to manage the disease.

There is promising evidence on the feasibility of inferring affective phenomena such as wellbeing, mental health, mood, and stress from a combination of physiological and behavioral data such as location patterns, surveys, smartphone usage, whether information, wearable electrodermal activity, accelerometer, and heart rate data [13, 23, 24, 88, 129, 142, 161, 236]. However, there remains a set of challenges. Most previous work has low detection accuracy, typically below 80% [12, 88, 142], which hinders its successful real-world deployment. Additionally, most of these machine learning solutions have been trained on data gathered through constrained conditions that don't resemble real-world scenarios and they have rarely been compared with clinically validated measurements of mental health.

Implementing capabilities of inferring humans' affective states can make technology more human-centered in day-to-day applications as well. Since affective states influence memory [84, 119, 136], perception [47, 98, 188, 255], attention [47], decision making [6], interpretation [9], creativity [33] and many of our cognitive processes [44, 190], taking them into account can lead to improved experiences by catering to human preferences and needs more effectively across applications. For example, a visual interface can present visual information differently depending on the user's affective state to ensure it is perceived [188, 255]; a todo list can suggest completing creative tasks more frequently when the user is in a positive mood [33]; an automated tutoring system can facilitate learning by taking into account the user's state and its interaction with memory formation and retrieval [43]; more generally, by inferring their affective states, machines can anticipate users' needs more accurately and be more effective across a spectrum of applications [190].

1.2 Inferring Human Feedback through Interaction

Inferring user opinion through implicit feedback has led to significant improvements in domains such as recommendation systems [177]. Collecting star ratings and gathering thumbs up/down button presses mostly known as explicit feedback has high quality; however, it is not as scalable as implicit feedback such as purchase history, browsing history, search patterns, or mouse movements. The sparsity and cost of explicit feedback ultimately renders it less useful than implicit signals in practice [106].

Such benefits can extend to affective and social cues that are ubiquitous, yet mostly an untapped resource. Humans are developed and socialized to communicate rich information through social cues, such as body language, facial expressions, and their tone of voice. Many of these social cues are not unique to human-human interaction, but also naturally arise when interacting with machines [201]. Being able to infer such signals could lead to improved systems that adapt to their users' needs and preferences, and learn from their feedback more efficiently and effectively.

One application area where harvesting these implicit affective cues can be of high impact is open-domain dialog. Current methods fail to produce key aspects of good quality conversation, such as staying on topic, not being repetitive, and generating emotionally appropriate responses. This is in part due to the widely used evaluation criteria that poorly capture conversation quality as judged by humans [145]. However, this is a setting in which humans naturally provide social cues in abundance. I show in this thesis that these implicit signals are an untapped resource and inferring such cues can significantly improve how we evaluate dialog systems.

1.3 Estimating Uncertainty of Machine Predictions

Uncertainty of predictions can arise for several reasons: measurement error, model parameters and structure, and the list goes on [58]. Quantifying and communicating the

uncertainty of automatic predictions is consequential for AI safety. It can potentially prevent a range of unintended failures across domains such as medical diagnostics and autonomous driving [53, 58, 164]. Communicating predictive uncertainty can help people interpret the context of predictions, anticipate when uncertainty might be irreducible, prioritize gathering more diverse data, changing the model, or soliciting other sources of information.

In common practice today, most deep learning models have been viewed as deterministic functions providing only point estimates. As eloquently reviewed in [61], probabilistic modeling, uncertainty, and Bayesian precursors to neural networks have existed for years (e.g. [102, 152, 174]). In the past few years, additional tools for practical uncertainty estimates to equip these models with Bayesian qualities have been proposed [58, 59, 118], leading to increased adoption of uncertainty estimation techniques for deep learning methods. The utility of these methods has been studied under dataset shift through empirical work [180] and compared to post hoc uncertainty [216]. However, there remain several open questions, especially in complex phenomena where annotators can legitimately disagree.

Consider the scenario of predicting facial expressions from a face image. Many different causes could result in low confidence predictions. A photo might be nuanced in its emotional expressions, making it even hard for humans to assign labels to it. In such cases, the subjectivity of annotations and their disagreement can lead to uncertain predictions by the machine, too. There might be occlusions in the image. For example, a face mask covering the mouth area makes it hard to identify the expression. The photo might be low quality. For instance, due to poor lighting conditions, it might be challenging to see the face, resulting in uncertain predictions. To what extent can we characterize and disentangle these sources of uncertainty, such as annotator disagreement and data bias and task difficulty, and how do we translate these into actionable directions for improvement?

1.4 Tools for Investigating Machine Predictions

Humans' ability to reason, imagine, draw insights from sparse interactions, and transfer their knowledge across contexts by far exceeds what our current computation methods are able to accomplish. While sharing machine predictions along with their predicted uncertainty is useful for human interpretation and intervention, it is not powerful enough to satisfy many of the questions raised by human insight and advanced cognition [3]. Developing computational tools for inspection and introspection in model development that are more aligned with humans' learning and decision making processes [3, 7, 18, 250] can be an impactful step toward a more human-centered approach.

It is argued that empowering humans with additional tools for investigation and interpretation of machine learning decision making processes is a promising venue that could lead to better outcomes in terms of scientific understanding, improving safety, uncovering hidden biases, evaluating fairness, and beyond [31, 45, 78]. Many efforts in machine learning interpretability methods have been working towards providing solutions for this challenging problem. One way to categorize them is by medium of explanations, some post hoc methods focusing on importance of individual features, such as saliency maps [51, 150, 233, 237], some on importance of individual examples [120, 122, 128, 260], some on importance of high-level concepts [123]. There has been active research into the shortcomings of explainability methods (e.g. [1, 111, 193, 224, 253]). For example, it has been shown that attention weights can be manipulated without hurting accuracy and result in misleading interpretations [193], adversarially constructed dissimilar attention distributions can lead to similar predictions [111], and some existing saliency methods are independent of the model and the data generating process [1] which renders them unfit for explaining the relationship between inputs and learned outputs. Scholars have also proposed tests to determine when attention can be used as an explanation [253].

All of these methods focus on mediums *that already exist* in the data–either by weighting features or concepts in training examples, or by selecting important training examples.

Guided by recent progress in generative models [27, 101, 124, 131, 147], another family of explainability methods have emerged that provide explanations by *generating* new examples or features [38, 117, 209, 230]. The goal of these methods is to use generated examples to highlight particular aspects or factors that contribute to classifier's decision or produce counterfactuals.

One of the key benefits of *counterfactual* generation is allowing users to explore "what-if" scenarios through what does not and cannot exist in the data, which makes them a great tool for making classifier decisions plausible [247]. Using counterfactual generation for investigating a classifier's decisions, one can ask: what if this sample were to be classified as the opposite class, and how would it differ? It has been argued that humans also justify decisions via counterfactuals [3], and children learn through a similar process [7, 18, 250]. Additionally, in-depth user studies have shown that examples have been the most preferred means of explanation by users across visual, auditory, and sensor data domains [115].

As I show in this thesis, current counterfactual explanation techniques fall short of simultaneously satisfying desired interpretability properties such as distinctness, compatibility, realism, substitutability, and stability. Distinctness [162] suggests that inputs should be representable with non-overlapping concepts. Compatibility with classifier [230] or classification model consistency [231] suggests that changing the explanation should produce the desired outcome from the classifier. Realism or data consistency [230] suggests that perturbed samples should lie on the data manifold to be consistent with real data. In other words, the generated samples should look realistic when compared to other samples. Substitutability suggests that explanations should preserve relevant information [209]. This quality has sometimes been referred to as fidelity [162, 192]. Stability [75, 162, 192] refers to the coherence of explanations for similar inputs. In this thesis, I introduce a method that addresses this challenge.

1.5 Thesis Roadmap

In **Chapter 2**, I present our work on approximating depressive symptom severity based on phone and wearable sensor data in an outpatient clinical setting. This chapter addresses several challenges mentioned in Section 1.1 by studying a real-world setting and validation against high-quality scores provided by clinicians. Collecting measurable real-world behavioral and physiological data allows for more scalable, accurate, and less burdensome symptom tracking and can overcome limitations of current office-based clinical interviews and self-reports in diagnosis and treatment of major depressive disorder. I show that using longitudinal data including electrodermal activity, heart rate and heart rate variability, motion, temperature, location patterns, social interactions, and phone usage, we achieve less than 8% error rate in predicting Hamilton Depression Rating Scale (HDRS) scores. I provide additional analyses identifying the most informative features regarding depressive symptomatology. In addition, we have investigated the association between depressive symptoms and several modalities in depth (e.g. location, incoming calls, sensor-recorded vs. self-reported sleep) to help make the human state reading more explainable.

In **Chapter 3**, I present our work on approximating human judgements of quality through inferring implicit signals in interactive text-based communications [63]. I introduce an automated metric that is better correlated with human evaluations than other alternatives and successfully address one of the challenges in open-domain dialog as mentioned in Section 1.2: poor correlation between automated evaluation metrics and human judgements [145]. While open-domain dialog is sometimes referred to as “non-goal oriented dialog"[212], many argue that it does serve a goal: responding to the human need for connection, affection, and social belonging [109]. Additionally, it could pave the way for more seamless language learning tools or computer game characters [222]. I identify limitations of static evaluation and provide evidence for added benefits of interactive human evaluation; I then introduce a novel, model-agnostic, and dataset-agnostic method to approximate it. While state-of-the-art methods predominantly use word-level

overlap [4, 143, 182] or distance metrics based on embedding resemblance to each turn of the conversation [56, 166, 204], we propose to take into account the trajectory of the conversation and use proxies such as sentiment, semantics, and user engagement that are psychologically motivated [11, 79, 95, 108, 110]. In particular, I propose a self-play scenario where the dialog system talks to itself and calculate a novel metric that is the combination of the aforementioned proxies. I show that this metric is better aligned with the human-rated quality of a dialog model than other automated metrics commonly used today [8, 56, 166, 204, 226], as measured by Pearson correlation. We perform extended experiments with a set of models, including several that make improvements to recent hierarchical dialog generation architectures through sentiment and semantic knowledge distillation on the utterance level.

In **Chapter 4**, I investigate uncertainty quantification and its connection to sample difficulty, data bias, and annotation disagreement to address the questions posed in Section 1.3 regarding the decomposition of sources of uncertainty and deriving actionable directions from them. We characterize interpretability benefits of uncertainty quantification for complex phenomena where annotators can legitimately disagree, such as facial expression identification [62]. Prior work includes numerous attempts to model annotators, their biases, and skill levels in crowd-sourcing literature to use labels more effectively and efficiently. (e.g. [121, 170, 200, 240, 251]). However, to what extent a simple uncertainty quantification technique can provide such insights is poorly understood. We demonstrate how adding Monte Carlo dropout to a classical network provides measures of uncertainty and helps disambiguate data bias and inter-rater disagreement. We confirm that this characterization also provides a proxy for Brier Score, a measure for the accuracy of probabilistic predictions [15].

In **Chapter 5**, I present a novel method that allows humans to investigate the decision making process of a predictor [64] and tackle the challenges discussed in Section 1.4: How to develop a method that simultaneously satisfies desirable qualities of an explanation tool and anticipates what changes in the input might lead to certain changes in the predictor's

output. Such a method allows counterfactual reasoning by answering what if an input sample were to be classified as the opposite class by the predictor-under-test. The proposed technique generates *Concept Traversals* (CTs), which are defined as a sequence of generated examples with increasing degree of concepts that matter for a classifier's decision. CTs are generated by jointly training a generator, a discriminator, and a CT disentangler, together to generate examples that (1) express one factor at a time that is influential to a classifier's decision that are distinct from each other, (2) are coupled to the classifier's reasoning due to joint training (3), are realistic (4), preserve relevant information, (5) and are stable across similar inputs. I compare our method against several baselines, of which some have been optimized for disentanglement, some are extensively used for explanation, and some fall in between. Our method is the only technique that performs well across all dimensions. I evaluate our method using these datasets: `3D Shapes` [19], `CelebA` [146], and a new synthetic dataset inspired by the challenges faced in dermatology domain [65]. I also discuss applications of this work for detecting potential biases of a classifier, investigating its alignment with expert domain knowledge, and identifying spurious artifacts that impact predictions using simulated experiments.

Finally, **Chapter 6** concludes by summarizing the contributions of this thesis and proposing future research directions.

Chapter 2

Estimating Depressive Symptoms from Patient's Physiological and Behavioral Data

Depression is the major cause of years lived in disability worldwide; however, its diagnosis and tracking methods still rely mainly on assessing self-reported depressive symptoms, which originated more than fifty years ago. These methods usually involve filling out surveys or engaging in face-to-face interviews. They are costly to track and scale and provide limited reliability and accuracy in predicting treatment response, remission, and relapse. Broader anatomical and neurophysiological understanding of emotion, behavior, and cognition and their disorders could lead to finding biomarkers that are scalable and have improved reliability and accuracy in disease prognosis. In this chapter, we develop and test the efficacy of machine learning techniques applied to objective data captured passively and continuously from E4 wearable wristbands and sensors in an Android phone to predict the Hamilton Depression Rating Scale (HDRS). Input data include electrodermal activity (EDA), sleep behavior, motion, phone-based communication, location changes, and phone usage patterns. We introduce our feature generation and transformation process, imputing missing clinical scores from self-reported measures, and predicting depression

severity from continuous sensor measurements. While HDRS ranges between 0 and 52, we were able to impute it with 2.8 RMSE and predict it with 4.5 RMSE, which are low relative errors. These error rates are calculated on a hold-out set of clinician-rated HDRS scores. Analyzing the features and their relation to depressive symptoms, we found that poor mental health was accompanied by more irregular sleep, less motion, fewer incoming messages, less variability in location patterns, and higher asymmetry of EDA between the right and the left wrists.

2.1 Introduction

Depression is the leading cause of ill health and disability worldwide: According to the latest estimates from WHO, more than 300 million people are now living with depression, an increase of more than 18% between 2005 and 2015 [252]. Historically, diagnosing and tracking depressive symptoms has been accomplished through periodic assessment with structured or unstructured clinical interviews using standardized symptom rating scales. This approach, which was invented in the 1960s, is based largely on subjective self-report, and has limited utility in fully characterizing clinically meaningful subtypes of depression. Also, this current “descriptive" way of diagnosing depression is limited in its ability to predict the course of illness or to capture variations of the disease over days.

An important paradigm shift is happening today: Psychiatry and the clinical neurosciences are moving from relatively narrow neurochemical models of disease, based on inferences about the pharmacological mechanisms of available psychotropic medications, to broader anatomical and neurophysiological understanding of emotion, behavior, cognition and their disorders [91]. This shift is important, not only because it provides a new understanding of the neuroscientific basis of psychiatric disorders, but also because it leads to the development of novel strategies for diagnosis and assessment. Researchers are increasingly developing objective mobile data-driven biomarkers for many healthcare conditions, including depression (e.g. [133]). We anticipate that the development of reliable

biomarkers will help improve the diagnosis and assessment of depression, prediction of treatment response, and early detection of response, remission and relapse. To date, there is no set of reliable biomarkers to assess depression.

In this work, we advance the state of the art in the development of biomarkers by providing a new way, based on passive sensing, to estimate depressive symptoms as measured by the Hamilton Depression Rating Scale (HDRS). The method utilizes data from E4 wearable sensors [49] and embedded sensors within an Android phone. Experience sampling, continuously capturing self-reported depressive symptoms, can be overwhelming for a patient in the long-run. Being able to estimate HDRS scores accurately using passive data could potentially improve the scalability of depression prognostication as well as its objectivity. In the meantime, it enables a fertile ground of research for providing timely interventions to individuals who show signs of relapse. Also, we believe that there is more value in a regression analysis as opposed to a classification between different severity levels of depressive symptoms. With regression, we may obtain a more accurate and precise understanding of the progression of the disease.

In our dataset, HDRS has been captured bi-weekly by a clinician, as part of their standard practice. Thus, we utilize a two-step prediction process: First, we use a surrogate (self-reported data) to predict HDRS and in doing so, impute the missing HDRS values (from the dates when the HDRS was not assessed by a clinician, approximately 13 values in between two consecutive visits) to construct an increased dataset “HDRS-I". Second, we use the passive phone and wearable sensor measures for predicting the HDRS-I values.

2.2 Background and Related Work

Over the past decade, affective computing researchers have utilized wearable sensors and phone usage patterns to detect stress, happiness, and mental wellbeing (e.g. [114, 156]). We hypothesize that similar underlying phenomena quantifying mood can help assess mood disorders as well.

Numerous researchers have demonstrated the use of mobile-based Experience Sampling Methods to monitor people's depression, e.g. [169, 241]. In these studies, the depressed patients are asked to fill out regular surveys about mood, behavior, sleep etc. on their mobile phones. The self-reports have several limitations. They can be unreliable as the response rate may depend on the current mood of the patients. Moreover, they are subjective since such logs are recorded by the patients themselves and the answers may vary with factors including mood, weather, social-demands, or patient's memory. Finally, frequently answering the mobile surveys is cumbersome, which may introduce bias or result in reduced adherence.

Several studies have proposed to measure passively objective parameters in controlled environments (hospital or laboratory). One of the first efforts to assess how long-term physiology and behavior of individuals are correlated with changes in depression was the LiveNet project [238]. The LiveNet platform, which monitored skin conductance, heart rate, activity and voice, was evaluated on six psychiatric inpatients. More recently, Valenza et al. [244] demonstrated the use of electrocardiogram and respiration signals collected in a hospital to assess depression. Although these studies show promising results, we aim at a harder problem: to continuously and unobtrusively monitor people during daily life in order to identify possible biomarkers of depression.

The MONARCA project [159], which developed tools for assessment and prediction of mood episodes in bipolar disorder, focused on analytics tools and validating them with a group of 20 patients. Also, scholars have studied phone usage correlates of mental health and depressive symptoms (e.g [194, 205]). Other researchers have looked at audio/visual cues including facial expressions, head movement, vocalization, and vowel production to predict depression severity (e.g. [41, 214, 245]). However, many of these studies have been validated based on self-reported standard depression scales, like Patient Health Questionnaire (PHQ-9) [130], rather than on clinical measurements. In this work, we aim to fill the gap by including clinical assessment of depressive symptoms using Hamilton Depression Rating Scale (HDRS) as scored by the expert clinician in a patient interview.

The clinical form of HDRS data is collected in a face-to-face meeting bi-weekly as it has been demonstrated that intensive assessment of depression may have a positive impact on the assessment score [16]. We then impute the depression level of the remaining dates using Machine Learning that incorporates daily patient self-reports.

Most previous work has addressed a classification problem, usually binary, within this area [89, 158]. Some captured only categorical label variables, while others transformed an inherent regression problem into a classification problem, and in doing so relaxed the problem; for example, they only included the highest and lowest values of the depression range and did not address the “middle". However, depressive symptoms change continuously and which way they are shifting is important. To better understand and prevent worsening of depression, it is not enough to distinguish between extremely severe and extremely mild depressive symptoms: We aim to measure progressive change of symptoms in order to enable just-in-time interventions before depression becomes severe.

2.3 Study Protocol

Patients diagnosed with MDD from Massachusetts (n=12) completed an 8-week protocol. Participants included 9 females and 3 males from white, hispanic, african-american, and asian races and aged between 20 and 73 years old (mean=37, std=17). The protocol involved tracking depressive symptoms and mobile phone usage. Movisens [172] was used to measure incoming and outgoing text messages and phone calls, location, app usage, and screen on/off behavior. Patients also wore Empatica E4 wristbands [49] that recorded accelerometer data and electrodermal activity 23 hours a day. Measurements were processed to obtain daily aggregate measures. Participants were clinically assessed for depression symptoms biweekly using the HDRS. Tab. 2.1 summarizes the number of observations for each modality. In the next section, we explain the detailed measurements in each modality and the feature generation.

Modality	# of Datapoints
Physiological signals	540
Phone passive usage data	605
Interactive surveys	503
Clinical measures	59

Table 2.1: Dataset summary after computing daily features.

2.4 Feature Architecture

2.4.1 Physiological Signals

E4 sensors worn on each wrist captured continuous electrodermal activity (EDA) via the measurement of skin conductance (4Hz sampling rate), temperature (4Hz sampling rate), and 3-axis accelerometer data (32 Hz sampling rate). In order to better understand the user's behavior within the day, we introduce 6-hour intervals, labeled as morning, afternoon, evening, and night. The 6-hour interval provides a balance between granularity and ratio of missing values. We also calculate aggregate daily measures. Any feature explained below has been calculated for all these intervals.

We first filtered out the EDA signal when the corresponding skin temperature was below 31°C to exclude the measurements when the sensor was not worn. Then we applied the 6th order Butterworth low-pass filter (1Hz cutoff frequency). We calculated mean EDA and the fraction of time the sensor was recording the signal. We also computed the number of skin conductance response (SCR) peaks and their average amplitude using the method from Gamboa [60]. There are indications that skin conductance level may distinguish between depressed and healthy individuals [248]. Also, previous research has shown that asymmetry in EDA between the wrists can provide extra affective information [191]. Thus, we also encoded asymmetry in different ways: the difference between average EDA value, difference between number of SCRs, and difference between SCL and SCR signals using a convex optimization approach recommended by [87].

We applied the 5th order Butterworth low-pass filter (10Hz cutoff frequency) to the

accelerometer data. We then translated the output into motion features by calculating the vector magnitude for the tth sample, V_t, of the z-axis acceleration data using the following formula:

$$V_t = \sum_{i=t-N}^{t} |R_{(z,i)} - M_{(z,i)}| \tag{2.1}$$

where $R_{(z,i)}$ is the raw z-axis (perpendicular to the skin) acceleration of ith sample, $M_{(z,i)}$ is the running mean in a 5-second window of the z-axis signal preceding the ith sample, and N is the number of raw data samples received in one second.

Next, we calculated average, median, and standard deviation of motion for the mentioned time intervals as well as the fraction of time in motion. We also kept meta-data such as the fraction of time within the time interval that the data were not missing.

We calculated objective sleep based on accelerometer data for 30 second epochs using the ESS method described in [14]. We calculated sleep duration, sleep onset time (time elapsed since noon), maximum duration of uninterrupted sleep, number of wake-ups during the night, and the time of waking up (time elapsed since midnight). We also computed a sleep regularity index (SRI) [28]:

$$SRI = \frac{1 + \frac{1}{T-\tau}\int_0^{T\tau} s(t)s(t+\tau)dt}{2} \tag{2.2}$$

where data were collected for $y = [0, T]$, $\tau = 24$, $s(t) = 1$ during sleep and $s(t) = -1$ during wake. The SRI ranges between 0 (highly irregular sleep) and 1 (consistent sleep every night). We also included meta-data such as the fraction of time that data were being recorded over nighttime (between 8pm-9am) as well as over the period of 24 hours.

2.4.2 Phone Passive Usage Data

We utilized Movisens [172] on Android to collect measures of how the participant is using his or her mobile phone and how s/he is interacting with other people using the mobile phone. More specifically, we captured meta-data of calls, text messages, app usage, display on/off behavior, and location. Passive data were captured 24/7. The content of the calls/texts, actual phone numbers, websites visited, and the content of the applications were not collected.

Following the steps of previous researchers in generating features from passive phone data [114], we introduce 3-hour intervals in order to better understand the user's day-time behavior. For example, [6am-9am] represents early morning while [9pm-12am] corresponds to late evening. We also calculate aggregate daily measures.

For quantifying call data, we calculate the number of incoming, outgoing, and missed calls daily and over the 3-hour periods within the day. In a similar manner, we calculate mean, median, and standard deviation (SD) of the duration of incoming, and outgoing calls. Finally, we calculate the incoming/outgoing ratio both for the number of calls and the duration of calls on a daily basis.

For quantifying SMS data, we use a similar approach, we calculate the number of incoming and outgoing texts daily and over 3-hour periods within the day. We also calculate a daily incoming/outgoing ratio of the number of text messages received or sent respectively.

Turning the display on/off is also an indication of phone usage. Thus, we look at the mean, median, and SD of duration of screen on within the mentioned intervals. We also calculate the number of the times the screen has been turned on over these periods. Note that these two correspond to different behaviors; Long screen-on duration is related to actively using the phone while a great number of screen-ons is related to consistently checking the phone which might be a sign of anxiety or anticipation.

For location data, we calculate mean, median, and SD of latitude and longitude along

with the number of data points that have been captured for each time period. We calculate total location mean, median, and SD by averaging values from latitude and longitude.

For app usage, we encode the app category using the following list: game, email, web, calendar, communication, social, maps, video streaming, photo, shopping, and clock. Then, we calculate the total duration and the number of app category usage in the different mentioned time intervals.

2.4.3 Interactive Surveys

Using the Movisens [172] on the mobile phone, we administer short questionnaires about overall health condition, sleep, mood, stress, anxiety, alcohol/drugs/caffeine usage, and social interaction; these should be completed each day upon awakening, at bedtime and twice during the day at random times, during the entire length of the study. For assessing mood, we have used Positive and Negative Affect Schedule (PANAS)[249], one of the most prevalently used scales for measuring affect. The 20 item questionnaire has been split into two 10-item questions that were administered twice during the day at random times. Each question is a five-point Likert scale, one indicating not at all and five indicating extremely. The average of two mid-day scores constitutes the daily PA and NA scores. The minimum possible value is five, and the maximum possible value is 25.

First, we preprocessed the data: we added how long it took the participant to fill in the survey and removed responses that took less than a second and are likely noise. This meta-data can also be informative; for example, long pauses while responding to surveys may represent motor slowing (a common symptom of depression), cognitive load, trouble remembering, or not being sure about the response. Short response time, on the other hand, may represent trivial answers or not reading through the questions. We calculate total alcohol (standard drink measure) and caffeine consumption (milligram) by summing the relevant features from the survey. We convert categorical features to their one-hot representation. We include day of the week as it has been shown to influence the aggregate

number of smiles which can be an indication of positive valence mood [96].

Since HDRS is closely related to self-reported mood, we add more detailed mood information. First, we calculate total positive affect (PA) and negative affect (NA) on a daily basis by averaging responses to relevant survey questions. We include an average of the past week's PA ($\sum_{i=t-7}^{t} PA_i/7$), and NA ($\sum_{i=t-7}^{t} NA_i/7$). We also include a weighted average of PA, when the effect of affect diminishes exponentially overtime when going back in history, e.g., yesterday's mood is half as important as today's mood in the weighted average measure: $\sum_{i=t-7}^{t} 2^{i-t} * PA_i$. We included a similar feature for NA, as well: $\sum_{i=t-7}^{t} 2^{i-t} * NA_i$.

We calculate the NA/PA ratio for the daily [1], average weekly, and weighted average weekly measures. To capture mood oscillation, we include the standard deviation of mood both PA and NA on a weekly basis and for the duration of the study.

2.4.4 Clinical Measures

During each biweekly visit, participants are assessed by the clinician for depressive symptoms using the HDRS. HDRS is a standard test for quantifying depressive symptoms which ranges between 0 and 52. Tab. 2.2 summarizes the depression severity in relation to HDRS.

HDRS	**Depression Severity**
0-7	Normal
8-13	Mild Depression
14-18	Moderate Depression
19-22	Severe Depression
$\geq$23	Very Severe Depression

Table 2.2: HDRS values and levels of depression severity.

[1] The minimum possible value for either PA or NA is five. Thus, division by zero would never happen.

2.5 Models

2.5.1 Feature Transformation and Selection

Combining the carefully-crafted features results in over 700 features for our dataset. Compared to the small number of data points we have, this number of features can easily result in over-fitting the model to the training set. One possibility is to use regularization tricks such as L1 to enforce selection of only a small number of features. However, for features that are non-linearly related, transforming the features into a new space through a non-linear transformation can be more beneficial. For example, several noisy measurements of a similar phenomena may not be informative on their own, but a transformed version of them can be a better predictor. Toward this end, we tested PCA, kernel PCA with radial-basis function kernel, and truncated SVD methods to reduce the dimensionality of our feature-set. We bound the number of selected features while keeping as few features as possible to explain the variance of data.

We created 3 datasets: one including all features (daily and over multiple intervals mentioned above), one including daily features, and one including the daily features concatenated with the features of the previous day. We conducted the feature transformations on these three datasets.

2.5.2 HDRS Imputation Based on Survey Data

Several studies have confirmed relationships between self-reported affect and clinical ratings of depression. In our dataset, we see a strong correlation between average weekly negative/positive affect ($M = 0.86, SD = 0.38$) and HDRS scores ($M = 19.64, SD = 7.60$), $r = 0.70, p = 0.00, n = 44$[2]. This observation suggests utilizing self-reported survey data to estimate the gold-standard measure in between clinical assessments. We impute these values and refer to them as HDRS-I to distinguish them from clinician-rated

[2]Data points with missing mood reports from surveys have been removed from this analysis. This reduces the number of data points from 59 available HDRS measurements to 44 for this section's analysis.

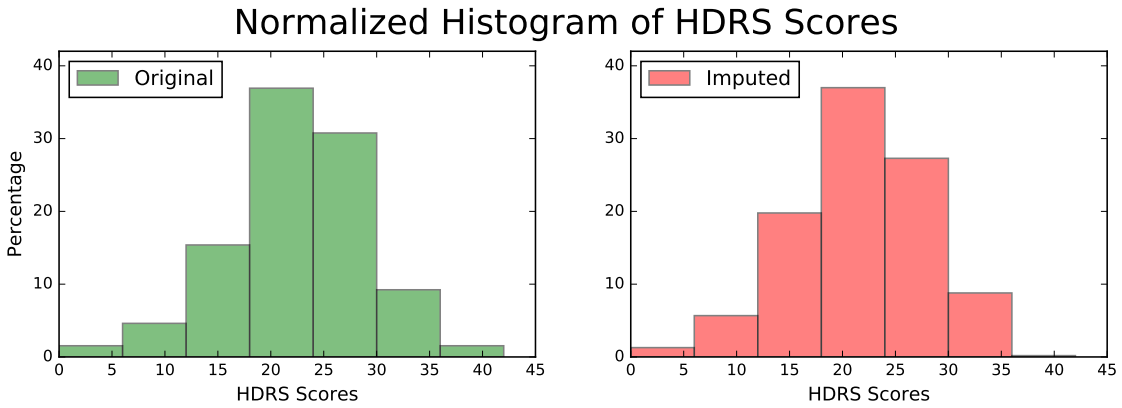


Figure 2-1: Normalized histogram of the original HDRS scores (left-green) and the imputed HDRS-I scores (right-red).

HDRS scores. The input features included daily PA, NA, and NA/PA ratio. We have also included average and standard deviation of these values over the past week and over the whole period of the study for the patient. Also, weekly weighted average of these values have been included where the effect of affect diminishes exponentially over time. We then impute the missing values to construct an almost 10-times-larger dataset [3]. We experiment with two methods to predict the HDRS score from survey data: regularized regression and robust-to-outlier methods. After choosing the best-performing model, we employ it for imputing HDRS-I scores used for training the model discussed in the next section.

Regression Models

The regression methods include lasso, ridge, and elasticNet which use L1, L2, and a combination of the two as regularization metrics, respectively. Note that the L1 regularization term acts as a feature selection mechanism by pushing coefficients of most of the variables to be exactly zero, while L2 pushes many coefficients to near zero values but does not remove them completely. We also included regression without regularization with the reduced and transformed features.

[3]Ideally, it would results in a 14-times-larger dataset. But due to missing values and clinical visits not happening exactly every 14 days, this step results in a almost 10-times-larger dataset.

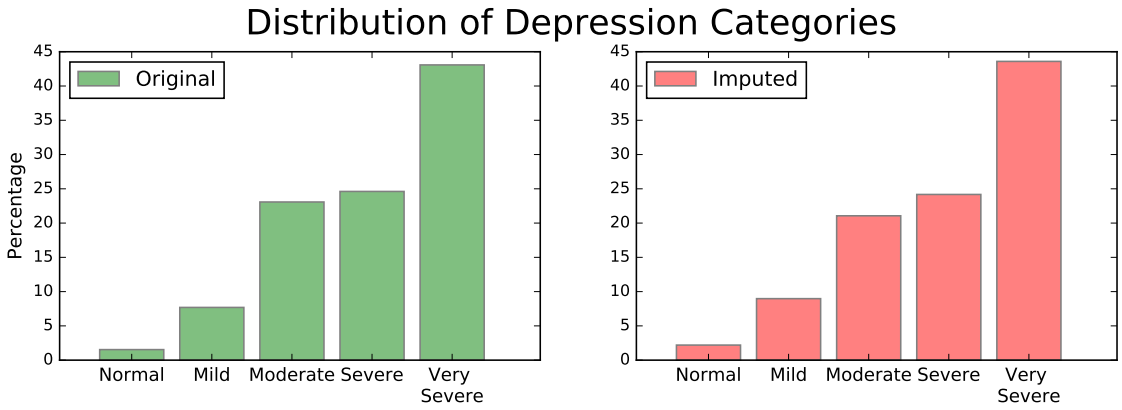


Figure 2-2: Distribution of depression categories based on original HDRS scores (left-green) and imputed HDRS-scores (right-red).

Robust Models

To be robust against outliers or errors in formulation of the model, we include Theil-Sen estimator, random sample consensus (RANSAC), and huber algorithms. These models have a built-in sampling procedure that allows a fraction of data points to be outliers.

Validation

For validation, we split the data into 90% training and 10% testing. We use leave-one-datapoint-out cross-validation on the training set to select the best model and use it for imputing missing HDRS values.

2.5.3 HDRS Prediction Based on Sensor Data

After imputing HDRS scores, the new dataset HDRS-I contains over 500 points. In this prediction phase, we train a model using HDRS-I data-points and test it on original HDRS scores. This dataset is still not large enough to be able to benefit from state-of-the-art neural network techniques. For example, long short-term memory (LSTM) network, a strong model that retains temporal information, performs as well as predicting the average value. We ran LSTM on the dataset as well as an augmented version of it. For augmentation, we

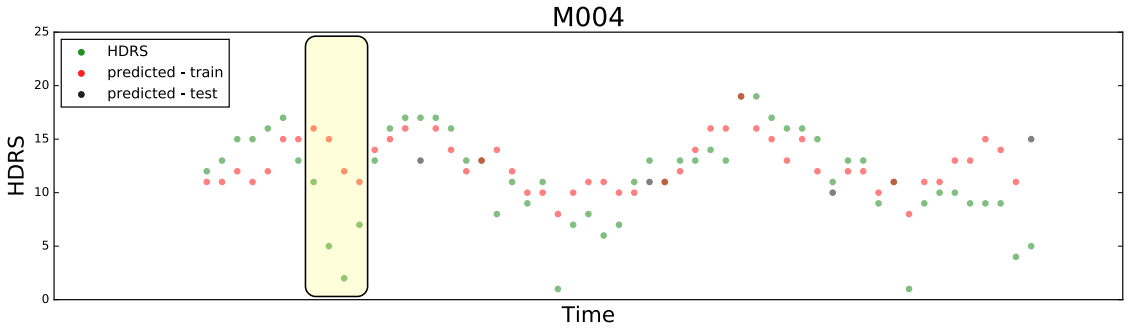


Figure 2-3: Time-series of Original (HDRS), imputed (HDRS-I), and predicted (HDRS-P) scores for one sample user over eight weeks. For simplicity, both HDRS and HDRS-I are shown in green. HDRS-P is shown in red and black for training on HDRS-I values and testing on HDRS, respectively.

have added $x * 0.01 * SD_f$ to each feature f where x is a random number between -0.5 and 0.5 and SD_f is the standard deviation of the values for that feature. Thus, we focus on models that do not require enormous amounts of training data. Note that self-reported affect measures have been used only in the imputation phase (HDRS-I) and are excluded from the current prediction step (HDRS). The HDRS prediction phase solely uses the passive wearable and phone sensor data.

Regression Models

Similar to the imputation phase, we use lasso, ridge, elasticNet, and unregularized regression. The following regularization coefficients have been considered: [0.1, 0.5, 1.0, 5.0, 10.0].

Robust Models

Similar to the imputation phase, we use Theil-Sen, RANSAC, and Huber methods. However, we loop through a larger list of parameters to optimize within each model. For RANSAC, we consider ratio of minimum samples in the range of [0.1, 0.2, 0.3, 0.4, 0.5]. For Huber method we consider epsilon values in the range of [1.0, 1.35, 1.5] and alphas in the range of [0.0001, 0.001, 0.01, 0.1, 0.5, 1.0, 5.0, 10.0]. We should note that these

				RMSE		Baseline	
Model Type	**Model**	**Parameters**	**Dataset**	**Validation**	**Test**	**Average**	**Median**
Regression	Regression		Kernel PCA subset	5.2	4.9	7.1	7.1
Robust	Ransac	ms=0.3	Kernel PCA subset	5.0	4.9	7.1	7.1
Boosting	AdaBoost	n=50, lr=1	Subset data	5.5	4.6	7.1	7.1
Random Forest	-	n=15	Subset data	5.4	4.6	7.1	7.1
Gaussian Process	-	α=0.1, n=5	Kernel PCA subset	5.3	5.5	7.1	7.1
Overall Ensemble		k=1	selected by individual models	5.8	4.5	7.1	7.1

Table 2.3: Best prediction model. For the values of hyperparameters used in these experiments, refer to the main text.

models do poorly when the feature set is large. Thus, we only use them for the subsets or the reduced version of the data.

Boosting

Boosting combines weak regressors sequentially to improve performance. We use adaptive boosting (AdaBoost) and Gaussian boosting in this category. We experimented with several hyperparameters such as the number of estimators in the range of [25, 50, 75, 100, 500], learning rates in the range of [5.0, 1.0, 0.1, 0.001, 0.0001], and linear and squared losses.

Random Forest

Random Forest is an ensemble method with multiple decision trees. We experimented with different numbers of estimators in the range of [5, 10, 15, 20, 25] for our Random Forest regressor.

Gaussian Process

We use a Gaussian Process with RBF kernel with length scale 1.0, different regularization parameters in the range of [1e-10, 1e-8, 1e-6, 0.0001, 0.001, 0.01, 0.1, 0.5, 1.0] and different numbers of restart points in the range of [5, 10, 50, 100] to model the data.

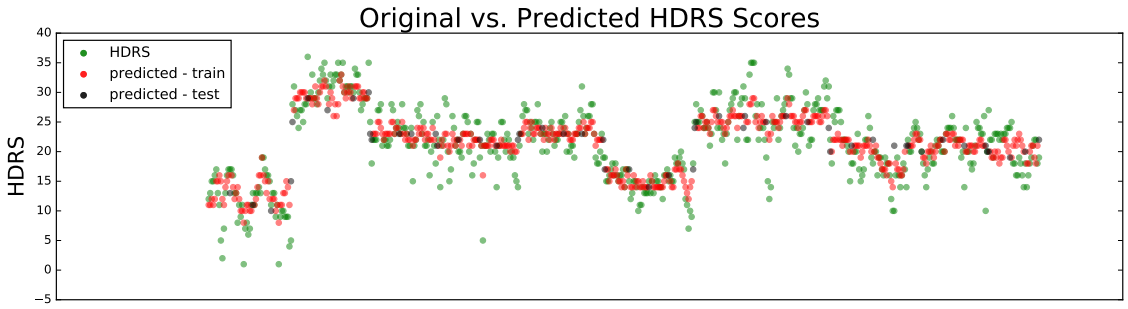


Figure 2-4: Original (HDRS), imputed (HDRS-I), and predicted (HDRS-P) scores for daily data from all patients over eight weeks. For simplicity, both HDRS and HDRS-I are shown in green. HDRS-P is shown in red and black for training on HDRS-I values and testing on HDRS, respectively.

Customized Ensemble Method

Finally, we combine the results from these different regressors to get a more robust estimator. The ensemble method first finds a set of k nearest neighbors from the training set for each point. It then chooses the model that performs best on that set as the estimator for this point. The heuristic behind this method is that slight modifications in the feature set do not change the output drastically. Thus, if a classifier is working well on similar points, chances are it works well for the current point, as well. Looking at k nearest points as opposed to only the most similar point is for smoothing purposes. Note that as the points become higher dimensional, the distance between them becomes less meaningful in explaining similarity between the points. Thus, we only use the first 5 reduced features based on kernel-PCA and create a KD tree and find the k nearest neighbors to the point at hand.

Validation

In real life, some depressed patients see a doctor and get clinical assessments at some point in their life. One major issue is a high relapse ratio and not being able to regularly visit the doctor to re-assess the improvement or worsening of depressive symptoms. In such cases, our method could be easily deployed in real life to passively monitor the patients after the

diagnosis. Thus, we will assume that we have at least some history for each user.

We use the imputed HDRS-I scores for training and reserve the original HDRS values for the hold-out test set. Additionally, to mimic the real-life deployment scenario, we do not include any ratings from the first two weeks in the test set. We use 10-fold-cross-validation on the training set to select the best model and predict HDRS values.

2.6 Results and Discussion

2.6.1 Imputation Phase

Root mean squared error (RMSE) is the primary metric used to validate the imputation phase. Table 2.4 shows the selected best model based on having the lowest RMSE on the validation set. Then we report the RMSE on the hold-out test set for each model. This model is ridge regression on the subset of mood features from the survey data , obtaining a test RMSE of 2.8. A baseline prediction of reporting the average or median HDRS score results in an RMSE of 6.8.

Looking more closely at the model provides insights into how the mood features correspond to the HDRS score. Consider the coefficients with the highest absolute values: The coefficient for weekly average positive affect is -9.3, confirming that reported positive affect is negatively associated with HDRS score. Another interesting observation is the -7.4 coefficient of the standard deviation of positive mood in the previous week.

	Name	Ridge (L2-Regularized Regression)
Model Information	**Dataset**	Mood Subset (PANAS)
RMSE	**Validation**	3.4
	Test	2.8
	Baseline 1 (Average)	6.8
	Baseline 2 (Median)	6.8

Table 2.4: Best performance for HDRS imputation on validation and hold-out test sets as measured by Root Mean Square Error (RMSE)

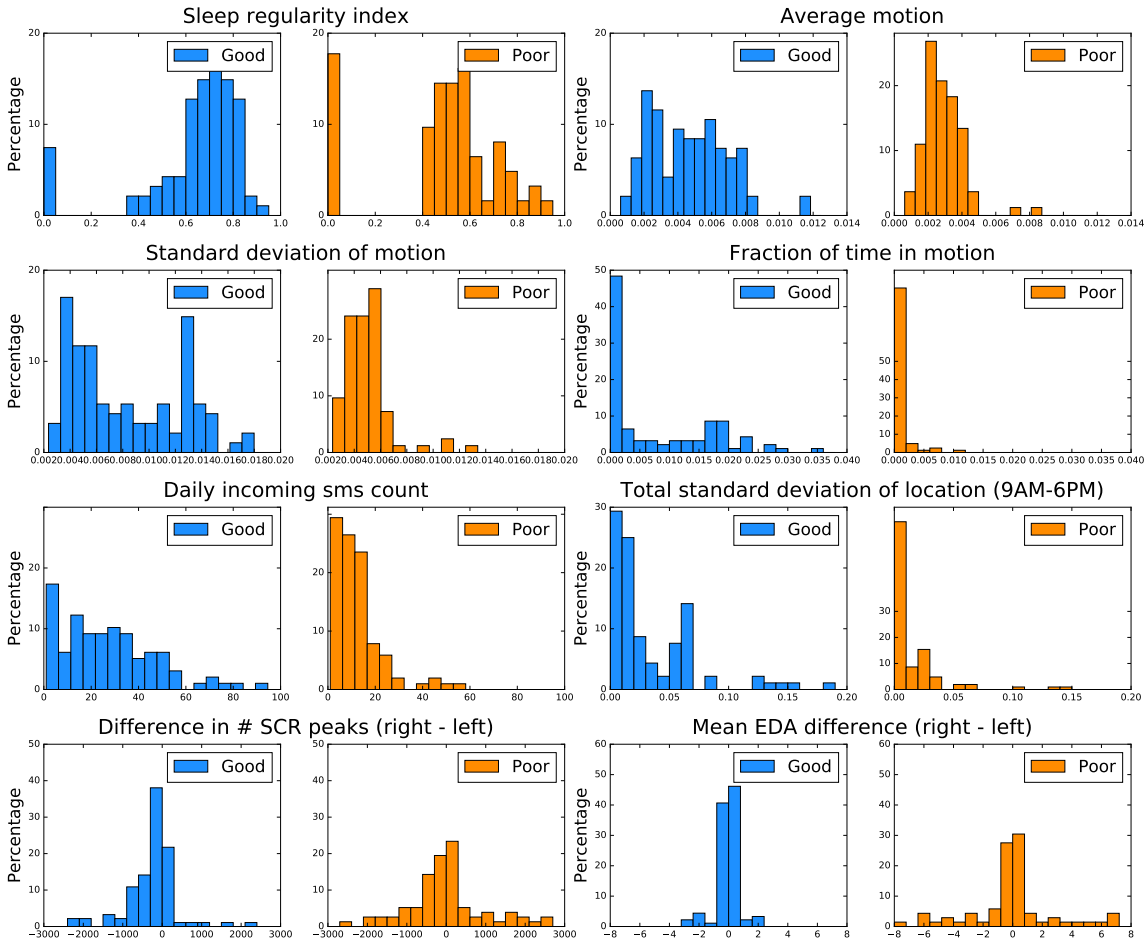


Figure 2-5: Distribution of features that are significantly different between days with good vs. poor mental health.

Depression is usually accompanied by anhedonia, withdrawal, and loss of engagement, resulting in a consistently low positive mood. Thus, a normal variation in positive mood is negatively associated with HDRS scores. This observation is aligned with previous work on characterization depressive symptoms [175]. At the same time, we see positive association between the average weekly negative affect and the HDRS score, shown by a positive 2.8 coefficient.

To further test the validity of the imputation model, we plotted the distribution of HDRS scores before and after the imputation (Fig. 2-1), and we used the Kolmogorov-Smirnov (KS) test to compare these two distributions. KS could not reject the null hypothesis of

samples coming from a common distribution. [4] Moreover, we examined the predicted levels (based on Table 2.2) of depression severity before and after imputation. Fig. 2-2 shows the bar chart of the distribution of depression severity categories. We also tested these two discrete-valued distributions and found they were not significantly different[5].

2.6.2 Prediction Phase

We primarily validate the new prediction model using RMSE. Table 2.3 shows the best performing model in each category and the overall customized ensemble method. The test RMSE for the ensemble method is 4.5 while it is 7.1 for the average or median baseline prediction.

To provide understanding of the predictions, we have visualized the time-series of HDRS-I values for a sample user (Fig. 2-3). Each point represents the HDRS-I value for a day. Green diamonds shows original values (either through clinical assessment or imputation). Red circles and gray triangles show the predictions for train and test points respectively. One interesting observation about this plot is the large prediction error in the highlighted area. A clinician we work with suggested it might be due to the placebo effect of being in the study. Many patients begin to feel better soon after joining the study, and report this, but they fall back into their depressed trend after the novelty effect wears off. We hypothesize that the placebo effect influences momentary assessment of mood quickly, while it is not adequate to influence behavioral or physiological signals . Thus, we see the red dots showing that the prediction based on the objective passive data, while it improves a little, does not improve as much as the self-reported (or their imputed) values. Fig. 2-4 visualizes the predicted and original values for all the data points from different users with the same color coding. As shown in both figures, the predictions follow the overall trend very well but miss the short term variations of HDRS-I. We should note that

[4] $D_{original}(M = 21.5, SD = 6.4), D_{imputed}(M = 21.2, SD = 6.3); ks - statistic = 0.08, p = 0.83$. The small ks-statistics and large p-value show that we cannot reject the null hypothesis.

[5] $ks - statistic = 0.01, p = 1.00$

HDRS is meant to measure depressive symptoms over the course of two weeks. Thus, from a clinical perspective, it is not supposed to vary much over consecutive days.

The final prediction based on the ensemble algorithm is a combination of different methods and sometimes non-linear feature transformations of the “subset data”. To gain deeper understanding of the relationship between the feature space and the resulting predictions we create two classes of points: the top 20% and the bottom 20% of the predicted HDRS-I scores. The former group represents days when the patient is doing very poorly and the latter represents the days when the patient is doing well or showing minimal depressive symptoms. We have compared the distribution of all the features from the “subset data" for these two groups using the KS test. Table 2.5 summarizes the 8 most significantly different distributions (highest ks-statistics and lowest p-values) and Fig. 2-5 depicts the differences where blue and orange show the good and poor mental health group respectively. The poor mental health group has more irregular sleep, moves much less on average, shows less motion variability, and is active a lower percentage of the time. Also, this group receives fewer incoming messages and has less variable location patterns. Another interesting finding is the EDA asymmetry. The number of skin conductance responses (SCR) between left and right wrist are mostly similar in the good mental health group. However, we see stronger asymmetry (more SCR peaks on the right wrist) for the poor mental health group. A similar trend is observed in average EDA magnitude.

Category	**Feature**	**ks-statistic**	**p-value**
Sleep	Sleep regularity index	0.51	$2e-9$
Motion	Average motion	0.49	$3e-10$
	SD of motion	0.47	$3e-9$
	Fraction of time in motion	0.44	$4e-8$
Communication	Daily # incoming SMS	0.44	$3e-9$
Location	Total SD of location (9AM-6PM)	0.34	$8e-6$
Physiology	Difference in #SCR peaks (right-left)	0.29	$8e-4$
	Mean EDA difference (right-left)	0.21	$4e-2$

Table 2.5: Most significantly different distributions of feature values for days with good vs. poor mental health.

These analyses are based on data from 12 participants from Massachusetts. Further studies are needed to confirm if the findings are generalizable to other populations, as well.

2.6.3 Limitations and Future Work

HDRS by definition is a biweekly measure, not intended to be captured daily. However, to be able to utilize the clinician-based ratings, we had to increase our dataset size by imputing daily HDRS values. Thus, theoretically, imputed values capture an aggregate measure for overlapping periods of time and are not independent.

In our imputation, we included multiple measures from self-reports, tried different models, and created a single dataset using the best-performing model. Given the high percentage of missing HDRS values, adding a stochastic perturbation to the regressed imputations, generating multiple datasets, and aggregating results on all of them may reduce bias introduced by the selected model.

We extended this work to a larger dataset with 31 patients, with about 1,500 days of data [186]. We ran several ablation experiments and simplified the technique without reducing performance by dropping the imputation phase and using an ensemble of random forest and boosting for prediction. Then, we compared our method's performance using different subsets of features against two individualized baselines: 1) Individual Screen Baseline, 2) Individual Median Baseline. Individual Screen Baseline is the screen measurement score of each patient, and Individual Median Baseline is the median of each patient's scores in the training set. We evaluated our method in two deployment scenarios: time-split and user-split. In the time-split scenario, the scores from the last visits of each patient were reserved for the hold-out test, and in the user-split setting, hold-out users were reserved for the test set. Our method performed better than the individual screen baseline in both scenarios. However, it did not outperform the personal median baseline in the time-split case (Figure 2-6). See [186] for more details. This result is not surprising. Many previous published works have not reported individualized baselines, and attempts at replicating

them have failed to outperform such baselines [35, 36]. In the future, we would like to investigate approaches to overcome this limitation, for example, by minimizing identifying information in the latent space to improve generalization and learning factors that lead to variation from personal baselines. Another avenue for exploration is incorporating few-shot learning approaches for personalization with per-person parametrization.

2.7 Supplementary Materials

2.7.1 Objective vs. Subjective Reports of Sleep Quality in Major Depressive Disorder

Sleep patterns in MDD are heterogeneous: both insomnia and hypersomnia are symptoms of depression. Assessment of sleep patterns in MDD is often limited by clinicians' reliance on subjective self-reported ratings of sleep. Objective measures, such as sleep regularity measured by accelerometer data, may provide a more accurate prognostication.

We hypothesized that:

- There is variability in sleep regularity and patterns among individuals with Major Depressive Disorder (MDD).
- There is a strong correlation between subjective self-reported sleep ratings and objective accelerometer-based measurements.
- Objective sleep measurement could detect differences among individuals with MDD.

We developed an algorithm to calculate objective sleep based on accelerometer data. We calculated sleep regularity indices (SRI) for both objective and subjective sleep using the following formula:

$$SRI = \frac{1 + \frac{1}{T-\tau}\int_0^{T\tau} s(t)s(t+\tau)dt}{2} \tag{2.3}$$

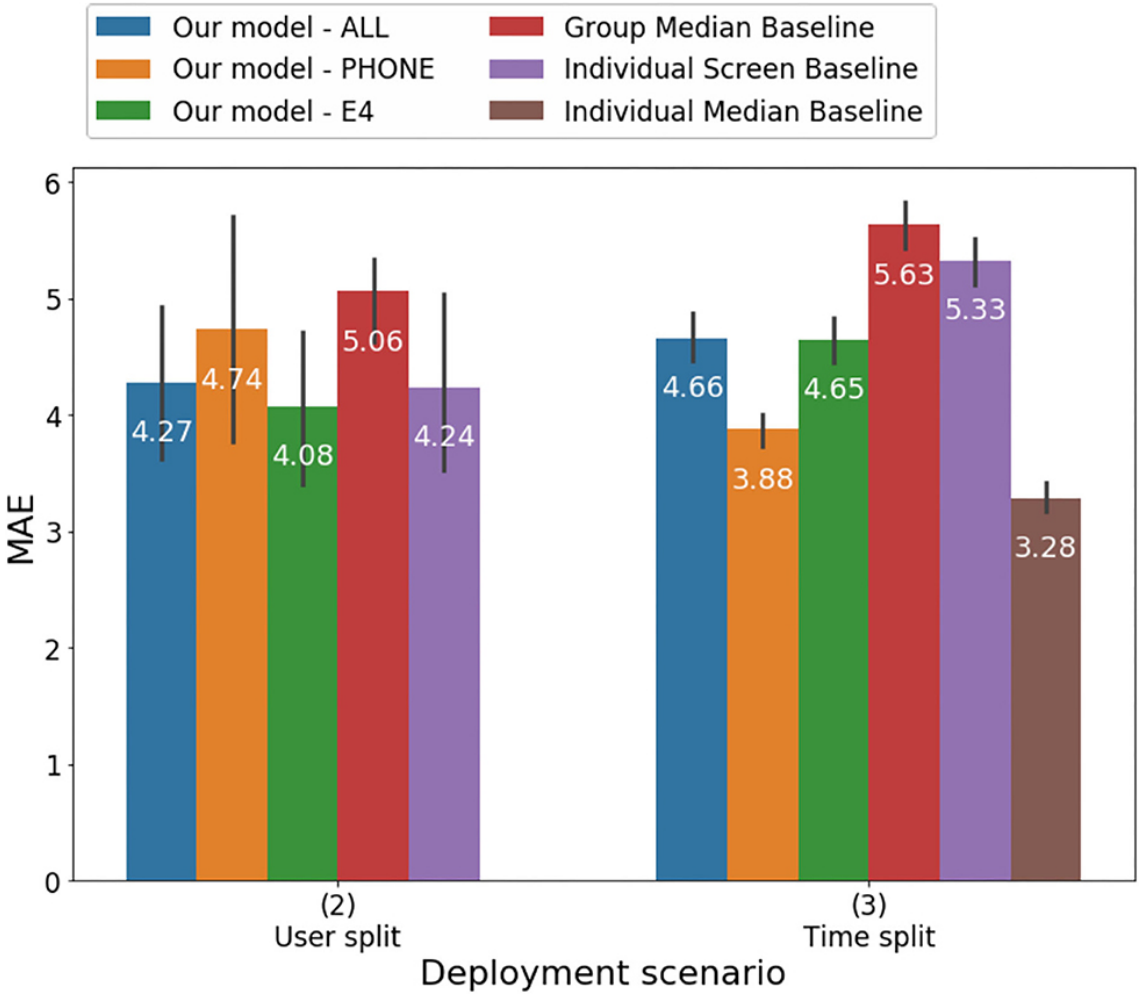


Figure 2-6: Mean absolute error of predicting HDRS using different models under the user-split and time-split scenarios [186]. In the time-split setting, the lowest mean absolute error (MAE) was obtained by the model that included only features from the phone $[F(2, 12) = 19.04, p < 0.002]$. In the user-split scenario, all modalities performed about the same $[F(2, 12) = 0.55, p < 0.59]$ with the lowest MAE obtained by the model using only the features from the wearable sensor. The best models in each deployment setting provided more accurate estimates than group median and individual screen baselines but not better than the individual median baseline in the time-split scenario. However, these differences were not significant.

Totally, the accelerometer-based (objective) and self-reported (subjective) sleep/awake time periods matched 60.94% of the time. Specifically for MDD patients, the algorithm overestimated accelerometer-based sleep epochs that were reported as awake.

Table 2.6: Objective vs. subjective sleep and awake epochs for HCs

HC Total Accuracy: 63.38%		Objective	
		Awake	Sleep
Subjective	Awake	49.28	18.86
	Sleep	17.76	14.10

Table 2.7: Objective vs. subjective sleep and awake epochs for MDD patients

MDD Total Accuracy: 59.32%		Objective	
		Awake	Sleep
Subjective	Awake	44.49	24.63
	Sleep	16.05	14.84

Based on t-statistics, MDD patients had a lower objective (t=3.09, p=0.012) and subjective SRI (t=3.37, p=0.005) compared to HCs. A trend toward positive Pearson correlation between objective and subjective SRI did not reach statistical significance in this small sample (r=0.37, p=0.17).

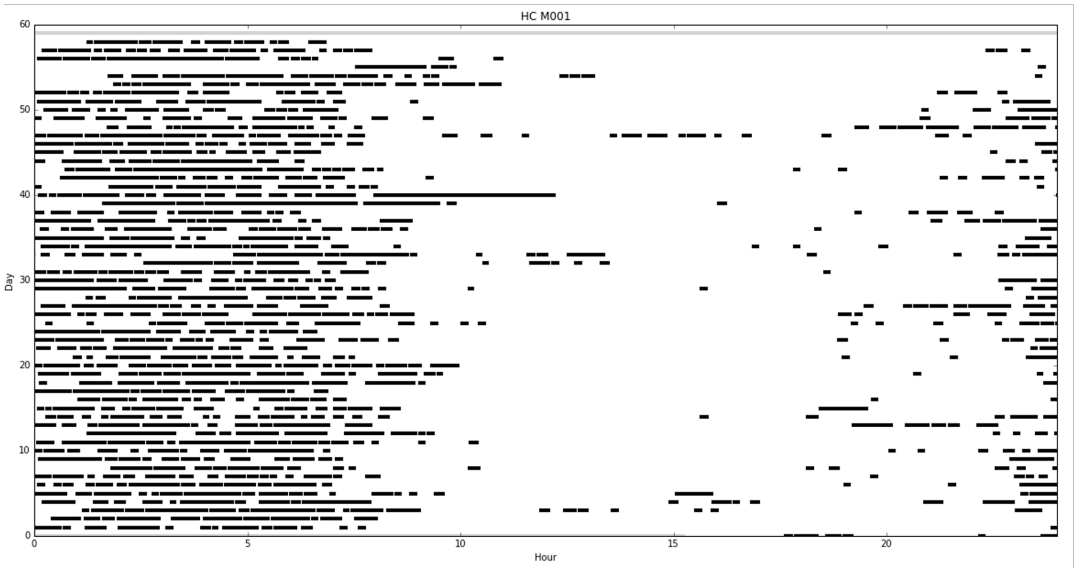
In summary, there are discrepancies between Individuals' subjective sleep ratings and objective data from the E4 sensors. Irregular sleep is associated with depression.

2.7.2 Association between Location Patterns from Commodity Phone Sensors and Depression Severity

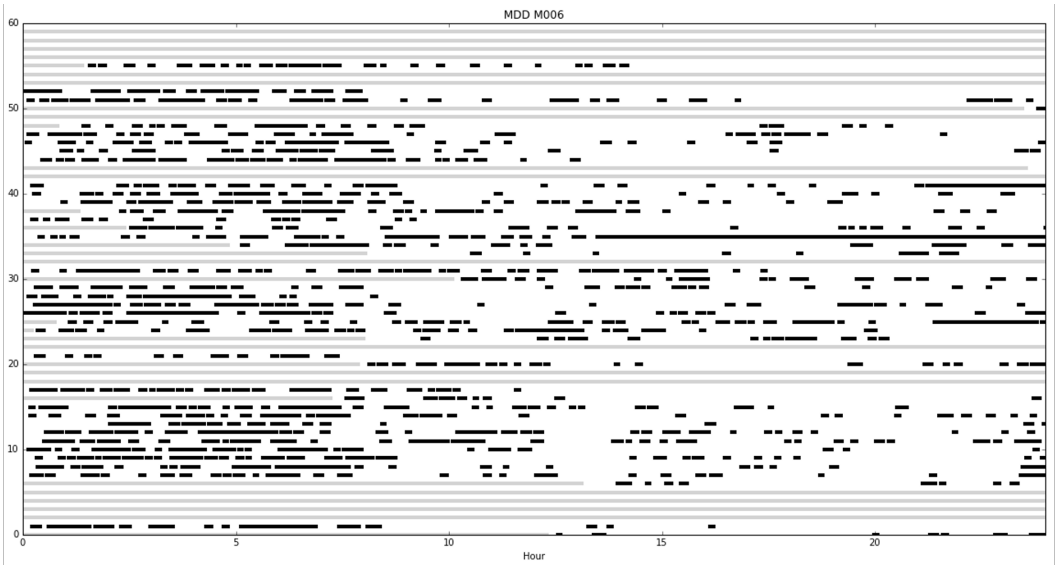
Smartphone technology is ubiquitous and can assist doctors by monitoring patients' symptoms and behavioral patterns. However, the extent to which the course of depression can be predicted with cell phone data remains unknown.

We hypothesized that location patterns from phone sensors are correlated with clinical depressive symptoms.

We calculated total standard deviation (SD) of location data, $(SD_{latitude}+SD_{longitude})/2$, in the week prior to the assessment. To remove the effect of the time spent at home around nighttime and while sleeping, we constrained the hours to between 9AM and 6PM to estimate the location changes only throughout the day. We used the full 24 hour for the



(a) A sample HC.



(b) A sample MDD patient.

Figure 2-7: Objective sleep from two sample patients. Black: sleep, white: awake, grey: missing.

weekend location SD to better represent user behavior when not obliged to show up at work.

The variable Transition Time represents the percentage of each day during which a participant was in a non-stationary state (moving faster than 0.3 m/s). We calculated the

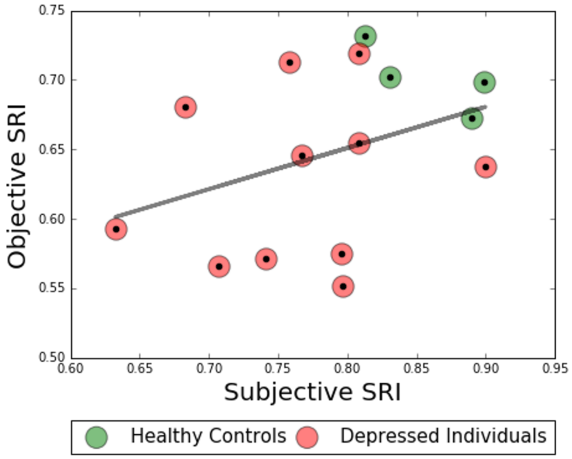


Figure 2-8: Objective vs. subjective sleep regularity index.

median of the Transition Time for each week prior to the HDRS assessment.

To asses the relationship between the HDRS total location variability in the week prior to clinical assessment while accounting for individual differences, we used linear mixed-effect (LME) models. We developed two models:

- M1: LME with random intercept
- M2: LME with random intercept and slope

We selected the model with a better balance between complexity and good fit based on Bayesian Information Criterion (BIC). For weekday: BICM1=454.6, BICM2=463.1, for weekend: BICM1=454.6, BICM2=462.5.

There was a statistically significant negative relationship between the total SD of location within day hours the week prior to the assessment (p=0.031) and HDRS total scores (M1 model).

Also, there was a statistically significant negative relationship between the total SD of location over the weekend prior to the assessment (p=0.036) and HDRS total scores (M1 model).

The variable Home Stay represents the percentage of time a participant spent at approximated home location (median location between 12am-6am), relative to other

location clusters. We calculated the median of the Home Stay for each week prior to the HDRS assessment dates.

We used the linear mixed-effects model with random intercepts and slopes to assess relationship between the HDRS and a) the Transition Time (Model 1) b) Home Stay (Model 2) using the following model:

$HDRS_i = \beta_{0i} + \beta_{1i} * LOC_i + \epsilon_i$

Where: $HDRS_i$ is the HDRS value for i-th person; LOC_i is the location metric (Transition Time for Model 1 and Home Stay for Model 2) for i-th person); β_{0i} is the i-th person intercept $\beta_{0i} = \beta_0 + \mu_{0i}, \mu_{0i} \sim N(0, \sigma_{02})$; β_{1i} is the i-th person slope, $\beta_{1i} = \beta_1 + \mu_{1i}, \mu_{1i} \sim N(0, \sigma_{12})$; ϵ_i is the i-th person error, and $\epsilon_i \sim N(0, \sigma_2)$.

There was a negative relationship trending towards significance between the median of the Transition Time metric calculated over the week prior to the assessment (p=.057) and HDRS total scores (M1 model).

There was a statistically significant positive relationship between the median of the Home Stay metric calculated over the week prior to the assessment (sample includes only MDD, p=.0393) and HDRS total scores. For all the participants p=.098. (M2 model).

In summary, location variability during day hours is negatively associated with HDRS score in the week prior to the assessment. The same trend is observed for location variations over the weekend prior to the assessment. Depression severity (measured with HDRS) is positively associated (p=.0393) with the % of time spent in home, and time in transition decreases with depression severity (p=.057)

2.7.3 Association Between Cell Phone Social Interactions and Depression Severity

Cell phone technology can assist doctors by monitoring patients' symptoms, and may eventually be useful in the prediction of depressive episode time courses. However, the extent to which the course of depression can be predicted with cell phone data remains

unknown. The quantitative measurement of communication patterns (i.e., number of text messages and phone calls) between depressed patients and their contacts may be useful for the prognostication of the course of depression. We hypothesized that more communication with social contacts via texts and phone calls would correlate with lower depression scores.

There was a statistically significant negative relationship between the average number of outgoing calls (p=.014) and HDRS total scores (M1 model).

To assess the relationship between the HDRS total score and the number of calls/texts in the week prior to clinical assessment while accounting for individual differences, we used linear mixed-effect (LME) models. We developed two models:

- M1: LME with random intercept

- M2: LME with random intercept and slope

We selected the model with a better balance between complexity and good fit based on Bayesian Information Criterion (BIC).

Furthermore, there was a statistically significant negative relationship between the average duration of outgoing calls (p=.047) and HDRS total scores (M2 model).

No significant relationship was observed between other hypothesized parameters (the number of incoming calls, texts, or the duration of the incoming calls in the week prior to the assessment) and HDRS total scores.

In summary, our results showed a significant negative relationship between number and duration of outgoing calls and subjective reporting of depression severity. Participants who were feeling less depressed may have been more inclined to reach out socially. Or, initiating more social interaction may have caused participants to feel less depressed. Longer interactions may be more meaningful and supporting and thus alleviating depressive symptoms.

2.7.4 Association Between Mood and Alcohol Use in Major Depressive Disorder

Heavy drinking often co-occurs with MDD, increasing disability and preventing the amelioration of symptoms. New tools such as ecological momentary assessment (EMA) and wearable sensors allow a more granular examination of the association between MDD and alcohol use. In this section, our objective is to examine the association between depressive symptoms and alcohol consumption and moderators of the association through active and passive data recording.

Surveys included 10 questions from the Positive and Negative Affect Scale (PANAS) and questions about number and type of drinks consumed daily.

Our dependent variables include:

- Low Mood: log total NA/PA;
- EDA asymmetry: difference in skin conductance level between right and left wrists over the course of the day;
- Home stay: percentage of time spent at home over the course of 24 hours;
- Alcohol use: daily number of drinks (SD).

We hypothesized that:

- Low mood and alcohol consumption are positively associated.
- Time spent at home moderates the association between mood and alcohol use.
- EDA asymmetry moderates the association between mood and alcohol use.

We used Linear Mixed Effects models with random intercept and slope to model the association between mood and alcohol consumption. Results showed a positive relationship between low mood and alcohol use ($p=0.003$).

A Linear Mixed Effects model with random intercept showed that the interaction between alcohol consumption and time at home was not significant (p=0.46). However, the percentage of time spent at home was directly associated with low mood (p=0.05).

A Linear Mixed Effects model with random intercept and slope showed a significant interaction between alcohol consumption and EDA asymmetry (p<0.0001): the greater EDA asymmetry, the stronger the influence of alcohol consumption on mood.

Findings are consistent with previous studies showing an association between mood and alcohol use. As expected, time spent at home and mood were associated. However, alcohol use did not affect this relationship. Higher arousal was associated with stronger association between alcohol consumption and mood. Integrating different technologies to assess alcohol use and mood is feasible. Daily passive and active recording will facilitate the development of complex models explaining the association of mood and alcohol use and moderating and mediating factors.

2.8 Conclusion

In this chapter, we showed the feasibility of continuously measuring depressive symptoms using a new method that requires only passive data. This data is captured from built-in sensors of a regular android phone and E4 wristbands, including measures of EDA, sleep patterns, motion, communication, location changes, and phone usage patterns. Using a novel combination of machine learning techniques and a day of data from wearable and phone sensor data, we could predict the Hamilton Depression Rating Scale (HDRS) values on a hold-out set, obtaining a low error rate of 4.5 RMSE. Moreover, a post hoc statistical analysis showed that poor mental health was associated with more irregular sleep and fewer incoming messages. Less motion measured by average motion, the standard deviation of motion, and the fraction of time in motion were associated with poor mental health. Additionally, less variability in location patterns measured by the standard deviation of location pattern changes between 9 am - 6 pm was associated with poor mental health.

Worse mental health was also associated with a higher asymmetry of EDA between the right and left wrists measured by the difference in skin conductance response peaks and mean EDA difference (right minus left).

In the future, we would like to explore the feasibility of our methods for other scenarios, for example having hold-out test subjects resembling when some patients have no observed data in the training phase. We know that there is some interdependency among patients. The variety of our prediction models and the ensemble methods can learn to account for individual differences. We would like to explicitly model that by comparing against mixed-effect models and modeling the patients' variations from their baseline. In this work, we included several post hoc analyses to discover more informative data streams. We would like to explore the discriminative power of those features further.

2.9 Statement of Contributions

This chapter, especially capturing the rich dataset, comes from a collaboration between Roz Picard's Affective Computing group and Depression and Clinical Research Program (DCRP) at Massachusetts General Hospital. My contributions include checking data quality and recruitment, resolving technology problems, cleaning data, feature extraction, and all the modeling and analyses provided in this chapter.

Szymon Fedor has made significant contributions to get this work off the ground. He wrote the original grant proposal that funded this work and contributed to several following grant proposals. He managed many aspects of the project, including coordinating with the medical team, preparing hardware, advising the project, and processing the physiological features. He has done several other analyses on this dataset that though not included in this thesis, have informed my thinking.

I started developing the predictive models when I took David Sontag's class on Machine Learning for Health Care. David has provided insightful advice throughout the course and afterward. Roz has provided advice and guidance throughout the project.

Our medical collaborators have played a significant role by helping us identify the right questions to ask, recruiting patients, gathering clinical ratings, and helping with grant proposal writing: Dr. David Mischoulon, Dr. Paola Pedrelli, Dr. Dawn Ionescu, Dr. Jonathan Alpert, Dr. Joshua Curtiss, Michael Pittman, Ashley K. Meyer, Bridget Wallace, Esther Howe, Lisa Sangermano, Chelsea Dale, John Lin.

MEng and undergraduate students have helped us along the way by building a visualization platform that made investigating this data more accessible, helping with data gathering, and several analyses that are not included in this thesis but have been impactful in addressing several questions in this space: Darian Bhathena, Noah Faro, Olivia Valle, Sarbari Sarkar, Marek Subernat.

Chapter 3

Approximating Interactive Human Evaluation in Open-Domain Dialog

Building an open-domain fully automated conversational agent is a challenging problem. Current evaluation methods, mostly post hoc judgments of static conversation, do not capture human notions of conversation quality. In this chapter, I describe the work I conducted with a team of collaborators (see section 3.9 for individual contributions). We investigate interactive human evaluation and provide evidence for its necessity; we then introduce a novel, model-agnostic, and dataset-agnostic method to approximate it. In particular, we propose a set of psychologically motivated proxies that capture sentiment, semantic coherence, and user engagement on the conversation trajectory. Then, we employ a self-play scenario where the dialog system talks to itself and we calculate the combination of the aforementioned proxies. While previous automated metrics at best only poorly correlate with human judgments of quality (r=.44) [149], we show that this newly developed hybrid metric is capable of capturing the human-rated quality of a dialog model better than any automated metric known to-date, achieving a significant Pearson correlation ($r > .7, p < .05$). To investigate the strengths of this novel metric and interactive evaluation in comparison to state-of-the-art metrics and human evaluation of static conversations, we perform extended experiments with a set of models, including

several that make novel improvements to recent hierarchical dialog generation architectures through sentiment and semantic knowledge distillation on the utterance level. Finally, we open-source the interactive evaluation platform we built and the dataset we collected to allow researchers to efficiently deploy and evaluate dialog models.

3.1 Introduction

The goal of an open-domain conversational agent is to carry out natural social interactions with humans. Current state-of-the-art generative neural networks fail in producing key aspects of good natural conversation, including staying on topic, not being repetitive, and generating emotionally appropriate responses. One of the biggest challenges in training better dialog systems relates to the difficulty of evaluating them. Automatic metrics such as BLEU score relate poorly to human judgment of dialog quality [145], and while embedding-distance based metrics provide an alternative [166], we will show that they also do not correlate well with human evaluation. Without a reliable metric to optimize, training high quality dialog models remains difficult.

Since humans are the ultimate authority on what constitutes a good conversation, many authors rely on human ratings to evaluate their methods [183, 222, 223]. The predominant procedure for obtaining human ratings uses *static evaluation*: a context of several sentences, often originating from the dataset that dialog model was trained on, is used as input to generate a response (either a single utterance or multiple consecutive utterances). This particular generated response for this particular context is then provided for a human to evaluate. However, such observational evaluation of a static text sample is limited in capturing common failure modes of open-domain dialog systems, such as a lack of diversity in the responses, inability to track long-term aspects of the conversation, and inability to maintain a consistent persona [42]. Despite this, static evaluation is commonly used for evaluating these exact qualities [137, 160].

Conversation is inherently a process. In this work, we argue that multi-turn interactive

Table 3.1: Static evaluation fails to capture a lack of diversity in a dialog model's responses, as well as its inability to track the conversation and respond in emotionally appropriate ways. We argue interactive evaluation is needed to evaluate dialog models, and show that our novel Emotion+Infersent (EI) models trained on a larger and more diverse corpus, produce better interactive dialog. We present strong evidence that our novel dialog self-play framework combined with psychologically motivated novel automated metrics can accurately estimate quality of a model with respect to its ability to carry out multi-turn open-domain conversations. Here, examples from one model category are included: Hierarchical Recurrent Encoder Decoder (HRED) [222]. Similar observations for other model categories are included in the appendix. * refers to novel elements of our work, including a new evaluation framework, new model, and dataset.

Evaluation Strategy	**Static**	**Interactive**	**Interactive**	**Self-play***
Model	**HRED**	**HRED**	**EI* HRED**	**EI* HRED**
Dataset	**Cornell**	**Cornell**	**Reddit***	**Reddit***
	[Context]: you're weird. [HRED]: i'm sorry.	[User]: i'm feeling sad today. [HRED]: no. [User]: yes i am! [HRED]: i'm sorry, i'm sorry. i don't want to know. [User]: you're weird. [HRED]: i'm sorry.	[User] talk to me [EI HRED]: it's my job. [User]: i love you [EI HRED]: thank you for the kind words. [User]: do you love me? [EI HRED]: it's a bit personal, so I sent you a dm. [User]: i'm blushing	[EI HRED] i can't wait to see what i can do. [EI HRED] that's a good idea. [EI HRED] how do you deal with people? [EI HRED] i've been on reddit for a long time. [EI HRED] i am in the same boat.

human evaluation is essential for evaluating this process, and for making progress in improving open-domain dialog models. A multi-turn interactive evaluation is open-ended, and allows the users to derive and control the trajectory of the conversation by saying whatever they want. As opposed to post hoc observations of a generated response to a static context, successful interactive dialog generation is an ultimate test of generalization. Table 3.1 illustrates an example where a bot generates a coherent static response, but interactive evaluation shows that the distribution of its responses has collapsed onto repeatedly saying the same phrase, a common problem in open domain dialog [139].

The relative sparsity of interactive human evaluation of dialog systems is partly due to the difficulty and expense of collecting human data. Therefore, we develop a way to approximate human judgment of interactive dialog quality using a novel form of dialog self-play. We begin by moving beyond superficial word-level mapping by proposing a series of human-centered metrics to evaluate the quality of conversation motivated by findings in psychology. Specifically, inspired by the effectiveness of sense of humor in creating solidarity [95], style matching for forming relationship stability and social

cohesiveness [79, 110], and the importance of active listening through forming follow up questions [108], we propose metrics to capture sentiment, semantics, and user engagement for conveying empathy and understanding. We then fit a function that predicts human assessments of conversation quality given these metrics. This function is used to predict bot quality through self-play: for a fixed number of turns, the bot generates utterances which are fed back into itself as input in the next turn. The same metrics described above are computed on the self-play generated conversation, and the same function fit to human data is used to predict the bot quality. We show a very high Pearson correlation ($r > .7, p < .05$) between the predicted quality scores and the ground-truth human judgments of bot quality, suggesting self-play is a good proxy for interactive conversation assessment.

To demonstrate the relevance of the interactive evaluation and the proposed self-play evaluation, we perform extended experiments with different hierarchical architectures. In particular, we compare three recent hierarchical baselines: HRED [222], VHRED [223], VHCR [183]. Motivated by sentiment and semantics being key aspects of producing high quality conversations, we regularize the top level of the hierarchy to ensure it encodes such information, using model distillation [103]. Our results show the effectiveness of the proposed regularization in interactive evaluation in both the human-bot and the self-play scenarios.

This work makes three main contributions: 1) demonstrates the benefits of multi-turn interactive evaluation to capture the quality of the dialog systems; 2) presents a novel self-play framework to estimate a new psychology-motivated hybrid quality score. These estimations are highly correlated with quality scores obtained from interactive human evaluation, more strongly than the state-of-the-art automated metrics; 3) proposes a new method of regularizing hierarchical seq2seq models with knowledge distillation. All the code, data, and interactive evaluation platform resulting from our work are publicly available.

3.2 Related Work

Interactive evaluation in dialog has been mostly limited to presenting the results of competitions (e.g. the Alexa prize [221, 246], or the Conversational Intelligence Challenge [42]). Those findings reveal that most bots do not perform well in interactive evaluation, due to repetitiveness, inability to balance dialog acts across the conversation, and inability to maintain a consistent persona [42]. Even work aimed at maintaining a persona does not test in an interactive setting [137, 160]. To the best of our knowledge, no prior work has compared multi-turn, interactive human evaluations of open-domain dialog models to traditional forms of evaluation.

Dialog systems remain difficult to train due to the lack of metrics that can effectively capture good dialog quality. Several authors have proposed training automatic predictors of human judgment or to combine human judgment with automatic metrics [93, 94, 149]. Before our work, the best model trained to predict human judgments achieved a Pearson correlation of .44 with the ground truth [149].

The lack of research into interactive evaluation relates to the difficulty and cost of collecting human ratings. We show that human judgments of the quality of an interactive evaluation can be automatically and reliably approximated using dialog model self-play. There is limited work investigating self-play for dialog systems: [225] use a task schema and user simulator to generate samples for input to a goal-directed dialog system, while [139] use a copy of a dialog model to compute a reward function that can be optimized with reinforcement learning. However, we are not aware of prior work using self-play for approximating interactive human evaluation.

Interactive conversation necessitates tracking long-term aspects of the dialog like the topic and tone. Hierarchical recurrent neural networks (RNNs) have been proposed as a way to improve long-term tracking of the conversation, through maintaining both a word- and utterance-level RNN [183, 222, 223, 226, 265]. Yet dialog is more than language modeling, it requires topic and social coherence. Prior performance improvements to

dialog models using topic information include appending topic as an additional input [77], or extracting topic information using Latent Dirichlet Allocation [138, 258]. Towards social and emotional coherence, previous works have investigated various features and loss functions based on emotion [107, 197, 198, 266, 267]. Given research highlighting the ineffectiveness of LDA for short texts [259], such as those involved in casual conversation, and the unavailability of topic and tone supervision at-scale, approaches overcoming these limitations are preferred. To the best of our knowledge, transferring sentiment and semantic information from a pre-trained model directly into a dialog model using knowledge distillation [103] has not been studied. Thus, we select a set of recent hierarchical dialog models and their improved versions through knowledge distillation for a thorough multi-turn interactive evaluation and comparison to traditional evaluation.

3.3 Knowledge Distillation for Sentiment and Semantic Regularization

To systematically compare multi-turn interactive evaluation of open-domain dialog with traditional forms of evaluation, we include a diverse set of models. Particularly, we build on three existing hierarchical seq2seq architectures designed for dialog. Here, we provide a brief summary; for detailed information, see [183, 222, 223]. The first baseline model, Hierarchical Recurrent Encoder Decoder (HRED) [222] extends a traditional seq2seq model by adding a third recurrent neural network (RNN), which is only updated after each dialog turn, or utterance. The idea behind this *Context RNN* is that it could potentially track longer term aspects of the conversation, such as the topic; however, there is no guarantee that it will learn to do so. The decoder of the HRED model conditions on both the embedding produced by the encoder for the current utterance, h_n^e, and the embedding of the Context RNN for the previous utterance, h_{n-1}^c.

The second baseline model, Variational HRED (VHRED) [223], extends HRED with a

variational constraint on the utterance embedding space z. Let $x_n = [w_{1n}, w_{2n} \dots w_{mn}]$ be the n-th utterance composed of tokens $w_{1..m}$. VHRED predicts x_n as follows:

$$h_n^e = f^e(x_{n-1}) \tag{3.1}$$

$$h_{n-1}^c = f^c(x_{n-1}, h_{n-1}^e) \tag{3.2}$$

$$\mu, \Sigma = f(h_{n-1}^c) \tag{3.3}$$

$$p_\theta(z_n|x_{<n}) = N(z|\mu, \Sigma) \tag{3.4}$$

$$p(x_n|x_{<n}) = f^d(h_{n-1}^c, z_n) \tag{3.5}$$

Equations (5.1)-(5.5) describe the computation of VHRED at inference time where f^e, f^c, and f^d are Gated Recurrent Unit (GRU) networks for the encoder, context, and decoder RNNs, respectively; at training time, it allows the computation of z, μ, and Σ to condition on the encoding of the target utterance, h_n^e, giving the posterior distribution $p_\Psi(z_n|x_{\leq n})$. A Kullback-Leibler (KL) divergence constraint is placed between the posterior and prior, $D_{KL}(p_\Psi||p_\theta)$.

The third model, Variational Hierarchical Conversation RNN (VHCR)[183] further extends VHRED by drawing a prior encoding $z^{conv} \sim N(0, I)$ for each conversation, allowing all parts of the model (f^c, μ, Σ) to condition on z^{conv}, which is unchanging throughout the conversation.

3.3.1 Emotion and Infersent Regularization (EI)

While the hierarchical design of these models is motivated by a desire to allow tracking high-level, slow-changing aspects of the conversation like topic or tone, it is unclear that the network will be able to model these aspects without additional structure or information. We thus propose a regularization to the top level of the hierarchy, the Context RNN, to force it to encode both the sentiment and semantics of the utterance. To do this, we leverage a state-of-the-art sentiment detection model trained on a large Twitter corpus [52], as well as

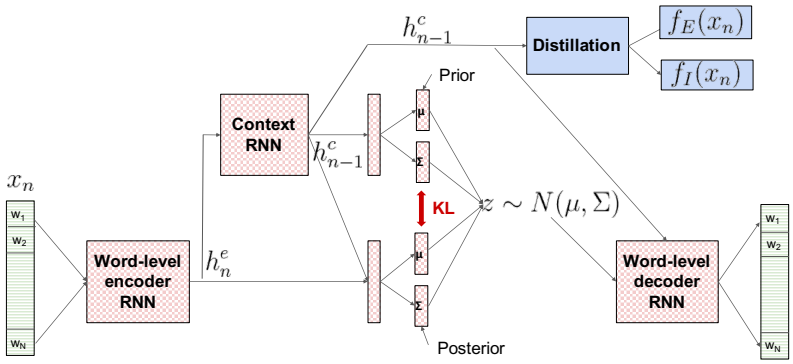


Figure 3-1: Illustration of the EI regularization (blue-solid) applied to VHRED baseline (red-checkered) to enforce encoding sentiment and semantics of an utterance in the Context RNN.

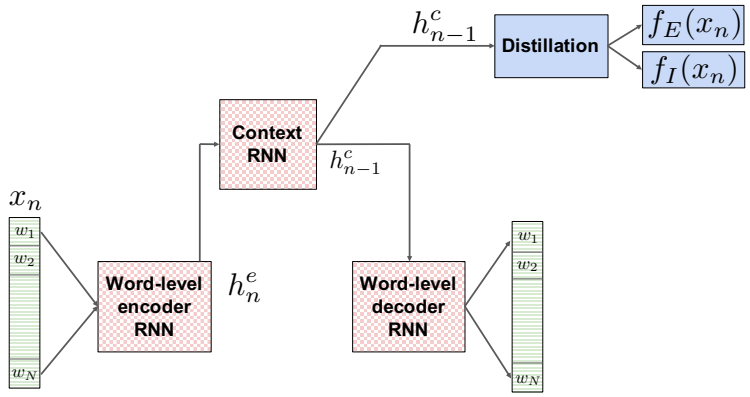


Figure 3-2: Illustration of EI regularization (blue-solid) applied to HRED baseline (red-checkered) to enforce encoding sentiment and semantics of an utterance in the Context RNN. The EI regularization can be similarly applied to VHCR.

the recently proposed *Infersent* sentence-embedding model trained to predict the meaning (i.e. entailment, contradiction) of sentences [29], and distill them into the *Context RNN*.

First, we use these models to predict the emotional content, $f_E(x_n)$, and infersent embedding, $f_I(x_n)$ of each input utterance. We then add an additional network to the hierarchical models which predicts these values based on the context RNN embedding of the utterance: $f^{distill}(h^c_n) = < f_E(x_n), f_I(x_n) >$. The goal is to transfer knowledge of emotion and semantics in text into the context RNN via knowledge distillation [103].

Figures 3-1 and 3-2 illustrate, in blue color, the EI regularization applied to the

VHRED model and HRED models respectively. The regularization can be similarly applied to VHCR. In our experiments we refer to the regularized models as HRED-EI, VHRED-EI, and VHCR-EI, respectively, or, more generally, EI models as opposed to baseline models. The code for all our models is available at `https://github.com/natashamjaques/neural_chat` and was originally based on [183]. For details regarding hyper-parameter tuning refer to §3.7.12.

3.4 Interactive Evaluation Methodologies

3.4.1 Traditional Evaluation

Automatic metrics Embedding-based metrics compare generated sentences to ground truth sentences using a vector representation of words [166]. In this work, we use three embedding metrics: embedding *average*, vector *extrema*, and *greedy* matching. These three metrics are used in previous open-domain dialog models [145, 183, 223]. We also use *perplexity* as a standard measure of the likelihood of the generated sentences with respect to the target outputs. Another common metric for variational models is the KL-Divergence between the posterior and the prior distribution, as a way of assessing the information encoded into the latent variables [226] (Figure 3-1 illustrates KL for the VHRED model). As I show in this chapter, most of these metrics ignore the trajectory of the conversation and are blind to human-centric qualities that make a conversation high-quality, such as sentiment, semantics, and user engagement. More information regarding embedding metrics can be found in §3.7.7.

Conventional static human evaluation We employ a similar method to previous work for our static human evaluation of generated responses [183, 223], sampling contexts from each corpus and asking humans to compare the generated responses. To reduce ambiguity, we exclude contexts shorter than 10 tokens and contexts containing <unknown> tokens. We recruited participants from Amazon Mechanical Turk (AMT) to compare generated

sentences. Annotators could also select a third "tied" option. For each example (context and pair of generated sentences), we asked annotators to compare generated sentences based on quality, fluency, diversity, contingency, and empathy. Each batch of 100 pairwise comparisons were labeled by 6 - 8 annotators.

3.4.2 Interactive Human Evaluation

To address the limitations of static human evaluation, we built a platform for conducting interactive evaluation of dialog models with humans, which we make available in open-source to the community (see Figure 3-5). Annotators rated quality, fluency, diversity, relatedness, and empathy of a bot after interacting with it for at least 3 turns. Participants can also upvote or downvote each bot response. For more information about this platform, see §3.7.10. Our goal is to make this work transparent and reproducible, while adding diversity to the platforms future practitioners can choose to use (e.g. ParlAI [165], Plato Research Dialog System [181], ChatEval [217]).

Figure 3-3: Consent form in the Interactive Evaluation Platform (available at `https://neural.chat`).

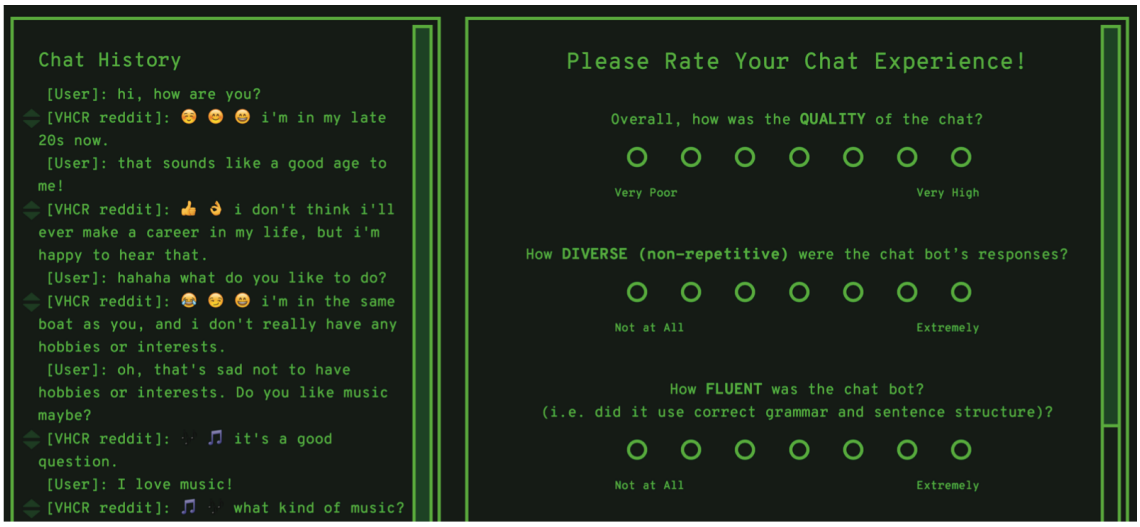


Figure 3-4: Interactive Evaluation Platform (available at `https://neural.chat`): Side-by-side view of chat history (left) and the first part of the evaluation form (right).

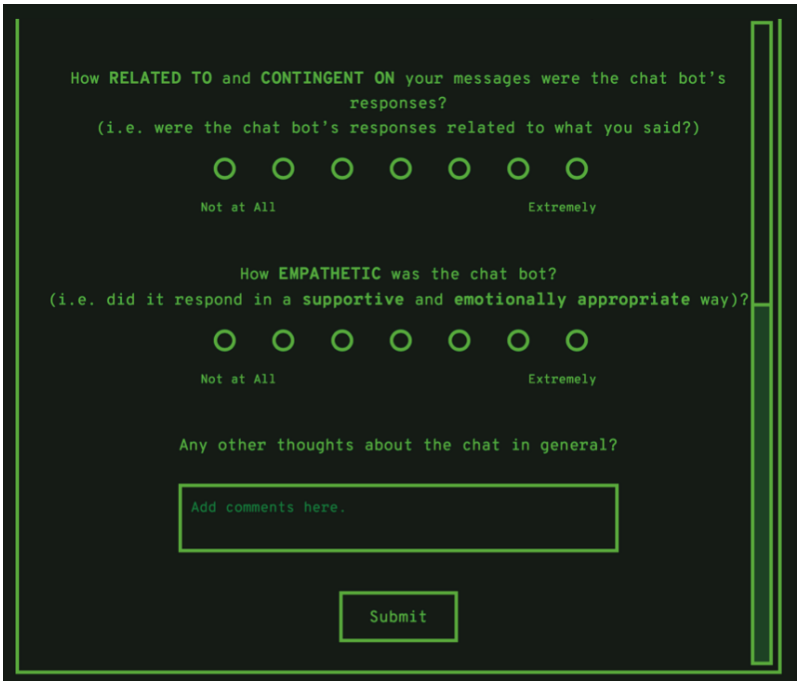


Figure 3-5: Interactive Evaluation Platform (available at `https://neural.chat`): The second part of the evaluation form showing the remaining questions.

3.4.3 Novel Metrics and Self-play

Inspired by real-world human interactions, we introduce novel metrics to capture the morphology of a conversation, i.e., how the users' responses progress over time and how the bot's responses interact with them. We propose a hybrid combination of these metrics, M_H, that is optimized to predict conversation quality on human data. We then apply M_H to self-play, i.e., the trajectory of bot-generated responses, and investigate how it relates to human ratings of conversation quality.

Sentiment metrics To approximate emotional tone of an utterance, we use a state-of-the-art sentiment detector trained on a large Twitter corpus [52]. This pre-trained model outputs an emotion embedding – a probability distribution over 64 most-frequently used emojis. To estimate the *Sentiment Coherence* between user's query and generated samples, we calculate the cosine similarity between their emotion embeddings. We define a set of weights over the 64 emojis and calculate the weighted sum over an emotion embedding vector to derive a *Sentiment* score which is higher for positive sentiment and lower for negative sentiment (See §3.7.11). We define *Sentiment Transition* as the change between user's *Sentiment* before and after a bot response. Additionally, *Sentiment Min-Max* is defined by the slope of change between min and max *Sentiment* in user utterances over the course of a conversation. Since humor can be used to create solidarity [95], we count the number of "ha"s in the user response as a proxy for *Laughter*. The combination of these metrics provides a snapshot of the trajectory of sentiment in a conversation and quantifies if the bot is able to elicit positive emotions in the user.

Semantic metrics Language style matching is a strong predictor of relationship stability [110] and social cohesiveness [79]; thus, we introduce metrics to capture lexical similarity. We use *Infersent*, a state-of-the-art sentence-embedding model to encode the user and bot responses into a 4096-dimensional embedding space [29]. Infersent was trained to distinguish if two sentences are supporting, contradicting, or have a neutral relationship. We estimate *Semantic Similarity* by calculating the cosine similarity between

the infersent embedding of the user's query and the generated bot sample. Additionally, we use the classic Word2Vec embeddings trained on Google News Corpus along with average, extrema, and greedy aggregation methods similar to Section 3.4.1 to derive *Average Word Coherence*, *Extrema Word Coherence*, and *Greedy Word Coherence* between user and bot responses.

Engagement metrics Asking questions is an important active listening skill which is linked to conversation management, attentiveness, and responsiveness [11, 108]. Therefore, we define *Question Score* to quantify if the bot is using question words and/or a question mark. We also introduce *# Words* as a proxy for user engagement that counts the number of words in their response.

Hybrid metric (M_H) We combine the aforementioned metrics (M_i) using linear regression, and optimize their coefficients (λ_i) to best predict human judgment of interactive conversation quality: $M_H = \sum \lambda_i * M_i + \lambda_0$. We use a leave-one-bot-out scenario where we isolate all the human conversations with one of the dialog models, χ_j, as the hold-out test set. We train the $\lambda_{i,j}$ on the remaining quality ratings. We found that the learned λ_is were stable across the training folds, only exhibiting small variations. Other researchers are encouraged to use our learned coefficients directly or adjust them according to their own interactive human evaluation dataset. See §3.7.2 for more details about the learned λ_is.

Self-play as an approximation for interactive evaluation Since interactive human evaluation is costly, we propose a *self-play* scenario where the dialog system talks to itself, i.e. the bot generated responses are fed back into it as the next turn input. For each model χ_j, we generate 100 random conversations, fixed at 10 turns. The self-play trajectories created using model χ_j are treated as the hold-out set. Therefore, the trained $\lambda_{i,j}$ values based on all conversations except for the ones with χ_j are used to calculate M_H on each generated bot-bot conversation trajectory for χ_j. The estimated M_H values are averaged across conversation samples for χ_j. This value is used for comparison against the ground-truth interactive quality ratings aggregated on the bot-level.

3.5 Experiments

3.5.1 Datasets

A common source of data for open-domain dialog systems is movie scripts, among which the CORNELL dataset [32] is the largest and most commonly used. Therefore, we use it to benchmark against previous state-of-the-art results [183]. Its median conversation length is 3 utterances and the conversations are strictly between pairs of speakers. Recognizing that movie lines have limited conversation diversity, we also built a new corpus, REDDIT. Between the many different subreddits available, the conversations vastly differ on topic, language style, and participation patterns. We select the Casual Conversations forum (`r/CasualConversations`), a community of $607K$ conversationalists discussing a variety of topics. We collect a dataset of $109K$ conversations of at least 3 turns with the median conversation containing 7 utterances from conversational exchanges on the platform in 2018[1]. More more details about this dataset refer to §3.7.6.

3.5.2 Interactive Human Evaluation

Table 3.1 (in §3.1) illustrates how EI regularization produces a higher quality conversation when compared to baseline. Rather than cherry-picking results, we make all of the bots evaluated in the study available at `https://neural.chat/BRFZACDCOA/` for readers to assess interactively.

Table 3.2 summarizes human ratings of baseline and EI models obtained via interactive evaluation. In total, 565 ratings were captured. We used a 7-point Likert scale to capture

[1]This REDDIT dataset is available at `https://affect.media.mit.edu/neural_chat/datasets`.

Table 3.2: Mean human ratings for Baseline and EI (Emotion+Infersent) models for HRED, VHRED, and VHCR architectures with 90% confidence intervals. See §3.5.2 for 3-factor ANOVA results.

		Cornell		Reddit	
Model	Metric	Baseline	EI	Baseline	EI
HRED	quality	2.182 ± 0.305	**2.347** ± 0.313	2.527 ± 0.310	**2.714** ± 0.299
	fluency	3.909 ± 0.387	**4.000** ± 0.381	4.436 ± 0.349	**4.786** ± 0.316
	diversity	**2.836** ± 0.374	2.735 ± 0.380	3.418 ± 0.386	**3.554** ± 0.372
	contingency	2.200 ± 0.291	**2.469** ± 0.336	2.382 ± 0.288	**2.536** ± 0.322
	empathy	**2.673** ± 0.352	2.490 ± 0.350	3.018 ± 0.329	**3.107** ± 0.337
VHRED	quality	2.022 ± 0.309	**2.333** ± 0.252	2.694 ± 0.392	**2.864** ± 0.341
	fluency	3.109 ± 0.351	**3.949** ± 0.396	4.250 ± 0.496	**4.477** ± 0.402
	diversity	3.565 ± 0.442	**4.385** ± 0.371	**5.00** ± 0.468	4.705 ± 0.353
	contingency	2.261 ± 0.287	**2.487** ± 0.346	2.472 ± 0.362	**2.773** ± 0.370
	empathy	**2.739** ± 0.374	2.564 ± 0.367	3.000 ± 0.393	**3.341** ± 0.385
VHCR	quality	2.132 ± 0.247	**2.548** ± 0.380	2.615 ± 0.350	**2.692** ± 0.298
	fluency	2.679 ± 0.306	**3.976** ± 0.380	3.923 ± 0.433	**4.308** ± 0.395
	diversity	3.755 ± 0.340	**4.238** ± 0.421	**4.436** ± 0.455	4.231 ± 0.382
	contingency	2.189 ± 0.270	**2.571** ± 0.356	2.077 ± 0.298	**2.692** ± 0.354
	empathy	2.340 ± 0.316	**2.714** ± 0.368	2.974 ± 0.434	**3.288** ± 0.379

quality[2], diversity [3], fluency [4], consistency [5], and empathy [6] of the chatbot, where one represented the worst and seven represented the best rating. See Figures **??** and 3-5 for the interface of user evaluation. Each dialog model has been evaluated by a number of annotators, ranging from 36 to 56. For additional information about human annotators refer to §3.7.9. We ran a 3-factor ANOVA on the sum of user scores, where the independent variables are model architecture (HRED, VHRED, VHCR), EI regularization (Baseline, EI), and dataset (CORNELL, REDDIT). We found a significant main effect of EI regularization and dataset, but no significant difference between the three types of hierarchical models. We found that adding emotion and infersent (EI) regularization to baseline models improved the interactive chat experience significantly, $F(554, 1) = 9.016, p = .003$. Further, the models trained on the REDDIT dataset performed significantly better, $F(554, 1) = 30.796, p < .001$. This finding validates the hypothesis that distilling information about topic and

[2]Overall, how was the quality of the chat?

[3]How diverse (non-repetitive) were the chatbot's responses?

[4]How fluent was the chatbot? (i.e. did it use correct grammar and sentence structure)?

[5]How related to and consistent on prior messages were the chatbot's responses? (i.e. were the chatbot's responses related to what you said?)

[6]How empathetic was the chatbot? (i.e. did it respond in a supportive and emotionally appropriate way)?

Table 3.3: Results of automatic traditional metrics for 1-turn responses of models per context of baseline and EI (Emotion + Infersent) models. PPL: perplexity, KL: KL divergence, Avg: Average, Ext: Extrema, Grd: Greedy

		Cornell					Reddit				
Model	Version	PPL	KL	Avg	Ext	Grd	PPL	KL	Avg	Ext	Grd
HRED	baseline	52.311	-	.471	.329	.331	41.730	-	.649	.394	.474
	EI	**47.636**	-	**.560**	**.383**	**.400**	**41.245**	-	**.651**	**.398**	**.482**
VHRED	baseline	**49.414**	.264	.539	.352	**.395**	36.240	**.188**	.635	.383	.464
	EI	50.526	**.517**	**.545**	**.355**	.394	**35.510**	.167	**.636**	**.392**	**.465**
VHCR	baseline	61.000	**.562**	.532	.345	.382	**36.736**	**.267**	.619	.371	.448
	EI	**49.243**	.475	**.588**	**.369**	**.444**	37.198	.231	**.639**	**.394**	**.469**

tone into the top level of the hierarchy is useful for good conversation, and suggests that the REDDIT dataset could provide more realistic training for open-domain dialog and be valuable to the community. Additional ablation results are provided in §3.7.1.

3.5.3 Traditional Metrics

Automatic metrics Several prior works have focused on ensuring that the variational KL term remains high in order to try to improve model quality (e.g. [183, 226]). However, we observe there is no consistency between human quality rating and KL (Table 3.3). See §3.7.8 for details about other human metrics, e.g. fluency, diversity, contingency, and empathy. Thus, it is not evident that KL as classically formulated for dialog captures human judgements of dialog quality. Even perplexity (a transformation of the cross-entropy loss used to train our models) falls short of capturing human quality judgments, underscoring the difficulty in effectively training good language models. We find embedding metrics show more promise in preserving the order of human quality ratings, but still have only weak correlation with human ratings. We present evidence for our novel hybrid metric being a much stronger alternative.

Human static evaluation As shown in Table 3.4, while static human evaluation suggests EI regularization is effective due to a higher number of win judgments[7], the results are noisy and difficult to interpret due to large confidence intervals and a high

[7]We follow [183] to highlight the higher value between wins/losses and reporting 90% confidence intervals.

Table 3.4: Results from human static evaluation for EI (Emotion+Infersent) vs. BL (baseline) models as measured by pairwise comparisons of **Quality** with 90% confidence intervals.

	Cornell			Reddit		
Model	Wins %	Losses %	Ties %	Wins %	Losses %	Ties %
HRED-EI	**40.8** $\pm$ 4.9	24.5 $\pm$ 4.9	34.8 $\pm$ 9.2	**31.3** $\pm$ 5.2	29.5 $\pm$ 6.6	39.3 $\pm$ 10.7
VHRED-EI	**36.9** $\pm$ 4.7	36.6 $\pm$ 5.6	26.6 $\pm$ 6.9	**39.0** $\pm$ 7.0	34.0 $\pm$ 5.3	27.0 $\pm$ 8.9
VHCR-EI	**33.0** $\pm$ 6.1	29.0 $\pm$ 5.4	38.0 $\pm$ 10.1	**33.7** $\pm$ 7.9	27.3 $\pm$ 3.3	39.0 $\pm$ 8.6

percentage of ties. The median inter-annotator agreement measured pairwise through Cohen's κ [55] for our human evaluation was only 0.176 and 0.120 for CORNELL and REDDIT respectively. This level of annotator agreement is lower than the median Cohen's κ of previous work [145] and explains the larger confidence intervals. Even after removing ambiguous examples (i.e. where equal number of annotators select each response as being better), large annotation variation persists. This may be due to subjectivity and ambiguity arising from different interpretations of <unknown> tokens or the short length of contexts in the CORNELL corpus (e.g. median length of conversation of 3 utterances). These findings further highlight the importance of an interactive evaluation as opposed to limited static responses.

3.5.4 Novel Metrics Applied to Human Data and Self-play

We examine how the novel psychologically-inspired metrics relate to the trajectories of the 100 best and 100 worst quality conversations. This is only feasible with interactive evaluation. As shown in Figure 3-6, we observe that appropriate sentiment, coherent semantics, and engaging users are indispensable to attaining high quality ratings in interactive interaction. Comparing EI and baseline conditions, we see a replication of these trends (Figure 3-7). For example, EI elicits longer responses from users (greater engagement), with more laughter and higher semantic coherence.

Figure 3-8 summarizes the relationships between interactive human ratings and the

automated metrics[8]. We observe that our sentiment metric applied to human data on its own has higher correlation with interactive human ratings than the commonly used metrics such as perplexity and embedding distance metrics. Most importantly, our novel hybrid metric, M_H, applied to self-play [9] aggregated on the model-level is strongly correlated with all human ratings ($r > .7$), while previous metrics achieved $r < .5$. This is a significant finding, suggesting that even without running interactive human evaluation, we can automatically approximate it through self-play. This metric is agnostic to the training set and model type and can be calculated on the trajectory of self-play utterances for any chatbot, regardless of its architecture. One interpretation is that the self-play framework keeps the conversation within the training set distribution, and the model is less likely to produce <unknown> tokens. Therefore, M_H and its sub-components have meaningful values and can be useful for quality approximation.

On a realistic conversation trajectory, M_H is a hybrid of conflicting objectives and thus is less susceptible to exploitation [34]. However, the purpose of the self-play metric ($\hat{M}_H$) in its current form is a post hoc evaluation of a dialog model. There are precautions if one intends to directly optimize for $\hat{M}_H$ or its sub-components, for example in a reinforcement learning scenario. The current formulation of self-play uses trajectories entirely generated by the same model. If one intends to optimize $\hat{M}_H$, we suggest calculating it on conversation trajectories between the bot and an external baseline model or a fixed copy [207], or adopting adversarial learning by maintaining a discriminator to distinguish between real/fake conversations [140]. This implicitly enforces generating realistic language. Additionally, we have shown how to successfully learn using sub-components of $\hat{M}_H$ as reward functions [112].

[8]For additional correlation results across the human metrics, between M_is and human metrics on a bot-level, and Spearman and Kendall rank coefficients, see §3.7.3, §3.7.4, and §3.7.5 respectively.

[9]Analyzing utterance overlap shows that these self-play conversations are distinct from the training corpus and exhibit high diversity for variational models. Details can be found in §3.7.13.

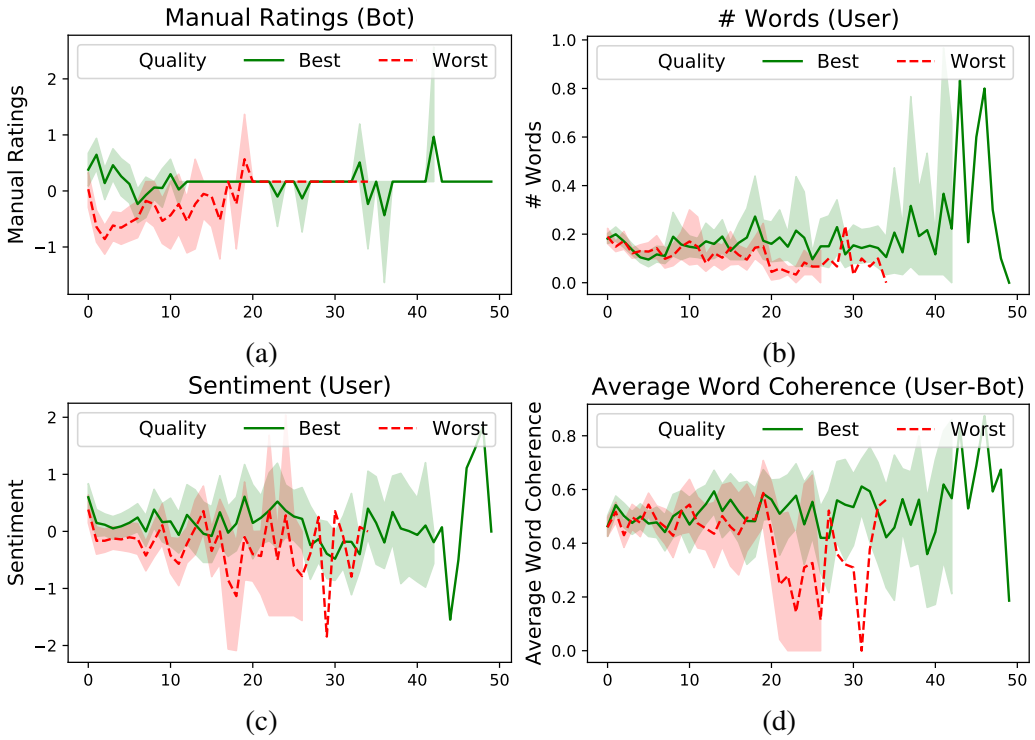


Figure 3-6: One hundred highest vs. lowest quality conversation trajectories; lines: mean, shaded area: 90% confidence intervals, x-axis: conversation turns. (a) Timing of upvote/downvote ratings: A bad first impression impedes overall rating. +1, -1, and 0 show upvotes, downvotes, and no manual feedback, respectively. (b) Participants talk longer and use more words in conversations rated higher. Number of words have been normalized between 0 and 1. (c) High-quality conversations elicit more positive user sentiment; many participants leave after expressing negative sentiment. Sentiment score ranges from -1 (the most negatively valenced emotion) to +1 (the most positively valenced emotion). (d) High-quality conversations are more semantically similar as measured by average word coherence between user query and bot responses. Users tend to leave the conversation when the bot responses are semantically dissimilar. Coherence score can range from 0 (no coherence) to 1 (maximum coherence).

3.6 Optimizing Human-centered Metrics in a Reinforcement Learning Framework

A natural question that arises is how to optimize these human-centered metrics and guard against Goodhart's law [83]. Goodhart's law suggests that "an observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." A

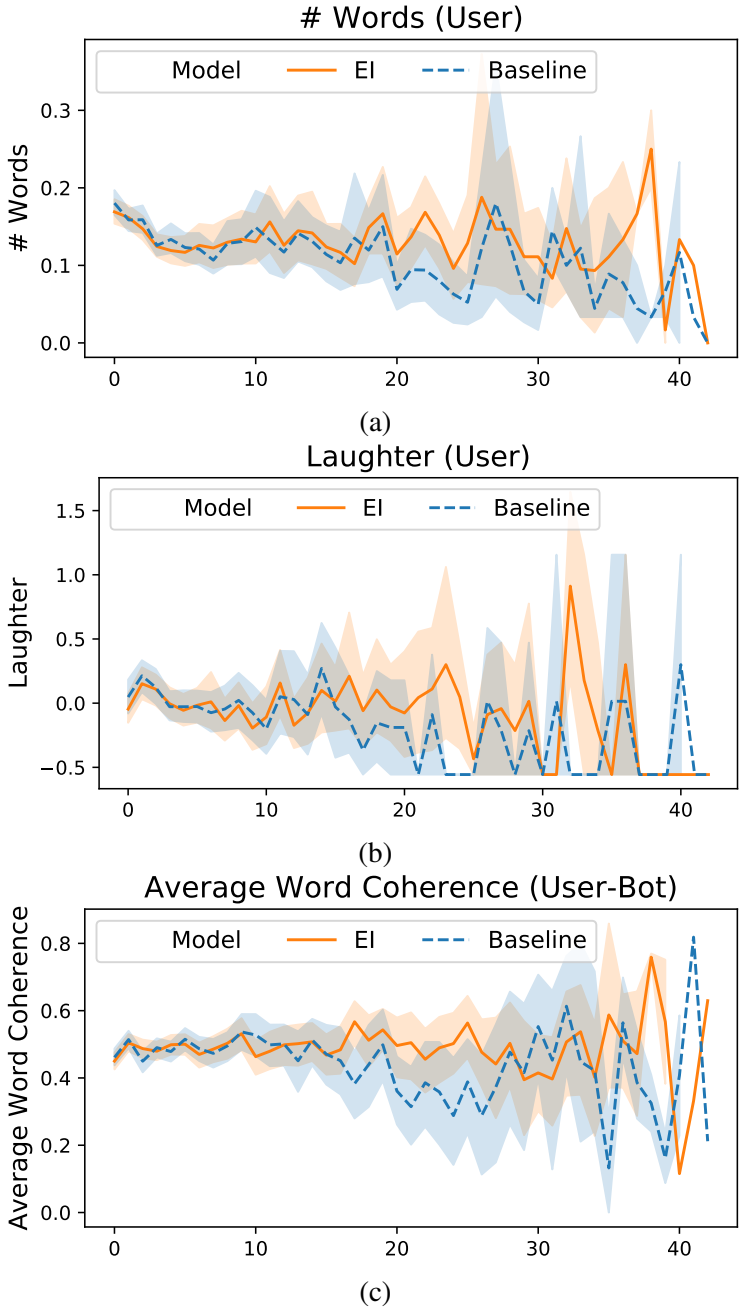


Figure 3-7: EI vs. baseline conversation trajectories; lines: mean, shaded area: 90% confidence intervals, x-axis: conversation turns. (a) EI elicits longer responses from users, suggesting that they are more engaged compared to the baseline models. (b) EI evokes more laughter from users compared to baseline. (c) EI has higher semantic coherence as measured by average word coherence.

more generalized interpretation of that is "when a measure becomes a target, it ceases to be a good measure." [235] Goodhart's law phenomenon has been observed across various

Figure 3-8: Pearson correlations between five human metrics and automated metrics. **Sentiment -U** has higher correlation with interactive human ratings than prior metrics. **Hybrid Metric $\mathbf{M_H}$ -B/B**, our novel self-play based metric, has higher correlation across all human metrics more than any other metric proposed to-date. **Notes:** -U: Calculated on user response, -B: Calculated on bot response, -U/B: Calculated between user and bot response, -B/B: Calculated between consecutive bot utterances.

domains such as neural machine translation [203, 257], summarization [185, 234], robotics [21], and beyond. For example, refining a translation model using reinforcement learning with BLEU score reward function resulted in improvements in BLEU score but did not reflect human evaluation [257]. While beam search significantly improves BLEU scores, it tends to hurt the diversity of text and amplify its biases [203]. Over-optimizing for a custom reward learned directly from human preferences ultimately becomes anti-correlated with human preferences [234]. We hypothesized that varying the trade-off between the incentive to get a higher reward against the incentive to remain close to an initial supervised model [10] or better long-term credit assignment could help overcome this challenge. This section briefly mentions our follow-up work in this space.

3.6.1 Follow Up I: Human-Centric Dialog Training via Offline Reinforcement Learning

How can we learn from human feedback to produce high-quality conversation without the risk of humans teaching it harmful chat behaviors? We approach this problem by optimizing the HC metrics we extensively studied in Section 3.4.3 using off-policy reinforcement

[10] Other researchers have also adopted this technique and reported its success [234].

learning (RL). A well-known challenge in off-policy RL is the inability to explore and over-optimistic estimation of future reward. The complexity of these problems increases as the action space and the number of reward functions grow, such as in language generation. We hypothesize that closeness to a strong pre-trained language model can generate natural and fluent language, and optimization for HC rewards can further improve qualities that matter to humans in an interactive dialog. Thus, we develop a novel class of offline RL algorithms. These algorithms use KL-control to penalize divergence from a pre-trained prior language model and use a new strategy to make the algorithm pessimistic, instead of optimistic, in the face of uncertainty. We test the resulting dialog model with ratings from 80 users in an open-domain setting and find it achieves significant improvements over existing deep offline RL approaches. The novel offline RL method combined with our proposed rewards can improve any existing generative dialog model using a static human feedback dataset. For more information about this work, see [113].

3.6.2 Follow Up II: Hierarchical Reinforcement Learning for Open-Domain Dialog

How to overcome the challenges of proper credit assignment for long-term conversational rewards? In this follow-up work, we propose a novel approach to hierarchical reinforcement learning (HRL), VHRL, which uses policy gradients to tune the utterance-level embedding of a variational sequence model. This hierarchical approach provides greater flexibility for learning long-term, conversational rewards than previous approaches that apply RL at the word-level [112, 139, 140, 196, 262]. We build our rewards functions off our findings from the hybrid HC metric discussed in Section 3.4.3 and other rewards associated with improved human judgments of conversation quality [218]. We also add a novel HC reward for minimizing the estimated toxicity of a conversation. This new reward aims to limit inappropriate, biased, and offensive responses. We use self-play and RL to optimize these HC rewards. We show that our approach provides significant improvements in both

human evaluation and automatic metrics over state-of-the-art dialog models, including Transformers. For more information about this work, see [207].

3.7 Supplementary Materials

3.7.1 Ablation models results

We conducted additional evaluations of ablations of our EI models to determine whether emotion or infersent regularization provided the most benefit. The results in Table 3.5 reveal that this depends on the dataset and the model in question. We also checked whether simply appending the emotion and infersent embedding of an utterance to the top level of the hierarchy could provide the same benefit as knowledge distillation, even though this would require retaining copies of the DeepMoji and Infersent models, and would be more computationally expensive at inference time. Table 3.5 reveals that the *input-only* models do not out-perform the knowledge-distillation EI models on automatic metrics.

Table 3.5: Automatic metrics computed on ablations of the EI models, trained with distillation from only the emotion recognition model (EI$_{emo}$), the infersent model (EI$_{inf}$), or receiving emotion and infersent only as input, without knowledge distillation (*input-only*). Whether emotion or semantics provides the most benefit depends on the dataset and the model.

		Cornell					Reddit				
Model	Version	PPL	KL	Avg	Ext	Grd	PPL	KL	Avg	Ext	Grd
HRED	baseline	52.311	-	.471	.329	.331	41.730	-	.649	.394	.474
	input only	47.911	-	.549	.381	.392	41.227	-	.644	.395	.469
	EI$_{emo}$	48.619	-	**.562**	.359	**.416**	47.395	-	.541	.310	.371
	EI$_{inf}$	47.988	-	**.562**	.381	.405	**41.083**	-	.646	.394	.472
	EI	**47.636**	-	.560	**.383**	.400	41.245	-	**.651**	**.398**	**.482**
VHRED	baseline	**49.414**	.264	.539	.352	.395	36.240	.188	.635	.383	.464
	input only	49.819	.442	.543	.353	.393	40.248	.312	.630	.377	.456
	EI$_{emo}$	51.346	.636	**.552**	**.358**	**.401**	36.212	.199	.631	.380	.458
	EI$_{inf}$	52.143	**.702**	.539	.346	.392	36.518	**.222**	**.637**	.381	.463
	EI	50.526	.517	.545	.355	.394	**35.510**	.167	.636	**.392**	**.465**
VHCR	baseline	61.000	.562	.532	.345	.382	**36.736**	.267	.619	.371	.448
	input only	50.966	.558	.531	.344	.382	37.342	**.287**	.608	.365	.431
	EI$_{emo}$	52.407	.590	.585	**.374**	.442	37.449	.254	.619	.366	.444
	EI$_{inf}$	53.085	**.575**	.544	.356	.390	37.109	.199	.629	.378	.457
	EI	**49.243**	.475	**.588**	.369	**.444**	37.198	.231	**.639**	**.394**	**.469**

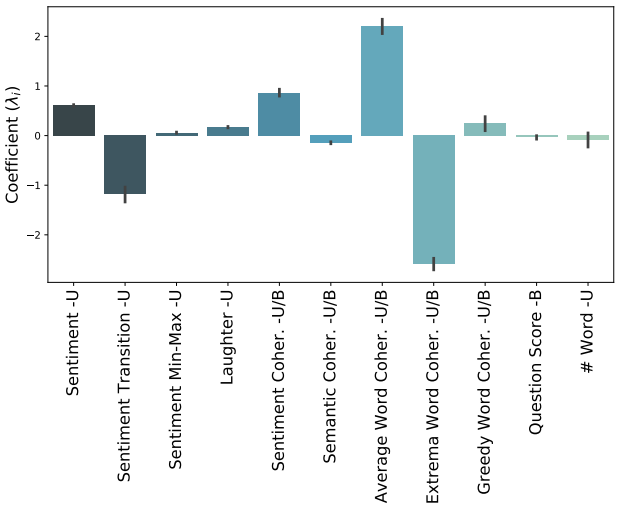


Figure 3-9: The learned coefficients (λ_i) within the hybrid metric (M_H). Using a leave-bot-out method, we observe that the λ_is are stable. The error bars show 90% confidence intervals. See Section 3.4.3 for details about calculation of these metrics.

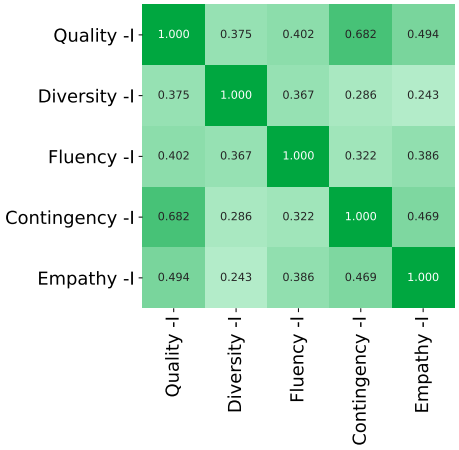


Figure 3-10: Correlation matrix showing the relationships between different aspects of interactive human evaluation. We observe a strong correlation across these aspects.

3.7.2 Hybrid metric coefficients

We optimized the coefficients of sub-components of the hybrid metric using a leave-bot-out scenario. As shown in Figure 3-9, we observe that λ_is are stable across these training iterations. However, because we have optimized a linear regression equation and some of the features have overlapping information, such as different aggregation methods for calculating word coherence, we do not suggest using λ_is for direct interpretation; further investigation is required.

3.7.3 Human interactive ratings correlation table

Figure 3-10 provides detailed information about different metrics from interactive human ratings. We observe that quality is highly correlated with other aspects of the conversation. Specifically, it is most strongly correlated with contingency, which further highlights the importance of semantic metrics of bot-generated responses in a good quality conversation. It also has high correlation with empathy.

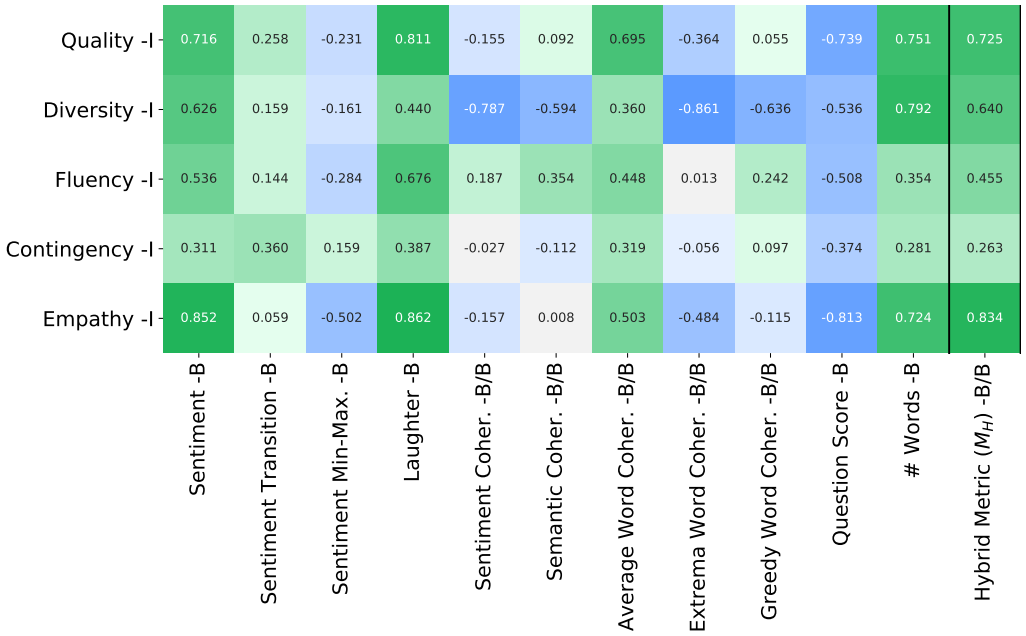


Figure 3-11: Correlation matrix showing the relationships between different automated metrics on self-play trajectories and interactive human ratings aggregated on the bot-level. We observe that inducing positive sentiment as measured by Sentiment and Laughter, and being able to generate longer sentences in self-play are associated with higher quality model ratings. It is worth mentioning that maintaining extreme similarity in sentiment or semantics or just asking questions in self-play conversation trajectories could backfire by reducing the diversity of generated responses, though applicable to interactive human data. Most importantly, our novel hybrid metric applied to self-play (M_H -B/B) is highly correlated with all human ratings of the dialog model. **Postfixes:** -I: Interactive human evaluation, -B: Calculated on bot response, -B/B: Metric applied to self-play on two consecutive bot generated utterances when the bot converses with itself. See Section 3.4.3 for details about calculation of these metrics.

3.7.4 Self-play correlation table

Figure 3-11 provides detailed information about the introduced metrics applied to self-play. We observe that several sentiment, semantic, and engagement metrics also transfer to self-play trajectories and the introduced hybrid metric, M_H, is highly correlated with human quality ratings aggregated on a bot-level. However, exploiting sentiment or semantic similarity in a self-play scenario should be avoided as it hurts ratings of the model, especially diversity of responses.

3.7.5 Additional correlation statistics

Figure 3-12 and 3-13 provide Spearman's ρ and Kendall's τ correlation coefficients between human metrics and automated metrics. These tests do not assume a linear correlation as opposed to the Pearson correlation. Similarly to the Pearson correlation results provided in Figure 3-8, these values provide additional evidence, further confirming the superiority of sentiment metric as well as the newly proposed self-play approximation of the hybrid metric M_H.

3.7.6 Reddit casual conversation corpus details

Using the 1.7 Billion post comments dataset hosted on Google BigQuery, we extracted post ids for all posts on `r/CasualConversation` from July 2018 to December 2018. For each post, we built a conversation tree of comments and subsequent replies to extract three-turn dialog. We removed links, excluded [REMOVED] and [DELETED] tag comments, and only used text before "*edit*" comments to preserve the original content in the conversation. We make this dataset available for public use [11].

3.7.7 Embedding-based metrics

Embedding Average Taking the mean word embedding of the generated sentence e_g and the target sentence e_t, the embedding average metric is the cosine distance between the two.

$$\bar{e}_t = \frac{\sum_{w \in t} e_w}{|\sum_{w' \in t} e_{w'}|} \tag{3.6}$$

$$\text{AVG}(\hat{e}_t, \hat{e}_g) = cos(\bar{e}_t, \bar{e}_g) \tag{3.7}$$

Vector Extrema The extrema vector for a sentence can be calculated by taking the most extreme value for each dimension ($e_w^{(d)}$) among the word vectors in the sentence.

[11] `https://affect.media.mit.edu/neural_chat/datasets`

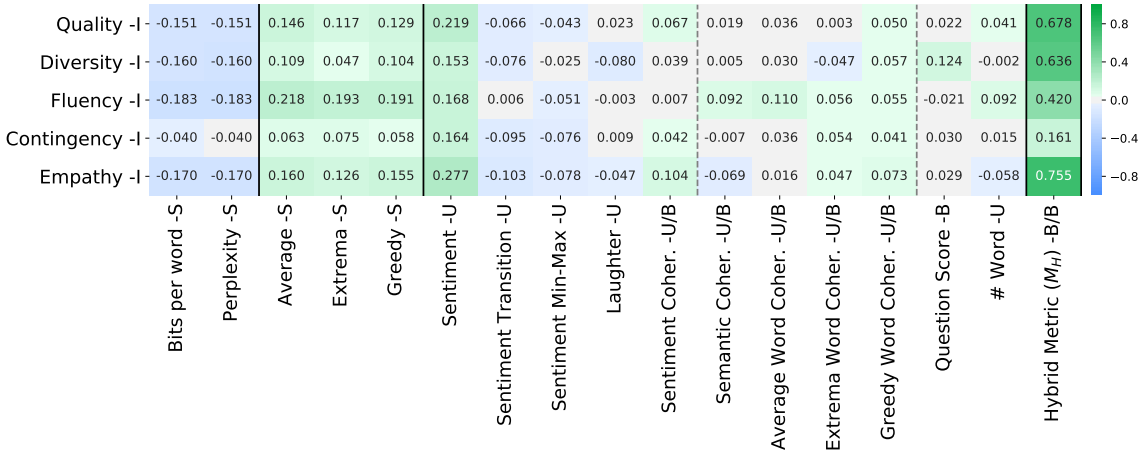


Figure 3-12: Spearman correlations between five human metrics and automated metrics. **Sentiment -U** has higher correlation with interactive human ratings than prior metrics. **Hybrid Metric $\mathbf{M_H}$ -B/B**, our novel self-play based metric, has higher correlation across all human metrics more than any other metric proposed to-date. **Notes:** -U: Calculated on user response, -B: Calculated on bot response, -U/B: Calculated between user and bot response, -B/B: Calculated between consecutive bot utterances.

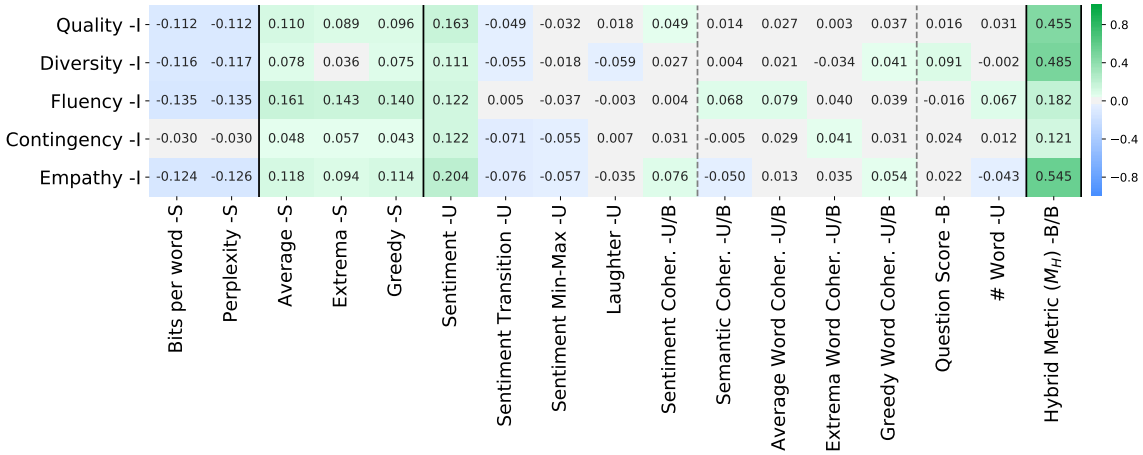


Figure 3-13: Kendall correlations between five human metrics and automated metrics. **Sentiment -U** has higher correlation with interactive human ratings than prior metrics. **Hybrid Metric $\mathbf{M_H}$ -B/B**, our novel self-play based metric, has higher correlation across all human metrics more than any other metric proposed to-date. **Notes:** -U: Calculated on user response, -B: Calculated on bot response, -U/B: Calculated between user and bot response, -B/B: Calculated between consecutive bot utterances.

The extrema embedding metric is again the cosine distance between the extrema sentence

Figure 3-14: Static single-turn evaluation interface crowdworkers see.

vectors.

$$\hat{e}_t^{(d)} = \begin{cases} \max_{w \in t} e_w^{(d)} & \text{if } e^{(d)} > |\min_{w' \in t} e_{w'}^{(d)}| \\ \min_{w \in t} e_w^{(d)} & \text{otherwise} \end{cases} \tag{3.8}$$

$$\text{EXT}(\hat{e}_t, \hat{e}_g) = cos(\hat{e}_t, \hat{e}_g) \tag{3.9}$$

Greedy Matching The greedy matching distance is computed by matching word vectors in a source sentence (s) with the closest words vectors in the target sentence(s).

$$G(r, \hat{r}) = \frac{\sum_{w \in r;} \max_{\hat{w} \in \hat{r}} cos(e_w, e_{\hat{w}})}{|r|} \tag{3.10}$$

$$\text{GRD}(s, t) = \frac{G(s, t) + G(t, s)}{2} \tag{3.11}$$

Table 3.6: Results from human static evaluation for EI vs. Baseline models for HRED, VHRED, and VHCR models across quality, fluency, relatedness and empathy pairwise comparisons with 90% confidence intervals

		Cornell			Reddit		
Model	Metric	Wins %	Losses %	Ties %	Wins %	Losses %	Ties %
HRED-EI	quality	**40.8** $\pm$ 4.9	24.5 $\pm$ 4.9	34.8 $\pm$ 9.2	**31.3** $\pm$ 5.2	29.5 $\pm$ 6.6	39.3 $\pm$ 10.7
	fluency	10.3 $\pm$ 4.4	**17.3** $\pm$ 4.1	72.5 $\pm$ 8.1	**22.8** $\pm$ 5.3	20.0 $\pm$ 7.1	57.3 $\pm$ 11.2
	relatedness	36.3 $\pm$ 6.5	28.7 $\pm$ 4.8	35.0 $\pm$ 7.9	**34.3** $\pm$ 2.8	30.3 $\pm$ 7.8	35.5 $\pm$ 9.7
	empathy	**37.8** $\pm$ 7.2	24.5 $\pm$ 5.6	37.8 $\pm$ 8.9	**32.5** $\pm$ 3.4	31.2 $\pm$ 5.9	36.3 $\pm$ 8.0
VHRED-EI	quality	**36.9** $\pm$ 4.7	36.6 $\pm$ 5.6	26.6 $\pm$ 6.9	**39.0** $\pm$ 7.0	34.0 $\pm$ 5.3	27.0 $\pm$ 8.9
	fluency	23.4 $\pm$ 9.6	**27.7** $\pm$ 8.3	48.9 $\pm$ 16.3	**29.0** $\pm$ 13.6	23.3 $\pm$ 9.3	47.7 $\pm$ 21.6
	relatedness	**37.4** $\pm$ 5.4	33.1 $\pm$ 7.2	29.7 $\pm$ 9.6	**38.3** $\pm$ 5.6	33.0 $\pm$ 5.1	28.7 $\pm$ 9.0
	empathy	**36.6** $\pm$ 9.4	34.0 $\pm$ 8.4	29.4 $\pm$ 15.8	**34.7** $\pm$ 8.7	33.7 $\pm$ 6.7	31.7 $\pm$ 10.9
VHCR-EI	quality	**33.0** $\pm$ 6.1	29.0 $\pm$ 5.4	38.0 $\pm$ 10.1	**33.7** $\pm$ 7.9	27.3 $\pm$ 3.3	39.0 $\pm$ 8.6
	fluency	13.5 $\pm$ 4.1	**25.5** $\pm$ 4.3	66.0 $\pm$ 7.7	**24.7** $\pm$ 7.2	18.3 $\pm$ 5.2	57.0 $\pm$ 10.2
	relatedness	**40.8** $\pm$ 4.8	26.8 $\pm$ 6.8	32.5 $\pm$ 10.5	28.3 $\pm$ 6.6	**31.3** $\pm$ 3.6	40.3 $\pm$ 8.4
	empathy	**32.8** $\pm$ 6.6	28.0 $\pm$ 7.8	39.3 $\pm$ 13.7	**30.3** $\pm$ 3.9	24.0 $\pm$ 4.6	45.7 $\pm$ 7.6

3.7.8 Static evaluation setup details

We replicated the static evaluation found in previous work [183, 223]. We sampled conversation contexts from the test set of each corpus and generated samples by each model based on these contexts. After filtering by context length (>10 tokens) and removing contexts which contain <unknown>tokens, we sampled 100 examples. We divided each set of 100 examples into two batches of 50 for annotators to rate. Annotators recruited through Amazon Mechanical Turk were first trained with an example question. Annotators must be in the United States and had to correctly answer all training questions before beginning the task. Figure 3-14 shows the interface displayed to crowd-workers in the static evaluation task. We asked annotators to select which sentence was better for quality, fluency, relatedness, and empathy. Note that in static single-turn evaluation, annotators only rate a single bot-generated response; thus they cannot judge the diversity of response generation in the dialog model and only rate the remaining four qualities. Table 3.6 summarizes the results for all 4 metrics and is an uncondensed version of table 3.4. One notable exception to the pattern of EI models winning is fluency; baseline models trained on the CORNELL corpus generated more fluency wins.

Noting the high disagreement between annotators in this task, we further examined the

Table 3.7: Count of ambiguous examples in human static evaluation.

	Cornell			Reddit		
	HRED	VHRED	VHCR	HRED	VHRED	VHCR
Quality	12	13	15	26	15	9
Fluency	4	10	10	12	20	6
Relatedness	11	12	10	15	13	7
Empathy	16	9	12	14	17	7

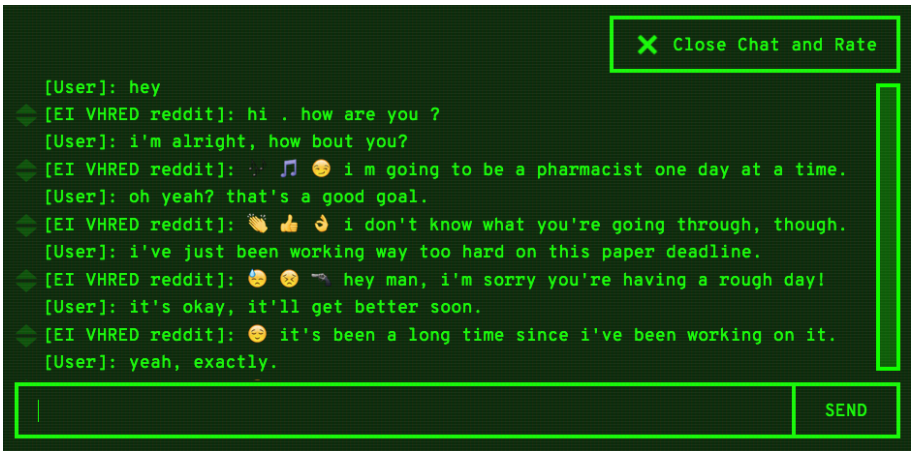


Figure 3-15: Interactive evaluation chat interface

ambiguous examples in the human evaluation test set. We define an ambiguous example as a question where an equal number of annotators select the first sentence as better as the second sentence. If the two examples were similar, annotators would select the "tied" option. An equal number of selections for each answer as the winner indicates a disagreement in perception. Table 3.7 summarizes the number of ambiguous examples per model and metric out of 100 in total for each box. After removing these ambiguous example from calculating wins, losses and ties, the results are similar to table 3.6. The number of ambiguous samples further highlights the noisy and unreliable nature of static single-turn evaluation.

Table 3.8: Summary table of number of human interactive ratings collected per model.

	Cornell			Reddit		
	HRED	VHRED	VHCR	HRED	VHRED	VHCR
Baseline	55	46	53	55	36	39
EI	49	39	42	56	44	52

3.7.9 Interactive evaluation details

For our interactive evaluation, we built a platform to mimic a natural chat setting. Figure 3-15 is an example conversation within the platform that interactive evaluation participants see. Annotators can optionally click the up and down arrows beside each chatbot response to give feedback on the specific utterance. Once 3 or more turns of the conversation has taken place, participants may click “Close Chat and Rate". This will take them to the rating page where the conversation to be rated is presented along side the 7 point Likert scale questions used to asses the conversation (Figure 3-5).

Participants both from Amazon Mechanical Turk and from the authors’ institution were recruited for interactive evaluation. Although the minimum required number of turns is 3, the average number of responses per conversation of participants varied between 3.00-10.58 turns with the average at 5.43 turns. Table 3.8 summarizes the number of ratings collected for each model.

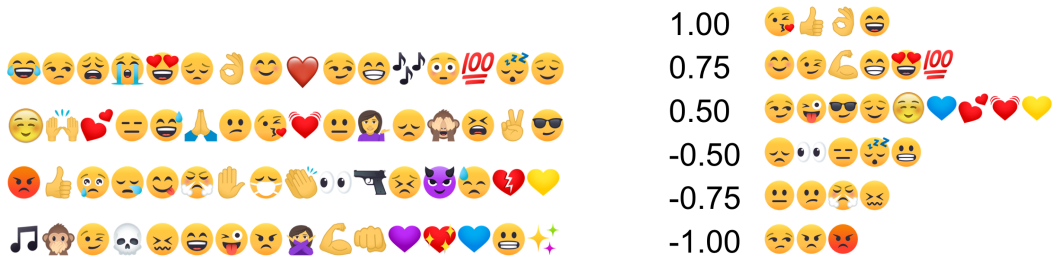
The average rating each annotator gave differed significantly between annotators. As a result, we also computed scores for interactive evaluation after normalizing each annotator’s scores. We restricted ratings down to only annotators who completed 10 or more ratings which left 301 ratings. Similar to the results without normalizing annotator scores in Table 3.2, the mean ratings for EI (Emotion+Infersent) models were higher than the mean ratings for the baseline models.

3.7.10 Website server setup and configuration

The server was hosted on a Google Cloud Platform virtual instance with 64GB of RAM and a NVIDIA Tesla P100 graphics card. The backend was a Django program being served by NGINX and uWSGI. For simplicity, we opted to have the Django process import the chatbots into the same Python process as Django, rather than have the two connect to each other via other means such as sockets. This configuration decreased development time and increased reliability, but it would need to be revisited if the server needed to scale several orders of magnitude past what was required for this study. The current configuration was still able to support hundreds of simultaneous users and host more than 30 bots concurrently.

The chatbots were kept in a separate project from the Django project and maintained separately from the server code. Each chatbot extended an abstract class that defined key methods for the Django program to use, and was registered to a globally accessible dictionary via a decorator. The Django project was provided the path to the Chatbots project in its PYTHONPATH, so it could import the dictionary in which all the chatbot objects had been registered and use that to dynamically determine which chatbots were available and to access them in its views.

It is important to note that the chatbots used PyCUDA, and PyCUDA does not work in a multiprocessing environment. Because of this, uWSGI needed to be configured to only have one python process and to disable any attempt at multiprocessing. Furthermore, the chatbots required substantial startup times, so all chatbots are kept in memory at all times in the Django process. In order to keep all the chatbots in memory concurrently, we needed a very high amount of RAM on our server and opted for a 64GB virtual instance, and a GPU with 16GB RAM. This combination of CUDA to run the chatbots on the GPU with a high amount of RAM to keep all bots in memory at the same time resulted in incredibly fast server response times, with effectively no increase in response time when using the bots in requests compared to requests that did not.



(a) (b)

Figure 3-16: (a) 64-most frequent emojis as predicted by [52] used for calculating emotion embeddings. (b) Assigned weights used for reducing the 64-dimensional emotion embedding into a *Sentiment* score.

For further information and instructions on server configuration, please read the server documentation available at `https://github.com/asmadotgh/neural_chat_web`.

3.7.11 Emotion embedding details

We calculate emotion embeddings of an utterance using a using a state-of-the-art sentiment-detection model [52]. This pre-trained model outputs a probability distribution over 64 most-frequently used emojis as presented in [52]). We define a set of weights over the emojis and calculate the weighted sum over an emotion embedding vector to derive a *Sentiment* score which is higher for positive sentiment and lower for negative sentiment (See Figure 3-16).

3.7.12 Hyper-parameter tuning details

For the baseline models that were trained on the CORNELL dataset, we used the parameters reported in [183, 222, 223] that achieved state-of-the-art results for HRED, VHRED, and VHCR models trained on the same dataset, respectively. For EI models, we compared a combination of values for encoder hidden size (400, 600, 800, 1250), decoder hidden size (400, 600, 800, 1250), context size (1000, 1250), embedding size (300, 400, 500),

word drop (0, .25), sentence drop (0, .25), beam size (1, 5). Learning rate (.0001), dropout (.2) were fixed. Batch size 80 was used. If due to memory limitation the job was not successfully completed, batch size 64 was used. Additionally, we tuned the EI parameters, i.e., emotion weight (25, 150), infersent weight (25K, 30K, 50K, 100K), emotion sizes (64, 128, 256), infersent sizes (128, 1000, 2000, 4000). Due to limited computational resources, we were not able to run a grid search on the aforementioned values. Instead we used combinations of the parameters that heuristically were more viable.

For the models that were trained on the REDDIT dataset, a set of properly tuned baseline parameters were non-existent. Thus, to ensure fair comparison, we used a similar approach for baseline and EI hyper-parameter tuning: We explored a combination of values for encoder hidden size (400, 600, 800, 1250), decoder hidden size (400, 600, 800, 1250), context size (1000, 1250), embedding size (300, 400, 500, 600), word drop (0, .25), sentence drop (0, .1, .25), and beam size (1, 5). Learning rate (.0001), dropout (.2) were fixed. Batch size 64 was used. If due to memory limitation the job was not successfully completed, batch size 32 was used. Due to limited computational resources, we were not able to run a grid search on all the aforementioned values. Instead we used combinations of the parameters that heuristically were more viable. To ensure fair comparison, any selected combination was tested for both baseline and EI models. Then, for EI models, we tuned the parameters that were solely relevant to the EI design, such as the weight of emotion and infersent term in the loss function and the size of the added discriminator networks: Emotion weight (25), infersent weight (25K, 50K, 100K), emotion sizes (64, 128, 256), infersent sizes (100, 128, 1000, 2000, 4000). See Table 3.9 for a summary of the final selected parameters.

3.7.13 Self-Play Overlap Analysis

As a post hoc sanity check on the conversations generated from self-play, we check whether there is i) overlap among generated conversations, and ii) overlap between these

Table 3.9: Hyper-parameters used for different models.

Dataset	Version	Model	Batch size	Dropout	Decoder hidden size	Encoder hidden size	Context size	Embedding size	Word drop	Sentence drop	Beam size	Emotion weight	Emotion discriminator layer size	Infersent weight	Infersent discriminator layer size
Cornell	Baseline	HRED	80	.2	400	400	1000	300	.0	.0	5	-	-	-	-
		VHRED	80	.0	1000	1000	1000	400	.25	.0	5	-	-	-	-
		VHCR	80	.2	1000	1000	1000	500	.25	.25	5	-	-	-	-
	EI	HRED	64	.2	1000	1000	1000	500	.0	.0	1	25	128	100K	4000
		VHRED	80	.2	1250	1250	1000	600	.0	.0	1	25	128	30K	128
		VHCR	32	.2	1000	1000	1250	600	.0	.0	1	25	128	25K	4000
Reddit	Baseline	HRED	64	.2	1000	1000	1000	500	.0	.0	1	-	-	-	-
		VHRED	32	.2	1250	1250	1000	600	.0	.0	1	-	-	-	-
		VHCR	32	.2	1000	1000	1250	600	.0	.25	1	-	-	-	-
	EI	HRED	64	.2	1000	1000	1000	500	.0	.0	1	25	128	25K	2000
		VHRED	32	.2	1250	1250	1250	600	.0	.0	1	25	128	100K	4000
		VHCR	32	.2	1000	1000	1250	600	.0	.0	1	25	128	100K	4000

Table 3.10: Percentage of pairs of conversations in each 100 sample for each model where there are 3 or 5 consecutive conversation turns that are exactly the same.

		Cornell		Reddit	
Model	Version	3-turn overlap	5-turn overlap	3-turn overlap	5-turn overlap
HRED	baseline	19.49%	1.76%	2.02%	0.24%
	EI	6.48%	0.30%	2.12%	0.16%
VHRED	baseline	0%	0%	0%	0%
	EI	0.16%	0%	0.16%	0%
VHCR	baseline	0%	0%	0%	0%
	EI	0%	0%	0%	0%

conversations and the training set. High overlap among generated conversations would indicate that there is a lack of diversity in the conversations generated by self-play while high overlap with the training set suggests self-play may be memorizing training dialog.

To measure overlap between the 100 conversations generated in each model, we consider all 3 and 5 consecutive conversational turns over the 10 turns in each conversation. We compare each pair of conversations in the 100 generated conversations in total to

Table 3.11: Percentage of of conversations (100 sample for each model) where there are 2 or 3 consecutive conversation turns that match the training set.

		Cornell		Reddit	
Model	Version	2-turn overlap	3-turn overlap	2-turn overlap	3-turn overlap
HRED	baseline	58%	0%	0%	0%
	EI	65%	0%	0%	0%
VHRED	baseline	8%	0%	5%	0%
	EI	5%	0%	12%	0%
VHCR	baseline	4%	0%	4%	0%
	EI	3%	0%	3%	0%

compute a percentage of conversations which contain overlap in this pairwise comparison. Table 3.10 summarizes these results and illustrates that overlap is not significant for most models. The exception is the non-variational models trained on the Cornell corpus (e.g. HRED Cornell). Qualitative evaluation reveals that these are degenerate cases where "what?" or "I don't know" or "I'm sorry" are repeated multiple turns.

To measure repetition with respect to the training set, we take all 2-turn and 3-turn windows in the self-play generated conversations and compare with the entire training set to check whether there is overlap. Table 3.11 shows the percentage of conversations (100 total for each model) where there is a 2-turn or 3-turn dialog appearing exactly in the training set. Since each conversation is 10 turns long, all of the conversations are distinct from the training set and no conversation contains more than 2-turns of overlap with the training set. The 2-turn overlap again appears due to cases where "what?" and "hi" are repeated for 2 turns.

3.8 Conclusions

A major obstacle in open-domain dialog generation is the predominant optimization of an objective function that does not closely match human judgment of conversation quality in a naturalistic chat. To alleviate this problem, we have combined interactive human data with psychologically-motivated measures and introduced a novel hybrid metric to

reflect human-centered optimality criteria. Using this metric in a self-play framework provides results that are strongly correlated with human judgment of chatbot empathy ($r > .8$) and quality ($r > .7$), and perform significantly better than the state of the art (r=.44, p<0.001) [149]. Additionally, we have demonstrated a significant improvement to several hierarchical seq2seq generative models using regularization of the utterance level of the hierarchy with knowledge distillation. Finally, we have open-sourced the platform together with a new REDDIT dataset. In follow-up works, we have successfully used the proposed metrics as reward functions in hierarchical and non-hierarchical off-policy learning scenarios and showed they resulted in higher-quality dialog based on human evaluation and automatic metrics compared to several baselines.

3.9 Statement of Contributions

A significant collaborative effort in Roz Picard's Affective Computing group has led to the results presented in this chapter. The work that I led along with Judy Shen and Natasha Jaques on approximating interactive evaluation with self-play is only a portion of this enormous effort. Discussing research ideas in depth and implementing this work has been truly collaborative and would not have been possible without my colleagues. Natasha initiated research in this direction, originally intending to utilize reinforcement learning to elicit positive *response* from *the user*. My contributions include investigating topic modeling, implementing interfacing with sentiment and semantic regularization, proposing and developing several sub-metrics, proposing and developing the self-play aspect, conducting the exploratory analyses, and preparing and obtaining COUHES approvals. I have also participated in training and evaluating several chatbots deployed on our platform and rated by crowd workers on Amazon Mechanical Turk (AMT). Natasha implemented EI regularization and took a major role in shaping the direction of this work. Judy conducted the human subject studies on AMT and calculated the statistical analyses. Craig Ferguson designed and implemented the `neural.chat` website. Noah Jones advised on

psychological aspects of empathy and good conversation that guided implementation of the hybrid metric that constitutes a major contribution of this work. Agata Lapedriza and Roz Picard provided guidance and advice throughout the project.

Natasha and Judy led the follow-up work with off-policy RL. Natasha proposed and implemented the modification with KL control, implemented baselines, and conducted most of the analyses. Judy ran additional AMT studies and analyses and shaped the framing of the work. My contributions include proposing and implementing Monte Carlo dropout estimates to overcome the over-optimism problem of off-policy RL, setting up additional AMT studies, and modification of the web platform for hosting new chatbots.

Abdul Saleh and Natasha led the hierarchical RL work, ranging from ideation to implementation. I advised on the direction and implementation of hierarchical RL work.

Chapter 4

Interpretability Benefits of Uncertainty Quantification

Supporting model interpretability for complex phenomena where annotators can legitimately disagree, such as emotion recognition, is a challenging machine learning task. In this work, we show that explicitly quantifying the *uncertainty* in such settings has interpretability benefits. We modify a classical network inference using Monte Carlo dropout to measure different types of uncertainty, such as epistemic and aleatoric uncertainty. Epistemic uncertainty is measured by the lack of confidence in one's knowledge and is attributed to missing information about the learning task. Aleatoric uncertainty is representative of an event's propensity and is attributed to the stochastic behavior of observations. We identify a significant correlation between aleatoric uncertainty and human annotator disagreement ($r \approx .3$). Additionally, we demonstrate how subjective and difficult training samples can be identified using aleatoric uncertainty and how epistemic uncertainty can reveal data bias resulting in unfair predictions. In addition to explainability benefits, we observe modest performance boosts from incorporating model uncertainty.

4.1 Introduction

Supporting interpretability of an automated prediction system in complex tasks where human experts disagree is a challenging machine learning problem. In such settings, answering the following questions can help understand model's predictions: is the model uncertain due to capturing annotator biases and their subjective perspective? Or is it error-prone for a specific set of samples due to a distribution shift from the training data? Can the predicted confidence scores of the model be trusted? Do they represent the true likelihood so that we can intuit and reason about their results?

Emotion understanding is an icon for a learning setting where label ambiguity abounds. Most researchers agree that *emotion* in itself is nuanced and the same input could be assigned different labels due to change in contextual information or the perspective of the reviewer [57]. Thus, disambiguating annotator and data bias and quantifying how well predictive confidence can be trusted is crucial to supporting explainability in emotion classification.

In this work, we extend beyond deterministic modeling of affect using Monte Carlo (MC) dropout [59], a technique that requires no changes to the neural network architecture and only minimal changes at inference time. This approach augments classification's per-class confidence scores with measures of uncertainty. We tease apart elements in the uncertainty estimates and investigate how each helps in interpreting model predictions and failure modes. We show this decomposition results in a proxy for inter-rater disagreement capturing annotators' bias, and a proxy highlighting bias in data that could potentially result in unfair predictions. These insights lead to better interpretation of model behavior and pave the way for HC optimality.

These techniques can also boost performance. We report significant improvement in Jensen-Shannon divergence (JSD) between predicted and true class probabilities. We show a strong correlation between total uncertainty and JSD ($r \approx .6$), identifying it as a proxy for performance. We study the influence on accuracy, especially if given the option to

reject classifying samples where the model lacks confidence.

To summarize, we use MC dropout with traditional neural network architectures and explore the benefits of resulting measures of uncertainty while disambiguating their source. Our contributions include: 1) introducing a proxy for inter-annotator disagreement, 2) demonstrating the power of such metrics in identifying difficult samples and bias in training data along with ways to alleviate them, 3) showing improvements in performance in addition to interpretability benefits.

4.2 Background & Related Work

Understanding what a model does not know is especially important to explain and understand its predictions. State-of-the-art classification results are mostly achieved by deep neural networks (DNN)—such as AlexNet, VGGNet, ResNet, etc.—that are deterministic in nature and not designed to model uncertainty. Bayesian Neural Networks (BNN) have been an alternative to DNNs, providing a distribution over model parameters at an extra computational cost while increasing difficulty of conducting inference [37, 151]. These computational challenges hinder scalability of BNNs.

MC dropout [59] has been introduced as an approximation of BNNs that can be achieved by keeping the same architecture of a deterministic DNN and only making minimal changes at inference time. Dropout, i.e. randomly dropping weights at training time, is commonly used in DNNs as a regularization method. Drawing random dropout masks at test time can approximate a BNN. Recently, [118] demonstrated ways to additionally learn the observation noise parameter σ, thus modeling epistemic and aleatoric uncertainty in parallel. Epistemic uncertainty represents lack of confidence in one's knowledge attributed to missing information about the learning task. Aleatoric uncertainty is attributed to the stochastic behavior of observations. They evaluated the approach for use in regression tasks in depth estimation. A partial version of the model, only modeling aleatoric uncertainty, was evaluated for classification for semantic segmentation.

These efforts show great potential at empowering deterministic DNNs with Bayesian properties with negligible computational overhead. However, we will show that complex, difficult tasks where reviewers disagree and data may not fully represent everyone, such as affect detection, can benefit from inferring different sources of uncertainty. Despite its importance, latent uncertainty quantification in emotion detection tasks is under-explored. However, there have been a few efforts regarding more realistic emotion recognition by incorporating explicit inter-annotator disagreement. For example, modeling *perception* of uncertainty as measured by the standard deviation of labels captured from crowd-sourced annotations has been studied in [92]. While such efforts are valuable in affective computing applications, these approaches are supervised, are prone to error when annotations are sparse and varied in number, are not capable of capturing uncertainty introduced by model parameters or sources of noise other than human judgment.

4.3 Technical Approach

The underlying architecture of our model is an Inception-ResNet-v1 [239] for extracting facial features, followed by a multi-layer perceptron for emotion classification. We built upon an open-source implementation [210] of FaceNet [215]. We pre-trained the model up to the `Mixed-8b` layer using cross-entropy loss on face identity classes using the CASIA-WebFace dataset [261]. The pre-processing step included using a Multitask CNN [264] to detect facial landmarks and extract facial bounding boxes in the form of 182×182 pixel images. Since the utility of this training mechanism is to identify faces, it learns to ignore features that are invariant to one's identity, e.g. facial expressions, in the later layers of the network while the earlier layers represent lower-level features. `Mixed-7a` best encoded and retained emotionally-relevant information based on our experiments (See §4.5.1).

4.3.1 Baseline

After extracting features from layer `Mixed-7a`, a fully-connected network with two hidden layers was used to infer basic emotions. We refer to this model as *Baseline*. Facial Expression Recognition (FER) is an established emotion detection dataset [82]. FER+ is the same set of images, expanded to include at least 10 annotations from crowd-sourced taggers [5]. We used FER+ train, private test, and public test subsets for training, hyper parameter tuning, and evaluation of our model performance, respectively. See §4.5.1 for details.

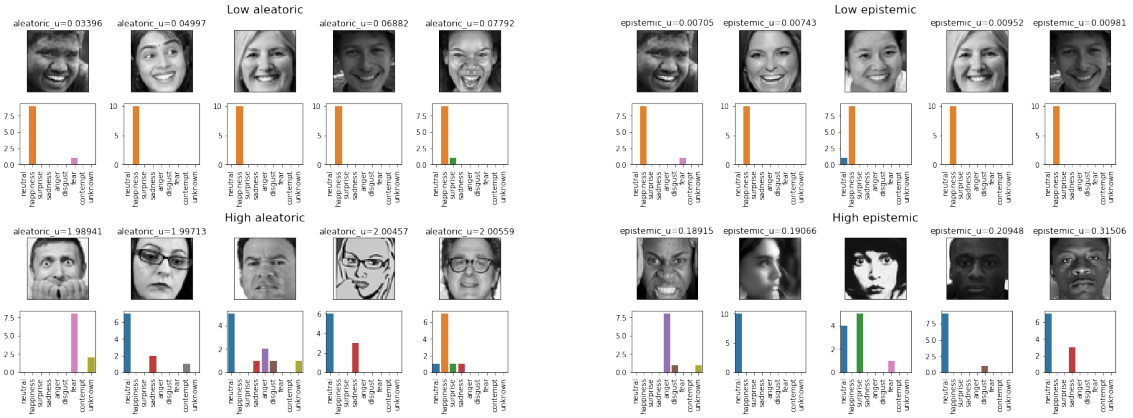


Figure 4-1: **Left**: Aleatoric uncertainty (U_a) - Samples with lowest U_a are stereotypical expressions of emotion where annotators (almost) unanimously agree on the assigned label. Conversely, images with the highest U_a either represent subjectivity involved in human annotations or low image quality, e.g. when the face is occluded by hands or the image is a drawing as opposed to a photograph. **Right**: Epistemic uncertainty (U_e) - Samples with lowest U_e show stereotypical expressions of emotion that are common in the training set. On the other hand, images with the highest U_e include dark-skinned subgroups, a non-frontalized photo, and a highly illuminated image, even when there is near-perfect agreement across human-annotators. We believe this is due to the skewed pre-training dataset, suggesting that it is not equipped to encode such samples.

4.3.2 Epistemic & Aleatoric Uncertainties

For each input image, *Baseline* predicts a length-C logits vector z which is then passed through a softmax operation to form a probability distribution p over a set of class labels.

For our new model, we move away from pointwise predictions, and put a Gaussian prior distribution over the network weights, $W \sim N(0, I)$. To overcome the intractability of computing the posterior distribution $p(W|X, Y)$, we use MC dropout [59], performing dropout both during training and test time before each weight layer, and approximate the posterior with the simple distribution q_θ^W. Here, q_θ^W is a mixture of two Gaussians, where the mean of one of the Gaussians is fixed at zero. We minimize the Kullback-Leibler (KL) divergence between q_θ^W and the $p(W|X, Y)$: $\mathcal{L}(\theta, p) = \frac{1}{N}\sum_{i=1}^{N} \log p(y_i|x_i, \theta, X, Y) + \frac{1-p}{2N}||\theta||^2$, where N is the number of data points, p is dropout probability, q_θ^W is the dropout distribution, and $\hat{W}_t \sim q_\theta^W$. Using MC integration with T sampled dropout masks, we have the approximation: $p(y = c|x, X, Y) \approx \frac{1}{T}\sum_{t=1}^{T} \frac{e^{z^{\hat{W}_{c,t}(x)}}}{\sum_{c=1}^{C} e^{z^{\hat{W}_{c',t}(x)}}}$.

Inspired by [153], we use entropy in the probability space as a proxy for classification uncertainty. To get an aggregate uncertainty measure, we marginalize over all parameters and use the entropy of the probability vector p: $H(p) = -\sum_{c=1}^{C} p_c \log p_c$. We then quantify the total (U_t) and aleatoric uncertainty (U_a) using:

$$U_t \approx H[E_{q(\theta|X,Y)}[p(y|x, \theta]] \approx H[\frac{1}{T}\sum_{t=1}^{T} p(\hat{y}|x, \hat{W}_t)];$$

$$U_a \approx E_{q(\theta|X,Y)}[H[p(y|x, \theta)]] \approx \frac{1}{T}\sum_{t=1}^{T} H[p(\hat{y}|x, \hat{W}_t)]$$

The epistemic uncertainty U_e is then defined $U_t - U_a$. Note that U_e will represent mutual information between true values and model parameters and thus has a different scale compared to U_a and U_t that each represent entropy of a probability distribution. For simplicity, we refer to this model as *UncNet* in the rest of the chapter. The code is available at `https://github.com/asmadotgh/unc-net`.

4.4 Results & Discussion

We show that modeling and disambiguating different sources of uncertainty provides a means to identify data that are more difficult to classify, and seek to provide interpretable reasons for why. Similar to [195], to represent task subjectivity, we compute the probability that two draws from the empirical histogram of human annotations disagree: $d_i = 1 - \sum_{c=1}^{C} p_{i,c}^2$, where C is the number of classes and $p_{i,c}$ is the probability of image i being rated as class c.

4.4.1 A Proxy for Inter-Rater Disagreement

Classification of perceived emotions is inherently a subjective task, with disagreement across human annotators. We hypothesize that aleatoric uncertainty is associated with inter-annotator disagreement. We used the Pearson correlation coefficient to assess the relationship between aleatoric uncertainty (U_a) and disagreement probability (d_i), resulting in a significant correlation: $r = 0.301, p \ll .001$. This finding suggests aleatoric uncertainty as a tool for quantifying degree of label subjectivity associated with an image.

Note that we observed no significant correlation between epistemic uncertainty and the annotators' disagreement probability: $r = -0.027$, $p = 0.105$. This is aligned with our hypothesis that epistemic uncertainty captures the uncertainty introduced by model parameters and is not able to capture the nuance in subjective annotations.

4.4.2 Task Subjectivity, Difficulty & Bias in Training

Figure 4-1 shows samples with the highest and lowest uncertainties. On the left, extreme cases in terms of aleatoric uncertainty (U_a) are listed. We observe that samples with low U_a are stereotypical expressions of emotion where annotators (almost) unanimously agree on the assigned label. The fact that "happiness" class is the second most common class in the dataset (after "neutral"), and has a stereotypical morphology in terms of the position of the eye corners, mouth, and teeth exposure may have contributed to the dominance of

“happiness" class in low U_a samples. On the other hand, we observe that samples with highest U_a either represent subjectivity involved in label assignment and lack of annotators' consensus; or low quality of an image. For example, the face occlusion or being a drawing as opposed to a photograph.

Figure 4-1, on the right, shows extreme cases in terms of epistemic uncertainty (U_e). Low U_e samples show similar patterns: samples with stereotypical expression of emotion that are common in the training set. On the other hand, we see different patterns in samples with high U_e. We observe that the model has low confidence in the predictions for dark-skinned subgroups. Our interpretation is that the CASIA-WebFace dataset that was used for pre-training the model is highly skewed. It contains faces of celebrities that IMDB lists as active between 1940 and 2014. Most of these celebrities are white. That may explain why the model has high U_e in making a prediction for non-white input images. We also see a sample that exemplifies a non-frontalized photo, which the human annotators were able to unanimously assign a “neutral” label despite its atypical viewpoint in the dataset. Since the pre-training process included a frontalization pre-processing step, we believe the current model is not capable of finding meaningful representations for non-frontalized photos and that is why this sample has high U_e. Factors such as different illumination may also result in higher U_e.

4.4.3 Performance

The Brier score [15, 50, 141] is a commonly-used metric for quantifying the accuracy of probabilistic predictions: $B = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}(y_{n,c} - \hat{y}_{n,c})^2$. N is the number of samples, C is the number of classes, y is a one-hot representation of true labels, and $\hat{y}$ is the predicted confidence scores. Since we have multiple annotations per data point, each pair of <annotation, sample> is treated separately. First, we confirm that samples with lower uncertainty measured by the MC dropout approach also have lower Brier scores. We sort samples based on predictive uncertainty estimates and calculate the correlation

between Brier score and U_a, U_e, and U_t. There was a significant Pearson correlation between each of these pairs: $r_{U_a,BCE} = .880, p \ll .00001$; $r_{U_e,BCE} = .710, p \ll .00001$; $r_{U_t,BCE} = .884, p = .00007$. This confirms that lower uncertainty measured by the MC dropout approach is associated with better probabilistic accuracy as measured by the Brier score.

We also hypothesized performance gains using *UncNet*. Due to task subjectivity and annotation spread (§4.5.2), we believe measures that rely on a binary true/false assumption for evaluation do not fully represent the nuance of our problem setting. Therefore, we use Jensen-Shannon divergence to quantify the distance between predicted and true class probabilities: $JSD(p, \hat{p}) = \frac{KL(p||m)+KL(\hat{p}||m))}{2}$. Here, m is the point-wise mean of p and $\hat{p}$ and KL is the Kullback-Leibler divergence. Lower $JSD(p, \hat{p})$ represents better performance. A paired-samples t-test was conducted to compare the JSDs in ***Baseline*** and *UncNet*. There was a significant difference in JSD for ***Baseline*** ($M = 0.473, SD = 0.131$) and ***UncNet*** ($M = 0.461, SD = 0.140$); $t(3578) = 9.335, p \ll .001$, confirming our hypothesis.

We take a more granular look and hypothesize that samples with higher uncertainty have higher $JSD(p, \hat{p})$. To test this, a Pearson correlation coefficient was computed to assess the relationship between U_a, U_e, U_t and JSD in *UncNet*. Each pair showed a significant correlation ($p \ll .00001$): $r_{U_a,JSD} = .583$; $r_{U_e,JSD} = .100$; $r_{U_t,JSD} = .591$. This finding further confirms our hypothesis: lower uncertainty is associated with a better match between prediction and groundtruth. Similar to findings of [118], we see aleatoric uncertainty plays a more significant role in such identification.

Though accuracy may not fully represent this nuanced problem setting, we also checked how *UncNet* compared to the ***Baseline*** as measured by accuracy. We observed that *UncNet* has the potential to improve performance modestly, but that if the model had the option to reject classifying samples it is not confident in up to 25%, it improves significantly in performance, by as much as 8%. See §4.5.4 for details.

Additionally, Figure 4-2 shows our investigation of the model calibration through

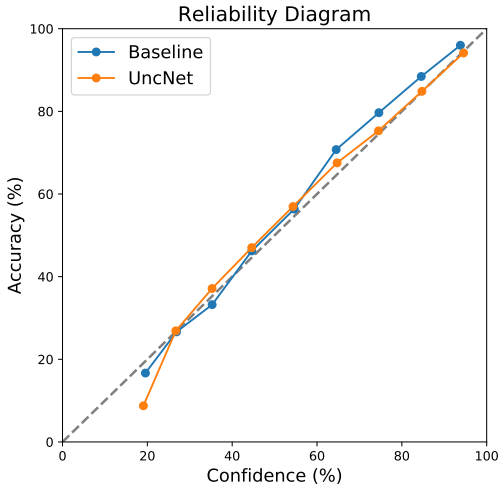


Figure 4-2: Reliability diagram for *Baseline* and *UncNet* of FER+ hold out test data [5]. Soft-labels result in well-calibrated predictions.

the reliability diagram for both *Baseline* and *UncNet*. As plotted, both models are close to the 45° line. This is aligned with previous research findings showing evidence of well-calibrated predictions when trained with soft labels [173, 220]. We observe that the near-perfect calibration in the *Baseline* does not leave space for further improvement.

4.5 Supplementary Materials

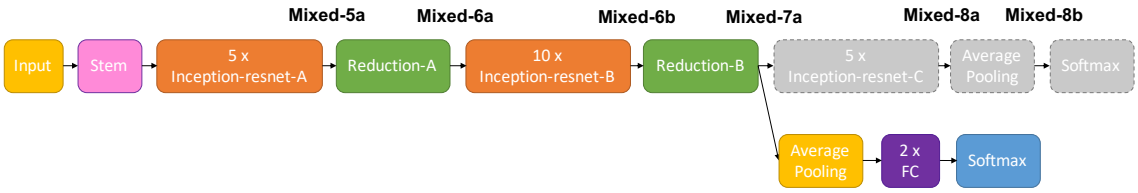


Figure 4-3: Model architecture: An Inception-ResNet-v1 followed by an average pooling layer and a fully-connected network with two hidden layers (FC). Pre-training on CASIA-WebFace dataset has been conducted on the full Inception-ResNet-V1. We froze the weights of the network and used up to the `Mixed-7a` layer to extract features from raw images. The remaining unused layers of Inception-ResNet-v1 are in grey. We then stack two FCs on the `Mixed-7a` layer after average pooling. Dropout is only applied to the FC layers.

Table 4.1: Validation accuracy and loss of predicting facial expression emotions on FER+ dataset, using the features extracted from different layers of FaceNet, pre-trained on two different datasets: CASIA-WebFace and VGGFace2.

		Mixed-5a	Mixed-6a	Mixed-6b	Mixed-7a	Mixed-8a	Mixed-8b
Accuracy (%)	CASIA-WebFace	49.85	54.27	52.81	**55.75**	52.45	52.50
	VGGFace2	50.60	54.80	55.11	**55.50**	50.69	50.46
Loss	CASIA-WebFace	1.589	1.516	1.555	**1.514**	1.580	1.554
	VGGFace2	1.607	1.527	1.519	**1.513**	1.589	1.598

4.5.1 Model Architecture and Pre-Training Details

Figure 4-3 shows the detailed model architecture. Note that the modules on the top represent an Inception-ResNet-v1 architecture. We have used up to layer `Mixed-7a` for feature extraction from raw images and added a fully connected (FC) network with two hidden layers of size 128×128. This represents the *Baseline* architecture. The main difference in *UncNet* is adding a dropout mask before each layer of FC, not only during training, but also at inference time.

For face similarity pre-training, we treated the intermediate layer that was used to export features from a raw image as a hyper-parameter that was tuned according to validation loss. Our experiments included `Mixed-5a`, `Mixed-6a`, `Mixed-6b`, `Mixed-7a`, `Mixed-8a`, and `Mixed-8b` layers. We observed that the `Mixed-7a` layer best encoded and retained emotional information from the input face crop. Table 4.1 summarizes our exploration results.

4.5.2 Annotation Disagreement Details

Figure 4-4 shows the distribution of disagreement probability (d_i) for all images in the training set. Histogram heights show a density rather than the absolute count, so that the area under the fitted curve is one.

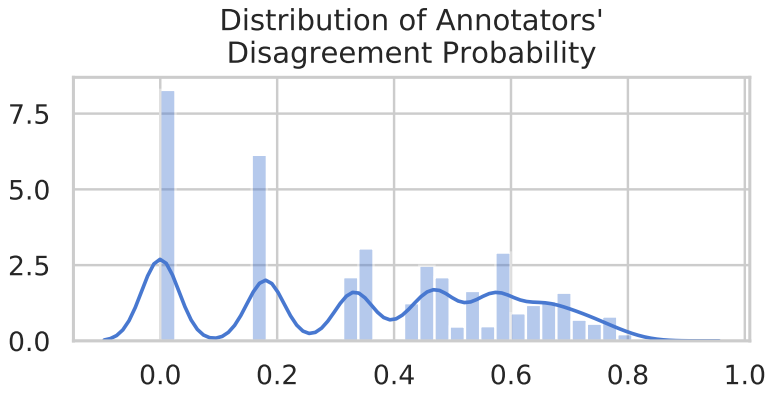


Figure 4-4: Distribution of annotators' disagreement probability (d_i) on FER+ training samples. The histogram heights are scaled to represent density rather than absolute count, so that the area under the fitted curve is one.

4.5.3 Detailed Calibration Results

In this section, we report a range of calibration scores for *Baseline* and *UncNet*. Further, we show how these scores are related to the predictive uncertainty estimates of *UncNet*.

Scholars have introduced a range of calibration scores. Maximum Calibration Error (MCE) and Expected Calibration Error (ECE) approximate calibration error by quantization of uncertainty bins and have been adopted in many recent publications [90]:

$$MCE = \max_{b=1}^{B} |acc(b) - conf(b)|$$

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)|$$

Here n_b is the number of predictions in bin b, N is the number of samples, $acc(b)$ is the accuracy of prediction in bin b, and $conf(b)$ is the average prediction confidence score in bin b. Recently, new metrics have been proposed to overcome the limited assumption of mutually exclusiveness of classes and improve robustness to label noise [176]. Static Calibration Error (SCE) is a metric where prediction for all classes is taken into account as opposed to only the argmax of softmax outputs. Adaptive Calibration Error (ACE)

is an extension of SCE where instead of equidistant bins, confidence scores are sorted and their percentiles represent “ranges", parallel to “bins" in SCE. Thresholded Adaptive Calibration Error (TACE) is an extension to ACE where values with at least ϵ confidence are taken into account. SCE, ACE, and TACE can be formally defined as the following:

$$SCE = \frac{1}{C}\sum_{c=1}^{C}\sum_{b=1}^{B}\frac{n_{bc}}{N}|acc(b,c) - conf(b,c)|$$

$$ACE = \frac{1}{CR}\sum_{c=1}^{C}\sum_{r=1}^{R}|acc(r,c) - conf(r,c)|$$

$$TACE = \frac{1}{CR}\sum_{c=1}^{C}\sum_{r\in R_\prime}|acc(r,c) - conf(r,c)|$$

$$where \quad \forall r \in R_\prime : conf(r,c) > \epsilon$$

Table 4.2: Summary of additional calibration error metrics for *Baseline* vs. *UncNet*. Near-perfect calibration with soft-labels and dependency of these metrics on quantization may be potential reasons for having inconclusive results.

Calib. Error	ECE (%)	MCE (%)	SCE (%)	ACE (%)	TACE (%)	BCE
Baseline	2.330	**6.231**	0.479	**0.390**	**0.393**	0.661
UncNet	**1.876**	10.326	**0.417**	0.526	0.506	**0.649**

Table 4.2 summarizes these metrics using B/R=10. We did not observe any conclusive results comparing *Baseline* and *UncNet* conditions or using uncertainty quantiles. Our interpretation is that the close-to-perfect calibration with soft-labels, as well as identified problems with the dependence of these metrics on quantization may have resulted in a null result. Further study in this area is required to better understand what these metrics can and cannot capture.

4.5.4 Detailed Performance Metrics

In this section, we report accuracy for a random run on the test set. Accuracy is defined as the percentage of samples where predicted maximum probability class maps to the annotated maximum probability class. Table 4.3 summarizes our findings. For future, we will add further performance metrics such as average precision or per-class accuracy and provide confidence bounds using bootstrapping.

Table 4.3: Summary of performance metrics for *Baseline* vs. *UncNet* and how it is influenced if given the possibility of rejecting classification of certain samples. U_e: Epistemic uncertainty, U_a: Aleatoric uncertainty, U_t: Total uncertainty.

Model	**Evaluation Dataset**	**Accuracy (%)**
Baseline	FER+ Test	54.848
UncNet	FER+ Test	56.943
UncNet - low U_e	75% of FER+ Test	57.452
UncNet - low U_a	75% of FER+ Test	62.481
UncNet - low U_t	75% of FER+ Test	62.332

4.6 Limitations and Future Work

This chapter investigated the relationship between decomposition of uncertainty using the Monte Carlo dropout technique and inter-rater disagreement in one particular domain, facial expression estimation. While we hypothesize these observations to be generalizable to other domains and techniques, further studies are required to confirm this hypothesis. For future work, we would like to consider other methods of uncertainty estimation ranging from Naive Bayes classifiers to weight uncertainty estimation [10] and other datasets such as COCO [144] and multi-annotator ImageNet [243].

Another avenue for future work is strengthening the results through additional quantitative approaches. An example experiment could be designed to confirm that epistemic and aleatoric uncertainties follow patterns that a secondary model can predict better than

random chance and are not just artifacts of the Monte Carlo sampling process. Another future experiment could complement the qualitative results with more quantitative approaches. For example, by labeling image attributes such as pose or skin tone using other classifiers, one can verify the patterns observed among the samples with the highest or lowest epistemic or aleatoric uncertainties quantitatively.

4.7 Conclusion

One of the dimensions of moving towards HC optimality is improving interpretation of models by humans. In this chapter, we focused on the often subjective task of perceived emotion classification and demonstrated how a classical network architecture can be altered to predict measures of epistemic and aleatoric uncertainties and how these measures can help interpretation of model's confidence scores. We presented evidence for aleatoric uncertainty being a proxy for inter-annotator disagreement and showcased how the measured aleatoric uncertainty can identify low quality inputs or more subjective samples. Additionally, we presented explorations of how epistemic uncertainty can represent bias in training data and suggest directions to alleviate that. Our results suggest that the predicted total uncertainty can act as a surrogate for degree of calibration, even on tasks without human-expert consensus. Finally, we showed there are other benefits such as potential performance improvements.

4.8 Statement of Contributions

I started broadly exploring uncertainty in the affective computing context when I began an internship at Google Research. I am grateful to my host Brendan Jou, and my co-host, Brian Eoff. Though this thesis does not include my work during that internship, it influenced this work's direction after returning to MIT. I framed this work, implemented it, and conducted the analyses. I reached out to Brendan and Brian to consult with their

experience, and Roz continued her advisement throughout the project.

Chapter 5

DISSECT: Disentangled Simultaneous Explanations via Concept Traversals

Explaining deep learning model inferences is a promising venue for scientific understanding, improving safety, uncovering hidden biases, evaluating fairness, and beyond, as argued by many scholars. Using counterfactual generation for investigating a classifier's decisions, one can ask: what if this sample were to be classified as the opposite class, and how would it differ? This is one of the principal benefits of counterfactual reasoning about what does not and cannot exist in the data, a quality that many other mediums of explanation such as heatmaps and influence functions are inherently incapable of doing. However, most previous work on generative explainability cannot disentangle important concepts effectively, produces poor quality or unrealistic examples, or fails to retain relevant information. We propose a novel approach, DISSECT, that trains a generator, a discriminator, and a concept disentangler simultaneously to overcome such challenges using little supervision. Our method generates Concept Traversals (CTs), defined as a sequence of generated examples with increasing degrees of concepts that influence a classifier's decision. By training a generative model from a classifier's signal, DISSECT offers a way to discover a classifier's inherent "notion" of distinct concepts automatically rather than rely on user-predefined concepts. We show that DISSECT produces CTs that (1) disentangle several concepts

that are influential to a classifier's decision, resulting in multiple distinct explanations that uncover blind-spots of alternative approaches providing a single plausible explanation (2) are coupled to the classifier's reasoning, due to joint training (3) are realistic (4) preserve relevant information, (5) and are stable across similar inputs. We validate our approach on several challenging synthetic and realistic datasets where previous methods fall short of satisfying desirable criteria for interpretability and show that our method performs consistently well across all. Finally, we discuss applications of DISSECT for detecting potential biases of a classifier, investigating its alignment with expert domain knowledge, and identifying spurious artifacts that impact predictions using simulated experiments.

5.1 Introduction

Many scholars have argued for promises of deep learning explainability from improving safety to evaluating fairness and beyond [31, 45, 78, 155]. Many efforts in explainability methods have been working towards providing solutions for this challenging problem. One way to categorize them is by the medium of explanations, some post hoc techniques focusing on the importance of individual features, such as saliency maps [51, 150, 233, 237], some on importance of individual examples [120, 122, 128, 260], some on importance of high-level concepts [123]. There has been active research into the shortcomings of explainability methods (e.g. [1, 111, 193, 224, 253]). For example, it has been shown that attention weights can be manipulated without hurting accuracy and result in misleading interpretations [193], adversarially constructed dissimilar attention distributions can lead to similar predictions [111], and some existing saliency methods are independent of the model and the data generating process [1] which renders them unfit for explaining the relationship between inputs and learned outputs. Scholars have also proposed tests to determine when attention can be used as an explanation [253].

These methods focus on information *that already exist* in the data–either by weighting features or concepts in training examples or by selecting important training examples.

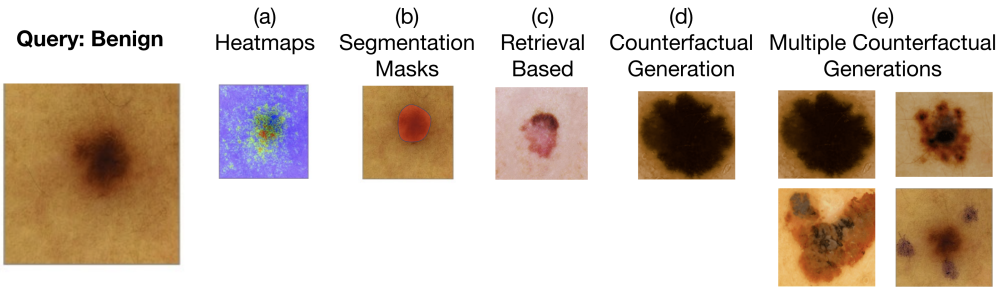


Figure 5-1: Applying explainability methods to a melanoma classifier in the dermatology domain. (a) explanation by heatmaps such as [51, 150, 233, 237]. (b) explanation by segmentation masks such as [76, 213]. Both heatmaps and segmentation masks only provide partial information. They might hint at what is influential within the sample, potentially focusing on the lesion area. However, they cannot show what kind of changes in color, texture, or inflammation could transform the input at hand from benign to malignant. (c) explanation by sample retrieval such as [228]. A retrieval-based technique might show input samples of malignant skin lesions that have similarities to a benign lesion in patient A, but from a different patient B, potentially from another body part or even a different skin tone. Such examples do not show what this benign lesion in patient A would have to look like if it were classified as malignant instead. (d) explanation by counterfactual generation such as [209, 230]. This method depicts *how* to modify the input sample to change its class membership. A counterfactual explanation visualizes what a malignant tumor could look like, in this case, by increasing the diameter of the lesion. (e) explanation by multiple counterfactual generations such as DISSECT. Multiple counterfactuals could highlight several different ways that changes in a skin lesion could reveal its malignancy and overcome some of the blind spots of a single explanation. For example, they can demonstrate that large lesions, jagged borders, and asymmetrical shapes lead to melanoma classification. They can even show potential biases of the classifier by revealing that surgical markings can spuriously lead to melanoma classification.

Guided by recent progress in generative models [27, 101, 124, 131, 147], another family of explainability methods has emerged that provide explanations by *generating* new examples or features [38, 117, 209, 230]. These methods aim to highlight particular aspects or factors contributing to a classifier's decision using generated examples or by producing counterfactuals.

One of the key benefits of *counterfactual* generation is allowing users to explore scenarios through what does not and cannot exist in the data, making them an excellent

tool for making classifier decisions plausible [247]. Using counterfactual generation for investigating a classifier's decisions, one can ask: what if this sample were to be classified as the opposite class, and how would it differ? For the rest of this chapter, I will refer to these types of questions for investigating a predictor as "what-if" questions [1]. Humans also justify decisions via counterfactuals [3], and children learn through a similar process [7, 18, 250]. Additionally, in-depth user studies have shown that examples have been the most preferred means of explanations by users across visual, auditory, and sensor data domains [115].

To illustrate the added benefits of counterfactual explanations, consider a dermatology task where an explanation method is used to highlight why a certain sample is classified as benign/malignant (Fig. 5-1). Explanations through mediums like heatmaps, saliency maps, or segmentation masks only provide partial information. Such methods might hint at what is influential within the sample, potentially focusing on the lesion area. However, they cannot show what kind of changes in color, texture, or inflammation could transform the input at hand from benign to malignant. Retrieval-based approaches that provide examples that show a concept are not enough for answering "what-if" questions either. A retrieval-based technique might show input samples of malignant skin lesions that have similarities to a benign lesion in patient A, but from a different patient B, potentially from another body part or even a different skin tone. Such examples do not show what this benign lesion in patient A would have to look like if it were classified as malignant instead. On the other hand, counterfactuals depict *how* to modify the input sample to change its class membership. A counterfactual explanation visualizes what a malignant tumor could look like in terms of potential color or texture, or inflammation changes on the skin. Better yet, multiple counterfactuals could highlight several different ways that changes in a skin lesion could reveal its malignancy.

Most previous work on generative explainability has focused on providing a single

[1] Note that this chapter uses this line of questioning only to refer to the predictor-under-test, which is different from the underlying data generative process. Therefore, causal claims about the real-world data cannot be derived from this method.

plausible explanation [209, 230]. However, a single explanation is not enough in many use-cases (Fig. 5-1). One scenario is recourse generation for decision-making systems. For example, consider a model deciding on granting loans to applicants. An individual that has been denied a loan aims to improve their unfavorable outcome. A single counterfactual explanation might suggest increasing their education. However, the individual might not be able to afford further investment in their education but may be capable of reducing their debt or increasing their income. Having multiple counterfactual explanations allows the user to incorporate their constraints and choose from more than one option to improve their chances of receiving a loan. Recourse generation has gained significant attention, and the benefits of multiple explanations have become evident. Several scholars have attempted to address this issue [46, 171]. Another use case is when explanations serve knowledge discovery and education. Providing several explanations depicts a more comprehensive view by showing the different possible ways that a classification outcome could flip instead of just converging to the most common one. For example, there are several ways that a benign skin lesion could become malignant. Another example is using explanations as an auditing tool before deploying a model to check its alignment with domain knowledge. Multiple distinct explanations can help uncover potential failure points that might be indistinguishable if merged into a single explanation. For example, consider a model that is influenced by the meaningful color/texture/size changes for classifying skin lesions as benign/malignant, but also relies on surgical markings [256] to make its decision. A single explanation might fail to reveal this flaw resulting from dependence on spurious features, but multiple distinct explanations shed light on this phenomenon.

Some of the most consistently agreed-upon properties desired for an explainability method include diversity, compatibility, realism, substitutability, and stability. Diversity [162] suggests that inputs should be representable with non-overlapping concepts. Compatibility with classifier [230], or classification model consistency [231] means that changing the explanation should produce the desired outcome from the classifier. Realism or data consistency [230] suggests that perturbed samples should lie on the data manifold to be

consistent with actual data. In other words, the generated samples should look realistic when compared to real samples. Substitutability suggests that explanations should preserve relevant information [209]. This quality has sometimes been referred to as fidelity [162, 192]. Stability [75, 162, 192] refers to the coherence of explanations for similar inputs. As we show in this chapter, current counterfactual generation techniques fail to satisfy these desired properties simultaneously. This call for developing a new method that can satisfy all of these properties.

In this work, we develop a generation-based explainability method that attempts to solve the challenges mentioned above. DISSECT generates *Concept Traversals* (CTs). We define a CT as a sequence of generated examples with increasing degrees of concepts that influence a classifier's decision. CTs are generated by jointly training a generator, a discriminator, and a CT disentangler, together to generate examples that (1) express one distinct factor at a time that is influential to a classifier's decision, (2) are coupled to the classifier's reasoning, due to joint training (3) are realistic (4) preserve relevant information, (5) are stable across similar inputs. We compare DISSECT with several baselines, some of which have been optimized for disentanglement, some used extensively for explanation, and some that fall in between. DISSECT is the only technique that performs well across all the aforementioned dimensions. Other baselines either have a hard time with influence, lack fidelity, generate poor quality and unrealistic samples, or are not disentangling properly.We evaluate DISSECT using `3D Shapes` [19], `CelebA` [146], and a new synthetic dataset inspired by the challenges faced in the dermatology domain [65]. We show that DISSECT successfully addresses all of these challenges. We also discuss this work's applications to detect a classifier's potential biases using a simulated experiment.

This chapter makes five main contributions: 1) presents a counterfactual explanation approach that manifests several desirable properties outperforming baselines, 2) demonstrate applications through experiments showcasing the effectiveness of this approach for detecting potential biases of a classifier, 3) presents a set of explainability

baselines inspired by approaches used for generative disentanglement, 4) translates desired properties commonly referred to in the literature across mediums of explanation into measurable quantities applicable to bench-marking and evaluating counterfactual generation approaches, 5) releases a new synthetic dataset inspired by the challenges faced in the dermatology domain. The code for all the models and metrics is publicly available at `https://github.com/asmadotgh/dissect`.

5.2 Related Work

We focus on reviewing post hoc explainability methods. One way to categorize them is by the medium of explanation. While met with criticisms [1, 232], many feature-based explainability methods exist [51, 150, 237] that assign a weight to each input feature to indicate their importance in classification. Example-based methods are another popular category [120, 122, 128, 260] that instead assign importance weights to individual examples. More recently, concept-based methods have emerged that attribute weights to concepts, i.e., higher-level representations of features [76, 85, 123] such as "long hair." Some of these methods have to provide multiple explanations [25, 179].

Our work leverages recent progress in the generative modeling community [81, 126], where the explanation is presented through several conditional generations [168, 178]. Efforts for the "discovery" of concepts are also related to learning disentangled representations [27, 101, 123, 124], which is particularly challenging without additional supervision [147]. Recent findings suggest that weak supervision in the form of how many factors of variation have changed [227], using labels for validation [148], or using sparse labels during training [148] are a few approaches that could make the problem identifiable. While some techniques like Conditional Subspace VAE (CSVAE) [127] started to look into conditional disentanglement by incorporating labels, their performance is far from their unconditional counterparts. Unlike previous work, we include weak supervision via the posterior probability and gradient of the classifier-under-test to aid discovery.

Many explainability methods have emerged that use generative models to modify existing examples or generate new examples [38, 40, 86, 117, 209, 230]. Most of these efforts use pre-trained generators and generate a new example by moving along the generator's embedding. Some aim to generate samples that would flip the classifier's decision [40, 117], while others aim to modify particular attribution (e.g., gender) of the image and observe the classifier's decision change [38]. While examples generated by an independently trained generator may look realistic, it might be missing a key piece– it is not explicitly coupled to the classifier's inner workings that it is trying to explain. More recent work addresses these issues by allowing the classifier's signal such as its predicted probabilities or gradients to flow through the generator during training [209, 230]. However, most assume that there is only one path that crosses the decision boundary, and they generate examples along that path. A single explanation would be adequate for simple classifiers/datasets. However, we argue that many classifiers trained on complex real datasets encode much more complex rationale that could benefit from decomposing the path into many paths or concepts.

Our work brings together the best of both worlds: we leverage a counterfactual explanation technique [230] that allows "what-if" questioning and successfully generates influential high-quality samples. Then, we endow it with qualities of disentanglement to promote diversity of explanations, by identifying multiple distinct paths/concepts to influence the classifier's decision. A sequence of generated examples expresses each concept called a Concept Traversal (CT). A generator trained while having access to the classifier's signal produces these examples.

5.3 Methods

Consider a classifier $f\colon X \to Y$ such that $f(x) = p(y|x)$ where $x \in X$ and $y \in Y$. We want to find K concepts that contribute to the decision-making of f. Consider a query

sample $x \sim p_{data}$, where p_{data} is the data distribution, a desired posterior probability[2] $\alpha \in \{0, \frac{1}{N}, \ldots, 1\}$, and a concept index $k \in \{1, 2, \ldots, K\}$. Given x, α, and k, we want to generate an image, $\bar{x}$, by perturbing latent concept k, such that the posterior probability $f(\bar{x}) = \alpha$. Putting it together, the function that generates $\bar{x}$ is defined as $I(x, \alpha, k; f) : X \times \{0, \frac{1}{N}, \ldots, 1\} \times \{1, 2, \ldots, K\} \rightarrow X$, where $f(I(x, \alpha, k; f)) \approx \alpha$. We define the k-th Concept Traversal (CT_k) as the series of $\bar{x}$'s generated from I, with k-th concept and an ordered set of α's, each of which resulting in a monotonic change in the $f(I(x, \alpha, k; f))$.

5.3.1 Baseline I: Multi-modal Explainability through VAE-based Disentanglement

Disentanglement approaches have demonstrated practical success in learning representations that correspond to factors for variation in data [227], though some gaps between theory and practice remain [147]. However, the extent to which these techniques can aid post hoc explainability in conjunction with an external model is not well understood. Thus, we consider a set of baseline approaches based on VAEs explicitly designed for disentanglement: β-VAE [101], Annealed-VAE [20], and DIPVAE [132]. We extend each of them to incorporate the classifier's signal during their training processes for a fair comparison with DISSECT. Intuitively speaking, this encourages the generative model to learn latent dimensions that could influence the classifier, i.e., learning *Influential* CTs.

More formally, consider a vanilla VAE that has an encoder e_θ with parameters θ, a decoder d_ϕ with parameters ϕ, and the M-dimensional latent code z with prior distribution $p(z)$. Recall that x denotes the input sample. The objective of a VAE is to minimize the loss:

$$\mathcal{L}_{\theta,\phi}^{\text{vanilla VAE}} = -\mathbb{E}_{z \sim e_\theta(z|x)}[\log d_\phi(x|z)] + \mathbb{KL}(e_\theta(z|x)||p(z)).$$

We introduce an additional loss term for incorporating the black-box classifier's signal:

[2]Following [230], we discretize the interval $[0, 1]$ into $N + 1$ steps with $\frac{1}{N}$ increments.

$1/K \sum_{k=1}^{K} \partial f(x)/\partial z_k$. We impose this only for the first K dimensions in the latent space[3], in other words, the number of desired CTs, $K \leq M$. Minimizing this term provides CT_ks, $k \in \{1, 2, \ldots, K\}$, with negative $\partial f(x)/\partial z_k$, with a high $|\partial f(x)/\partial z_k|$. The final loss is:

$$\mathcal{L}_{\theta,\phi} = \mathcal{L}_{\theta,\phi}^{\text{vanilla VAE}} + \lambda * \frac{\sum_{k=1}^{K} \partial f(x)/\partial z_k}{K},$$

where λ is a hyper-parameter. We apply this modification to the four aforementioned VAE-based approaches and refer to them with a -mod postfix, e.g., β-VAE-mod.

5.3.2 Baseline II: Multi-modal Explainability through Conditional Subspace VAE

Another relevant area of work is conditional generation. In particular, Conditional subspace VAE (CSVAE) is a method aiming to solve unsupervised learning of features associated with a specific label using a low-dimensional latent subspace that can be independently manipulated [127]. CSVAE partitions the latent space into two parts: w learns representations correlated with the label, and z covers the remaining characteristics for data generation. An assumption of independence between z and w is made. To explicitly enforce independence in the learned model, we minimize the mutual information between Y and Z. CSVAE has proven successful in providing counterfactual scenarios to reverse unfavorable decisions of an algorithm, also known as algorithmic recourse [46]. To adjust CSVAE to explain the decision-making of an external classifier f, we treat the predictions of the classifier as the label of interest.

More formally, the generative model can be summarized as:

$$w|y \sim N(\mu_y, \sigma_y^2.I), y \sim Bern(p),$$
$$x|w, z \sim N(d_{\phi_\mu}(w, z), \sigma_\epsilon^2.I), z \sim N(0, \sigma_z^2.I)$$

[3]Without loss of generality, the additional term can be applied to the first K dimensions, and there is no need to consider $\binom{K}{M}$ potential selections.

Conducting inference leads to the following objective function:

$$M_1 = \mathbb{E}_{D(x,y)}[-\mathbb{E}_{q_\phi(z,w|x,y)}[\log p_\theta(x|w,z)]+\mathbb{KL}(q_\phi(w|x,y)||p_\gamma(w|y))+\mathbb{KL}(q_\phi(z|x,y)||p(z))-\log p(y)]$$

$$M_2 = \mathbb{E}_{q_\phi(z|x)} D(x)[\int_Y q_\delta(y|z)\log q_\delta(y|z)\, dy]$$

$$M_3 = \mathbb{E}_{q(z|x)D(x,y)}[q_\delta(y|z)]$$

$$\min_{\theta,\phi,\gamma} \beta_1 M_1 + \beta_2 M_2; \quad \max_{\delta} \beta_3 M_3$$

5.3.3 Baseline III: Multi-modal Explainability through Progressive Exaggeration

Explanation by Progressive Exaggeration (**EPE**) [230] is a recent successful generative approach that learns to generate one series of counterfactual and realistic samples that change the prediction of f, given data and the classifier's signal. It is particularly relevant to our work as it explicitly optimizes *Influence* and *Realism*. EPE is a type of Generative Adversarial Network (GAN) [81] consisting of a discriminator (D) and a generator (G) that is based on Projection GAN [168]. It incorporates the amount of desired perturbation α on the outcome of f as:

$$\mathcal{L}_{\text{cGAN}}(D) = -\mathbb{E}_{x\sim p_{data}}[\min(0, -1 + D(x,0)] - \mathbb{E}_{x\sim p_{data}}[\min(0, -1 - D(G(x,\alpha),\alpha))] \tag{5.1}$$

$$\mathcal{L}_{\text{cGAN}}(G) = -\mathbb{E}_{x\sim p_{data}}[D(G(x,\alpha),\alpha)] \tag{5.2}$$

A Kullback–Leibler divergence (KL) term in the objective function between the desired perturbation (α) and the achieved one ($f(G(x,\alpha))$) promotes Importance [230]:

$$\mathcal{L}_f(D,G) = r(D, G(x,\alpha)) + \mathbb{KL}(\alpha|f(G(x,\alpha))), \tag{5.3}$$

where the first term is the likelihood ratio defined in projection GAN [81], and [230] uses an ordinal-regression parameterization of it.

A reconstruction loss and a cycle loss promote self-consistency in the model, meaning that applying a reverse perturbation or no perturbation should reconstruct the query sample:

$$\mathcal{L}_{\text{rec}}(G) = ||x - G(x, f(x))||_1 \tag{5.4}$$

$$\mathcal{L}_{\text{cyc}}(G) = ||x - G(G(x, \alpha), f(x))||_1. \tag{5.5}$$

Thus, the overall objective function of EPE is the following:

$$\min_G \max_D \lambda_{\text{cGAN}}\mathcal{L}_{\text{cGAN}}(D, G) + \lambda_f\mathcal{L}_f(D, G) + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}(G) + \lambda_{\text{rec}}\mathcal{L}_{\text{cyc}}(G), \tag{5.6}$$

where λ_{cGAN}, λ_f, and λ_{rec} are the hyper-parameters.

Note that EPE only finds one pathway to switch the classifier's outcome. We argue that classifiers learned from challenging and realistic datasets will have complex reasoning pathways that could enhance model explainability if revealed. Decomposing this complexity is needed to make reasoning comprehensible for humans. We compare DISSECT to a more powerful baseline, an EPE-variant, **EPE-mod**. EPE-mod learns multiple pathways by making the generator conditional on another variable: the CT dimension. More formally, EPE-mod updates $G(\cdot, \cdot)$ to $G(\cdot, \cdot, k)$ in Eq. (5.1)-(5.5), while Eq. (5.6) remains unchanged.

5.3.4 Our Method: Enforcing Distinctness of Discovered Concepts

We build our proposed method on EPE-mod and further promote distinctness across CTs by adding a disentangler network, R. The disentangler is a classifier with K classes. Given a pair of $< x, x' >$ images, R tries to predict which $\text{CT}_{k;k\in\{1,\ldots,K\}}$ has perturbed query x to produce x'. Note that R can return close to 0 probability for all classes if x' is just a reconstruction of x, indicating no tweaked dimensions. The disentangler also penalizes

the generator if any CTs use similar pathways to cross the decision boundary. See the appendix for schematics of our method.

To formalize this, let:

$$\hat{x}_k = G(x, f(x), k)$$
$$\bar{x}_k = G(x, \alpha, k)$$
$$\tilde{x}_k = G(\bar{x}_k, f(x), k).$$

Note that $\hat{x}_k$ and $\tilde{x}_k$ are reconstructions of x while $\bar{x}_k$ is perturbed to change the classifier output from $f(x)$ to α. Therefore, x, $\bar{x}_k$ and $\tilde{x}_k$ form a cycle, and k represents CT$_k$. $R(.,.)$ is the predicted probabilities of the perturbed concept given a pair of examples, which is a vector of size K, where each element is a value in $[0, 1]$. We define the following cross entropy loss that is a function of both R and G:

$$\begin{aligned}\mathcal{L}_r(G, R) &= CE(\mathbf{0}, R(x, \hat{x}_k)) + CE(\mathbf{0}, R(x, \tilde{x}_k)) + CE(e_k, R(x, \bar{x}_k)) + CE(e_k, R(\bar{x}_k, \tilde{x}_k)) \\ &= -\mathbb{E}_{x \sim p_{data}} \sum_{k=1}^{K} [e_k log R(x, \bar{x}_k) + e_k log R(\bar{x}_k, \tilde{x}_k)]\end{aligned} \tag{5.7}$$

Here, $\mathbf{0}$ refers to a vector of size K with all zeros, and e_k refers to a one-hot vector of size K where the k-th element is one and the remaining elements are zero. This term enforces R to identify no change when receiving reconstructions of the same image as input and utilizes the cycle and promotes determining the correct dimension when a non-zero change has happened, either increasing or decreasing the outcome of f. In summary, the overall objective function of our method is:

$$\begin{aligned}\min_{G,R} \max_{D} [&\lambda_{\text{cGAN}} \mathcal{L}_{\text{cGAN}}(D, G) + \lambda_f \mathcal{L}_f(D, G) + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}(G) + \lambda_{\text{rec}} \mathcal{L}_{\text{cyc}}(G) \\ &+ \lambda_r \mathcal{L}_r(G, R)]\end{aligned} \tag{5.8}$$

For this adversarial min-max optimization, we use the Adam optimizer [125].

5.4 Experiments

We evaluate DISSECT on several datasets (e.g. [19, 146]) and experimental designs.

5.4.1 Datasets

3D Shapes

We first use `3D Shapes` [19], a synthetic dataset composed of 480K 3D shapes procedurally generated from 6 ground-truth factors of variation. These factors are floor hue, wall hue, object hue, scale, shape, and orientation. Note that this dataset is purely for validation and demonstration purposes due to the controllability of all these factors. We also include datasets inspired by real-world problems (Section 5.4.1) as well as real datasets (Section 5.4.1).

Figure 5-2: Illustration of `SynthDerm` dataset that we algorithmically generated. Fitzpatrick scale of skin classification based on melanin density and corresponding samples representing different characteristics in the dataset are visualized.

SynthDerm

Second, we create a new dataset, SynthDerm (Figure 5-2). Real-world characteristics of melanoma skin lesions in dermatology settings inspire the generation of this dataset [242]. These characteristics include whether the lesion is asymmetrical, its border is irregular or jagged, is unevenly colored, has a diameter more than 0.25 inches, or is evolving in size, shape, or color over time. These qualities are usually referred to as the ABCDE of melanoma [202]. We generate SynthDerm algorithmically by varying several factors: skin tone, lesion shape, lesion size, lesion location (vertical and horizontal), and whether there are surgical markings present. We randomly assign one of the following to the lesion shape: round, asymmetrical, with jagged borders, or multi-colored (two different shades of colors overlaid with salt-and-pepper noise). For skin tone values, we simulate Fitzpatrick ratings [54]. Fitzpatrick scale is a commonly used approach to classify the skin by its reaction to sunlight exposure modulated by the density of melanin pigments in the skin. This rating has six values, where 1 represents skin that always burns (lowest melanin) and 6 represents skin that never burns in sunlight (highest melanin). For our synthetic generation, we consider six base skin tones that similarly resemble different amounts of melanin. We also add a small amount of random noise to the base color to add further variety. Overall, SynthDerm includes more than 2,600 images of size 64x64. We have made this dataset publicly available at https://affect.media.mit.edu/dissect/synthderm.

CelebA

We also include the CelebA dataset [146], where the attributes are nuanced and not truly independent. This dataset contains images of celebrities, is realistic, and closely resembles real-world settings. CelebA includes more than 200K celebrity images with 40 annotated face attributes, such as smiling, hair color, bangs, and glasses.

5.4.2 Evaluation Strategy

To evaluate the quality of the discovered CTs, we consider several measures that formalize *Importance* [230, 231], *Realism* [230], *Distinctness* [162], *Substitutability* [162, 192, 199, 209], and *Stability* [75, 162, 192], which commonly appear as desired qualities in the explainability literature.

Importance

Explanations should produce the desired outcome from the black-box classifier f. Previous work has referred to this quality using different names, such as importance [76], compatibility with classifier [230], and classification model consistency [231].

While most previous methods have relied on visual inspection, we introduce a quantitative metric to measure the gradual increase of the target class's posterior probability through a CT. Notably, we compute the correlation between α and $f(I(x, \alpha, k; f))$ introduced in Sec. 5.3. For brevity, we refer to $f(I(x, \alpha, k; f))$ as $f(\bar{x})$ in the remainder of the chapter. We also report the mean-squared error and the Kullback–Leibler divergence between α and $f(\bar{x})$.

We also calculate an empirical proxy for the generalization of black-box classifier f to counterfactual explanations. Specifically, we replace the test set of real images with their DISSECT-generated counterfactual explanations and quantify the performance of the pre-trained black-box classifier f on the DISSECT-generated test set. Better generalization to the counterfactual samples suggests that they are compatible with f and lie on the correct side of the classifier's boundary.

Realism

We need the generated samples that form a CT to look realistic to enable users/humans to identify concepts they represent. It means the counterfactual explanations should lie on the data manifold. This quality has been referred to as realism or data consistency [230].

Inspired by [48], we train a post hoc classifier that predicts if a given sample is real or generated. Although its objective is identical to that of the discriminator in our architecture, it is essential to do this step post hoc and independent from the training procedure. That is because relying on the discriminator's accuracy in an adversarial training framework can be misleading [48].

Distinctness

A desirable quality for explanations is to represent inputs with non-overlapping concepts, often referred to as diversity [162]. Others have suggested similar properties such as coherency, meaning examples of a concept should be similar to each other but different from examples of other concepts [76]. To quantify this quality in a way applicable to counterfactual generation, we introduce a distinctness metric. We also include metrics commonly used in disentanglement literature.

To measure distinctness, we report the performance of a secondary classifier that distinguishes between CTs. Mainly, we train a classifier post hoc that given a query image x and a generated image x' and K number of CTs, predicts one of the following: (1) x' is the reconstruction of x, (2) x' is a perturbation of x and CT_k has produced it, $k \in \{1, 2, \ldots, K\}$. This classifier is agnostic to our model and only uses its pair of input images.

Substitutability

The representation of a sample in terms of concepts should preserve relevant information [162, 192]. Previous work has formalized this quality for counterfactual generation contexts through a proxy called substitutability [209] . Substitutability measures an external classifier's performance on real data when it is trained using only synthetic images.[4] Higher performance when tested on real data suggests that explanations have

[4]This metric has been used in other contexts outside of explainability and has been called Classification Accuracy Score (CAS) [199] CAS is more broadly applicable than Frechet Inception Distance [100] and

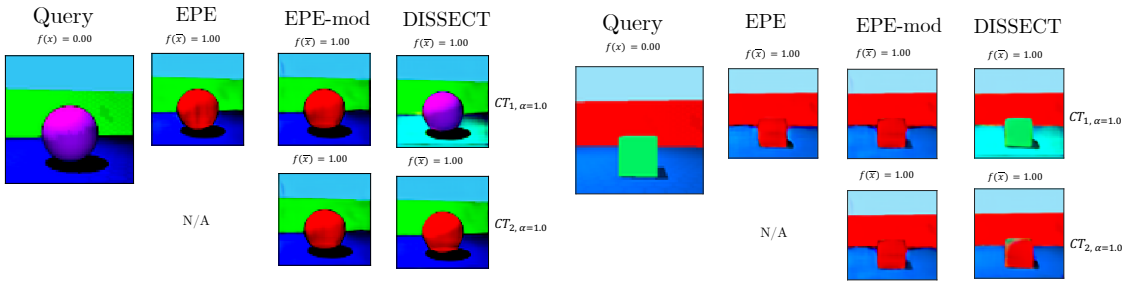


Figure 5-3: Qualitative results on `3D Shapes`. We observe that EPE and EPE-mod converge to finding the same single concept, despite EPE-mod having the ability to express multiple pathways to switch the classifier outcome from False to True. However, DISSECT is capable of discovering the two distinct ground-truth concepts: CT_1 flips the floor color to cyan and CT_2 flips the shape color to red.

retained relevant information and are of high quality.

Stability

Explanations should be coherent for similar inputs, a quality known as stability [75, 162, 192]. To quantify stability in the context of counterfactual explanations, we augment the test set by adding random noise to each sample x and produce several copies $\hat{x}_i$ where $i \in \{1, .., S\}$; Here, S is the number of random augmentations. Then, we generate counterfactual explanations $\bar{x}$ and $\bar{\hat{x}}_i$, respectively. We calculate the mean-squared difference between counterfactual images $\bar{x}$ and $\bar{\hat{x}}_i$ and the resulting Jensen Shannon distance between the predicted probabilities $f(\bar{x})$ and $f(\bar{\hat{x}}_i)$.

5.4.3 Case Study I: Validating the Qualities of Concept Traversals

Considering `3D Shapes` [19], we define an image as "colored correctly" if the shape hue is red or the floor hue is cyan. We train a classifier to detect whether a sample image is "colored correctly" or not. In this synthetic experiment, only these two independent factors contribute to the decision of this classifier.

Inception Score [208] that are only useful for evaluating GAN models. Furthermore, CAS can reveal information that none of these inception scores successfully capture [199].

Table 5.1: Quantitative results on `3D Shapes`. DISSECT performs significantly better or on par with the strongest baselines in each category of evaluation criteria. We observe that the modified variants of disentanglement VAEs perform poorly in terms of *Importance*, worse than CSVAE, and significantly worse than EPE, EPE-mod, and DISSECT. CSVAE, along with other VAE variants, cannot produce high-quality images, thus achieving poor *Realism* scores. On the other hand, EPE, EPE-mod, and DISSECT generate realistic samples indistinguishable from real images. While the aggregated metrics for *Importance* are useful for discarding VAE baselines with poor performance, they do not show a consistent order across EPE, EPE-mod, and DISSECT. Our approach greatly improves *Distinctness*, especially compared to EPE-mod. The EPE baseline is inherently incapable of doing this, and the extension EPE-mod does, but poorly. For contextualizing the *Substitutability* scores, note that the classifier's precision, recall, and accuracy when training on actual data is 100.0%. * Certain VAE methods fail to change the classification outcome. They only generate samples that produce $f(\bar{x}) = 0.0$. Correlation with a constant value is undefined.

	Importance							Realism			Distinctness					Substitutability			Stability	
	↑ R	↑ ρ	↓ KL	↓ MSE	↑ Gen Acc	↑ Gen Prec	↑ Gen Rec	↓ Acc	↓ Prec	↓ Rec	↑ Acc	↑ Prec (micro)	↑ Prec (macro)	↑ Rec (micro)	↑ Rec (macro)	↑ Acc Sub	↑ Prec Sub	↑ Rec Sub	↓ CF MSE	↓ Prob JSD
VAE-mod	0.1	0.3	inf	0.42	50.0	0	0	94.9	92.1	99.6	86.3	95.1	95.3	80.6	80.6	20.2	19.1	99.2	0.151	**0.0000**
β-VAE-mod	0.0	0.0	inf	0.42	50.0	0	0	99.5	99.0	100	90.7	92.2	92.2	91.0	91.0	33.0	10.5	33.4	0.197	**0.0000**
Annealed-VAE-mod	N/A*	N/A*	inf	0.42	50.0	0	0	100	100	100	34.0	53.2	49.2	1.20	1.2	46.4	15.9	42.5	0.175	**0.0000**
DIPVAE-mod	0.1	0.5	inf	0.41	52.3	100	4.6	100	100	100	97.2	96.7	96.9	96.5	96.5	18.2	18.4	96.2	0.127	0.0001
βTCVAE-mod	0.5	0.7	inf	0.42	50.0	0	0	100	100	100	**100**	**100**	**100**	**100**	**100**	20.2	19.1	98.8	0.143	**0.0000**
CSVAE	0.3	0.3	5.5	0.28	64.6	**100**	29.3	100	100	100	71.4	74.7	76.4	87.2	87.2	47.0	23.8	81.3	19.544	0.0274
EPE	0.8	**0.8**	**1.54**	0.09	98.4	**100**	96.7	50.1	**0**	**0**	-	-	-	-	-	99.2	95.9	**100**	0.134	0.0004
EPE-mod	**0.9**	0.7	2.2	**0.08**	**99.7**	**100**	**99.4**	**49.3**	**0**	**0**	45.3	49.6	49.8	30.3	30.3	91.0	**100**	52.5	0.128	0.0002
DISSECT	0.8	**0.8**	1.61	**0.08**	98.7	**100**	97.5	**49.3**	**0**	**0**	**100**	**100**	**100**	**100**	**100**	**100**	99.7	**100**	**0.102**	0.0003

Given a not "colored correctly" query, we would like to recover a CT related to the shape color and another CT associated with the floor color–two different pathways that lead to switching the classifier outcome for that sample.[5]

Table 5.1 summarizes the quantitative results on `3D Shapes`. Most VAE-variants perform poorly in terms of *Importance*, CSVAE performs slightly better, and EPE, EPE-mod, and DISSECT perform best. Our results suggest that DISSECT performs similarly to EPE that has been geared explicitly toward exhibiting *Importance* and its extension, EPE-mod. Additionally, DISSECT still keeps *Realism* intact. Also, it notably improves the *Distinctness* of CTs compared to relevant baselines.

Figure 5-3 shows the qualitative results for EPE, EPE-mod, and DISSECT.[6] Our results reveal that EPE converges to finding only one of these concepts. Similarly, both CTs generated by EPE-mod converge to finding the same concept, despite being given the capability to explore two pathways to switch the classifier outcome. However, DISSECT is capable of recovering the two separate concepts through its two generated CTs. Similarly, to flip the class of a "colored correctly" query to the opposite class, EPE-mod finds one possible transformation rather than two distinct ones that DISSECT finds. For brevity, only two sample queries are visualized; however, the observation is consistent across samples.

5.4.4 Case Study II: Investigating Alignment with Expert Domain Knowledge and Identifying Spurious Artifacts

A "high performance" model could learn to make its decisions based on irrelevant features that only happen to correlate with the desired outcome, known as label leakage [254]. One of the applications of DISSECT is to uncover such spurious concepts and allow probing a black-box classifier. Motivated by real-world examples that revealed classifier dependency

[5]In this scenario, these two ground-truth concepts do not directly apply to switching the classifier outcome from True to False.For example, if an image has a red shape and a cyan floor, both of these colors need to be changed to switch the classification outcome. We still observe that applying DISSECT to such cases results in CTs that change different combinations of colors. However, the baseline methods converge to the same CT. See appendix for more details.

[6]This figure only includes methods with minimally acceptable *Realism* scores.

on surgical markings in identifying melanoma [256], we design this experiment.

Given the synthetic nature of `SynthDerm` and how it has been designed based on real-world characteristics of melanoma [202, 242], each sample has a deterministic label of melanoma or benign. If the image is asymmetrical, has jagged borders, has different colors represented by salt-and-pepper noise, or has a large diameter (i.e., does not fit in a 40x40 square), the sample is melanoma. Otherwise, the image represents the benign class. Similar to in-situ dermatology images, melanoma samples have surgical markings more frequently than benign samples. We train a classifier to detect whether a sample image is melanoma or a benign lesion.

Given a benign query, we would like to produce counterfactual explanations that depict *how* to modify the input sample to change its class membership. We want DISSECT to recover CTs that disentangle meaningful characteristics of melanoma identification in terms of color, texture, or shape [202], and identify potential spurious artifacts that impact the classifier's predictions.

Table 5.2 summarizes the quantitative results on `SynthDerm`. Our method performs consistently well across all the metrics, significantly boosting *Distinctness* and *Substitutability* scores and making meaningful improvements on *Importance* scores. Our approach has higher performance compared to EPE-mod and EPE baselines and substantially improves upon CSVAE. Our method's high *Distinctness* and *Substitutability* scores show that DISSECT covers the landscape of potential concepts very well and retains the variety seen in real images strongly better than all the other baselines.

Figure 5-4 illustrates a few examples to showcase DISSECT's improvements over the strongest baseline, EPE-mod. We observe that EPE-mod converges to finding a single concept that only vaguely represents meaningful ground-truth concepts. However, DISSECT successfully finds concepts describing asymmetrical shapes, jagged borders, and uneven colors that align with ABCDE of melanoma [202]. DISSECT also identifies surgical markings as a concept that impacts the classifier's decisions. Overall, the qualitative results show that DISSECT uncovers several critical blind spots of the baseline techniques.

Table 5.2: Quantitative results on `SynthDerm`. The new DISSECT performs consistently best in terms of *Importance*, *Realism*, *Distinctness*, *Substitutability*, and *Stability*. Note that the precision, recall, and accuracy of the classifier when training on actual data is 97.685%, 100.0%, and 95.381%, respectively. Anchoring the *Substitutability* scores to these original values provides additional context, showing the meaningfully high performance of DISSECT compared to EPE-mod and EPE and a much larger improvement compared to CSVAE.

	Importance							Realism			Distinctness					Substitutability			Stability	
	↑ R	↑ ρ	↓ KL	↓ MSE	↑ Gen Acc	↑ Gen Prec	↑ Gen Rec	↓ Acc	↓ Prec	↓ Rec	↑ Acc	↑ Prec (micro)	↑ Prec (macro)	↑ Rec (micro)	↑ Rec (macro)	↑ Acc Sub	↑ Prec Sub	↑ Rec Sub	↓ CF MSE	↓ Prob JSD
CSVAE	0.25	0.64	1.78	0.12	86.7	43.9	5.2	54.6	69.9	16.3	85.1	0	0	0	0	29.9	36.8	55.4	2.318	0.006
EPE	0.87	0.23	inf	0.03	80.7	55.1	86.9	50.1	**0**	**0**	-	-	-	-	-	74.4	83.8	60.7	**0.111**	**0.001**
EPE-mod	0.81	0.73	0.92	0.04	95.3	83.9	79.5	**50.0**	**0**	**0**	85.1	71.1	26.3	0.7	0.7	74.3	83.5	60.7	0.239	0.002
DISSECT	**0.92**	**0.75**	**0.35**	**0.02**	**97.8**	**92.3**	**91.1**	**50.0**	**0**	**0**	**96.0**	**96.5**	**96.5**	**74.6**	**74.6**	**81.0**	**97.2**	**64.0**	0.338	**0.001**

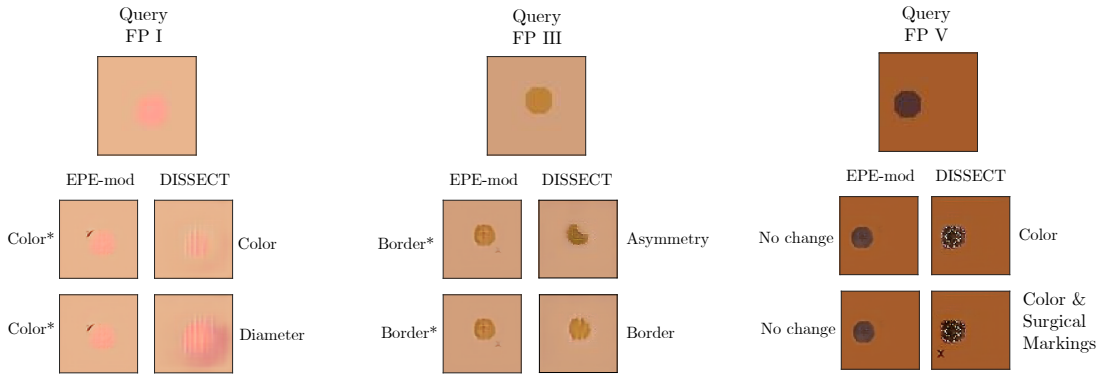


Figure 5-4: Qualitative results on `SynthDerm` comparing DISSECT with the strongest baseline, EPE-mod. We illustrate a few queries with different Fitzpatrick ratings [54] and visualize two of the most prominent concepts for each technique. We observe that EPE-mod converges to finding a single concept that only vaguely represents meaningful ground-truth concepts. However, DISSECT successfully finds concepts describing asymmetrical shapes, jagged borders, and uneven colors that align with the ABCDE of melanoma [202]. DISSECT also identifies concepts for surgical markings that impact the classifier's decisions. Basing melanoma classification on such spurious concepts is incongruent with expert domain knowledge. Successfully surfacing that the model has learned these false associations could inform actions to improve the model-under-test.

5.4.5 Case Study III: Identifying Biases

Another potential use case of DISSECT is to identify biases that might need to be rectified. Since our approach does not depend on predefined user concepts, it may help discover biases that were not identified beforehand. We design a simulated experiment to test DISSECT in such a setting. We sub-sample `CelebA` to create a training dataset such that smiling correlates with "blond hair" and "bangs" attributes. In particular, positive samples either have blond hair or have bangs, and negative examples are all dark-haired and do not have bangs. We use this dataset to train a purposefully biased classifier. We employ DISSECT to generate two CTs. Figure 5-5 shows the qualitative results, which depict that DISSECT discovers the two biases, which other techniques fail to do. Table 5.3 summarizes the quantitative results that replicate our finding from Table 5.1 in Sec. 5.4.3 and Table 5.2 in Sec. 5.4.4 in a real-world dataset, confirming that DISSECT outperforms

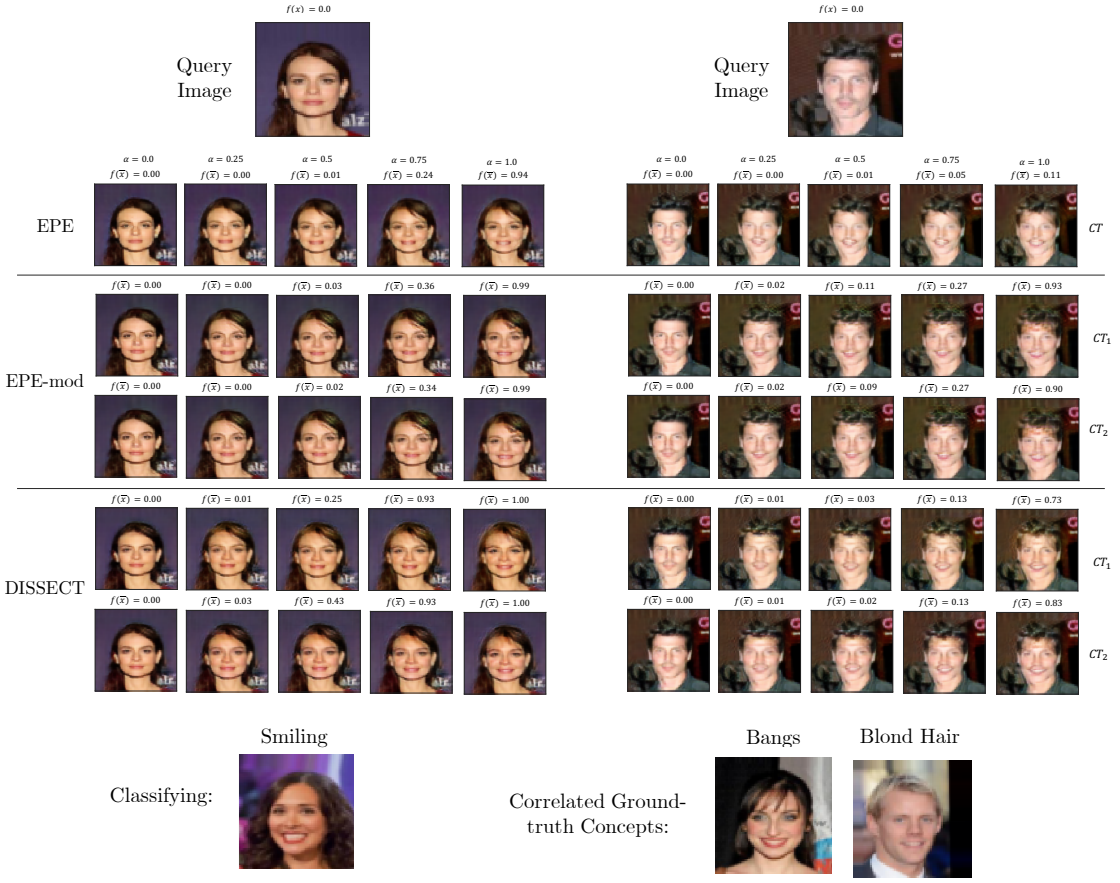


Figure 5-5: Qualitative results on `CelebA`. A biased classifier has been trained to predict smile probability, where the training dataset has been sub-sampled such that smiling co-occurs only with "bangs" and "blond hair" attributes. EPE does not support multiple CTs. We observe that EPE-mod converges to finding the same concept, despite having the ability to express various pathways to change $f(\bar{x})$ through CT_1 and CT_2. However, DISSECT discovers distinct pathways: CT_1 mainly changes hair color to blond, and CT_2 does not alter hair color but focuses more on hairstyle and tries to add bangs. Thus, DISSECT identifies two otherwise hidden biases.

all the other baselines in *Distinctness* without negatively impacting *Importance* or *Realism*.

Table 5.3: Quantitative results on `CelebA`. **Importance**: We observe that DISSECT performs similarly and even slightly outperforms the baselines in terms of *Importance* scores. **Realism**: DISSECT achieves a higher *Realism* score, suggesting disentangling CTs does not diminish the quality of generated images and may even improve them. **Distinctness**: DISSECT strongly improves the *Distinctness* of CTs compared to EPE-mod. The EPE baseline is inherently incapable of doing this, and the extension EPE-mod does, but poorly. For anchoring *Substitutability* scores, note that the classifier's precision, recall, and accuracy when training on actual data is 95.387%, 98.55%, and 92.662%, respectively.

	Importance							Realism			Distinctness					Substitutability			Stability	
	↑ R	↑ ρ	↓ KL	↓ MSE	↑ Gen Acc	↑ Gen Prec	↑ Gen Rec	↓ Acc	↓ Prec	↓ Rec	↑ Acc	↑ Prec (micro)	↑ Prec (macro)	↑ Rec (micro)	↑ Rec (macro)	↑ Acc Sub	↑ Prec Sub	↑ Rec Sub	↓ CF MSE	↓ Prob JSD
CSVAE	0.00	0.00	1.25	0.284	50.2	50.2	34.1	99.7	100	99.5	15.1	0.0	0.0	0.0	0.0	52.8	53.0	**98.6**	23.722	0.039
EPE	0.85	**0.91**	0.28	0.060	99.2	**99.9**	98.5	49.9	32.0	0.3	-	-	-	-	-	**93.0**	95.4	91.2	**0.411**	**0.003**
EPE-mod	**0.86**	0.90	0.21	0.048	**99.5**	99.7	**99.2**	49.3	33.3	0.1	18.7	50.0	50.1	15.2	15.2	91.8	94.8	89.4	0.446	0.004
DISSECT	0.84	0.88	**0.19**	**0.047**	99.2	99.8	98.5	**49.2**	**0.0**	**0.0**	**95.0**	**98.0**	**98.1**	**96.1**	**96.1**	91.9	**96.9**	87.6	0.567	0.005

5.5 Supplementary Materials

5.5.1 DISSECT Details

See Figure 5-6 for a detailed visualization of the components of our method or Figure 5-7 for a simplified version.

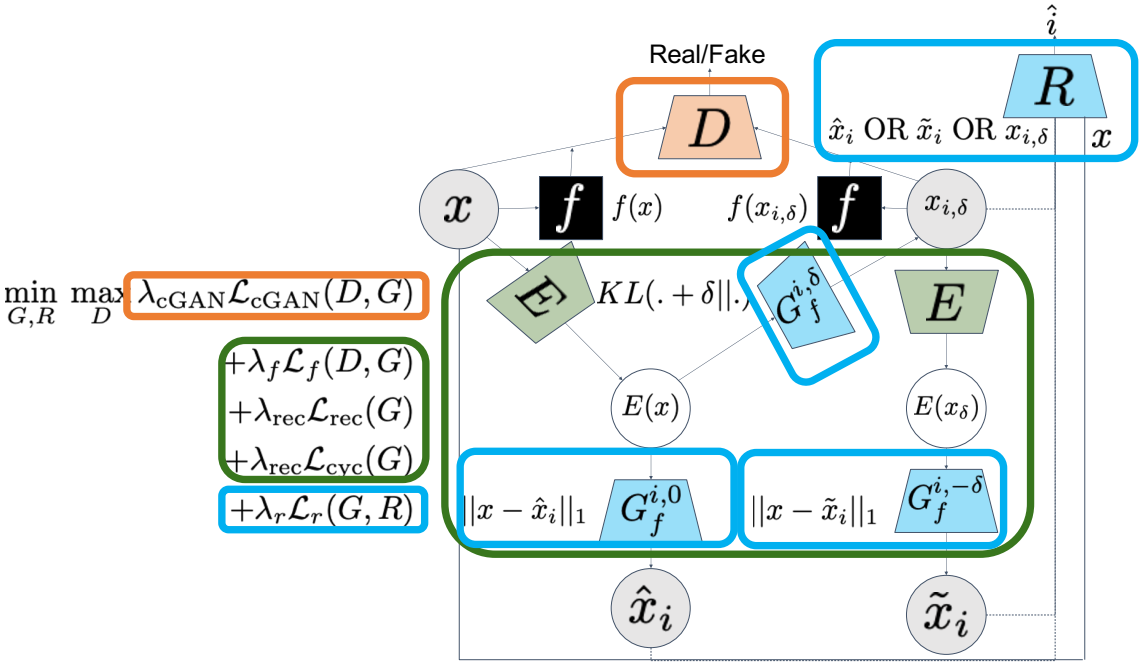


Figure 5-6: Illustration of DISSECT. Orange, Green, and Blue show elements related to the discriminator, generator, and CT disentangler, respectively.

5.5.2 Development of Modified VAE Baselines

To promote discovering *Important* CTs, we introduced $\mathcal{L}_{\text{aux}} = \frac{\sum_{d=1}^{K} \partial f(x)/\partial z_d}{K}$, which incorporated the directional derivative of f with respect to the latent dimensions of interest into the loss function of VAE. Despite experimentation with many variants of $\mathcal{L}_{\text{aux}}$, we observed two common themes.

First, a monotonic increase of $f(\bar{x})$ through traversing one latent dimension and keeping the rest static was hardly achieved. Second, while the purpose of $\mathcal{L}_{\text{aux}}$ was to

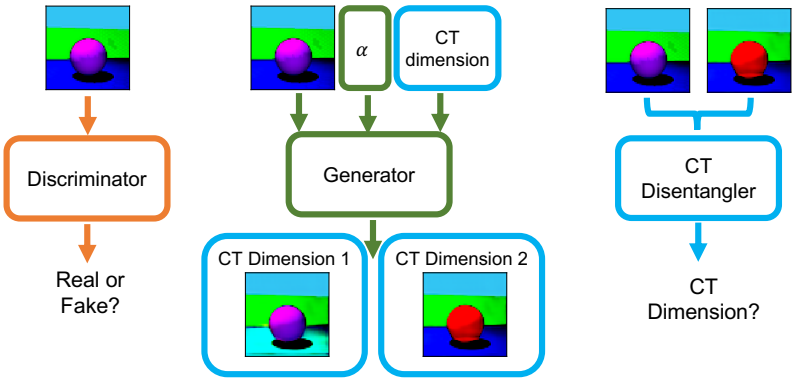


Figure 5-7: Simplified illustration of DISSECT. Orange, Green, and Blue show elements related to the discriminator, generator, and CT disentangler, respectively.

promote exerting *Importance* only in the first K dimensions of the latent space, $\partial f(x)/\partial z_d$ for $d \in \{K+1, K+2, \cdots, M\}$ were impacted similarly. Having strongly correlated dimensions is a failure in achieving the very goal of disentanglement approaches. Table 5.4 summarizes a subset of the variants of $\mathcal{L}_{\text{aux}}$ studied.

Table 5.4: Summary of a subset of $\mathcal{L}_{\text{aux}}$ iterations. The development goal is to make the first K dimensions of the latent space *Important*. In some iterations, we encouraged the remaining $M-K$ dimensions not to be *Important* to reduce potential correlation across latent dimensions.

	$\mathcal{L}_{\text{aux}}$
1	$\frac{\sum_{k=1}^{K} \partial f(x)/\partial z_k}{K}$
2	$\frac{\sum_{k=1}^{K} \partial f(x)/\partial z_k}{K} + \frac{\sum_{d=K+1}^{M} \lvert\partial f(x)/\partial z_d\rvert}{M-K}$
3	$\frac{\sum_{k=1}^{K} \partial f(x)/\partial z_k}{K} + \frac{\sum_{d=K+1}^{M} [\partial f(x)/\partial z_d]^2}{M-K}$
4	$\frac{\sum_{d=K+1}^{M} \lvert\partial f(x)/\partial z_d\rvert}{M-K}$
5	$\frac{\sum_{d=K+1}^{M} [\partial f(x)/\partial z_d]^2}{M-K}$
6	$\frac{\sum_{k,d} \lvert\partial f(x)/\partial z_d\rvert / \lvert\partial f(x)/\partial z_k\rvert}{K*(M-K)}$ where $k \in \{1, 2, \cdots, K\}, d \in \{K+1, K+2, \cdots, M\}$
7	$\frac{\sum_{k,d} log(\lvert\partial f(x)/\partial z_d\rvert / \lvert\partial f(x)/\partial z_k\rvert)}{K*(M-K)}$ where $k \in \{1, 2, \cdots, K\}, d \in \{K+1, K+2, \cdots, M\}$

5.5.3 Evaluation Metrics Details

Following [230], we conduct a more granular analysis to investigate if DISSECT works similarly across different queries, e.g., when flipping classification outcome from "True" to "False" or the other way around. We plot α vs. $f(\bar{x})$ for query samples with $f(x) < 0.5$ and $f(x) \geq 0.5$ separately.

The VAE-based baselines support continuous values for latent dimensions z_k. Also, we can directly sample latent code values and produce CT_k by keeping z_j ($j \neq k$) constant and monotonically increasing z_k values. However, to calculate the evaluation metrics comparably to EPE, EPE-mod, and DISSECT, we do the following: We encode each query sample using the probabilistic encoder. We set $z_j = \mu_j$, $j \neq k$ where μ_j is the mean of the fitted Gaussian distribution for z_j. For dimension k, we produce $N + 1$ linearly spaced values between $\mu_p \pm 2 * \sigma_p$, where μ_p and σ_p are the mean and standard deviation of the prior normal distribution, in our case 0.0 and 1.0 respectively. Note that these different values for z_k map out to α, $\alpha \in \{0, \frac{1}{N}, \cdots, 1\}$ in EPE, EPE-mod, and DISSECT models. After this step, calculating all the metrics related to *Importance*, *Realism*, and *Distinctness* is identical across all the models.

5.5.4 Experiment Setup and Hyper-parameter Tuning Details

We seeded the model's parameters from [230] based on the reported values in their accompanying open-sourced repository[7]. We used the same parameters for `3D Shapes`, except for the number of bins, N, used for ordinal regression transformation of the classifier's posterior probability. The largest number of bins that resulted in non-zero samples per bin, 3, was selected. We kept all the parameters shared between EPE, EPE-mod, and DISSECT the same.

Given the experiments' design, we fixed the number of dimensions K in DISSECT and EPE-mod to 2. We experimented with a few values for λ_r, 1, 10, 20, 50. Based on manual

[7] `https://github.com/batmanlab/Explanation_by_Progressive_Exaggeration`

inspection after 30k training batches, λ_r was selected. Factors considered for selection included inspecting the perceived quality of generated samples and the learning curves of $\mathcal{L}_{\text{cGAN}}(D)$, $\mathcal{L}_{\text{cGAN}}(G)$, $\mathcal{L}_{\text{cyc}}(G)$, $\mathcal{L}_{\text{rec}}(G)$, and $\mathcal{L}_r(G, R)$.

For evaluation, we used a hold-out set including 10K samples. For post hoc evaluation classifiers predicting *Distinctness* and *Realism*, 75% of the samples were used for training, and the results were reported on the remaining 25%. See Table 5.5 for the summary of the hyper-parameter values.

Table 5.5: Summary of hyper-parameter values. Discriminator optimization happens once every D steps. Similarly, generator optimization happens once every G steps. λ_r is specific to DISSECT, and K is specific to EPE-mod and DISSECT. All the remaining parameters are shared across EPE, EPE-mod, and DISSECT. Note that samples used for evaluation are not included in the training process.

	Preprocessing		Training									Evaluation Metrics			
	N	max samples per bin	λ_{cGAN}	λ_{rec}	λ_f	D steps	G steps	batch size	epochs	K	λ_r	max # samples	batch size	epochs	hold-out test ratio
`3D Shapes`	3	5,000	1	100	1	1	5	32	300	2	10	10,000	32	10	0.25
`SynthDerm`	2	1,350	2	100	1	5	1	32	300	5	2	10,000	8	10	0.25
`CelebA`	10	5,000	1	100	1	1	5	32	300	2	10	10,000	32	10	0.25

5.5.5 Additional Qualitative Results for Case Study I

Recall that considering `3D Shapes`, we define an image as "colored correctly" if the shape hue is red *or* the floor hue is cyan. Given a not "colored correctly" query, we recover a CT related to the shape color and another CT associated with the floor color–two different pathways leading to switching the classifier outcome for that sample. See Figure 5-8 for additional qualitative examples where classification outcome is flipped from False to True.

However, these two ground-truth concepts do not directly apply to switching the classifier outcome from True to False in this scenario. For example, if an image has a red shape *and* a cyan floor, both colors need to be changed to switch the classification outcome. As shown in Figure 5-9, we still observe that applying DISSECT to such cases results in two discovered CTs that change different combinations of colors while EPE-mod converges to the same CT.

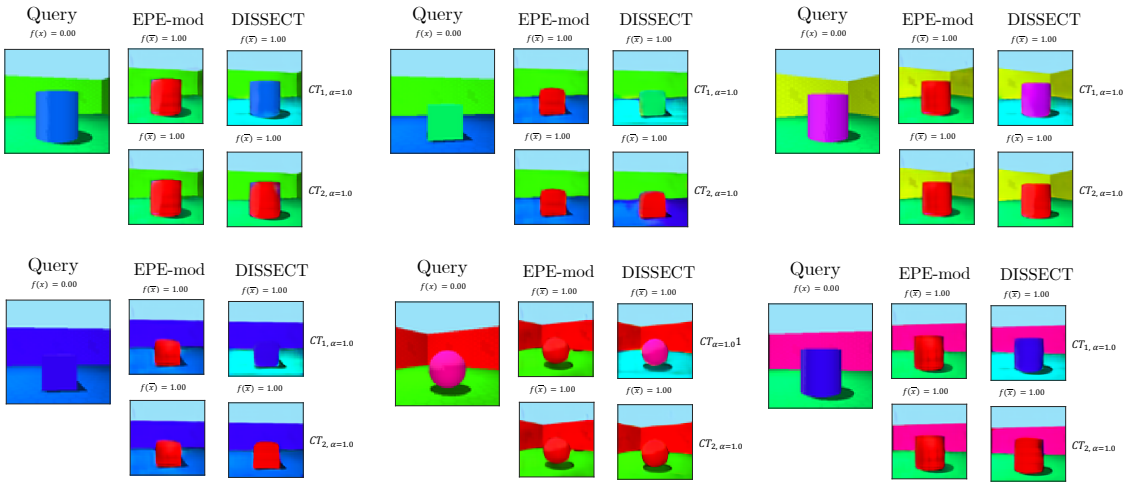


Figure 5-8: Qualitative results on `3D Shapes` when flipping classification outcome from "False" to "True." We observe that EPE-mod converges to finding the same concept, despite having the ability to express multiple pathways to switch the classifier outcome. However, DISSECT can discover the two Distinct ground-truth concepts: CT_1 flips the floor color to cyan, and CT_2 converts the shape color to red.

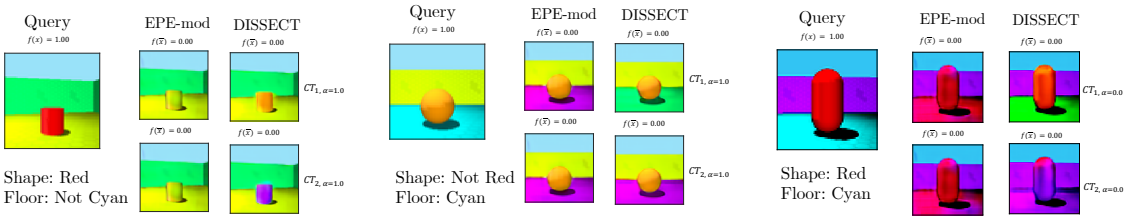


Figure 5-9: Qualitative results on `3D Shapes` when flipping classification outcome from "True" to "False." We observe that EPE-mod converges to finding the same concept, despite having the ability to express multiple pathways to switch the classifier outcome. However, DISSECT is capable of discovering Distinct paths to do so. **Left**: When the input query has a red shape, but the floor color is not cyan, CT_1 flips the shape color to orange and CT_2 flips it to violet. **Middle**: When the input query has a cyan floor, but the shape color is not red, CT_1 flips the floor color to lime, and CT_2 converts it to magenta. **Right**: When the input query has a red shape and cyan floor, CT_1 changes the shape color to dark orange and floor color to lime, and CT_2 flips the shape color to violet and floor color to magenta.

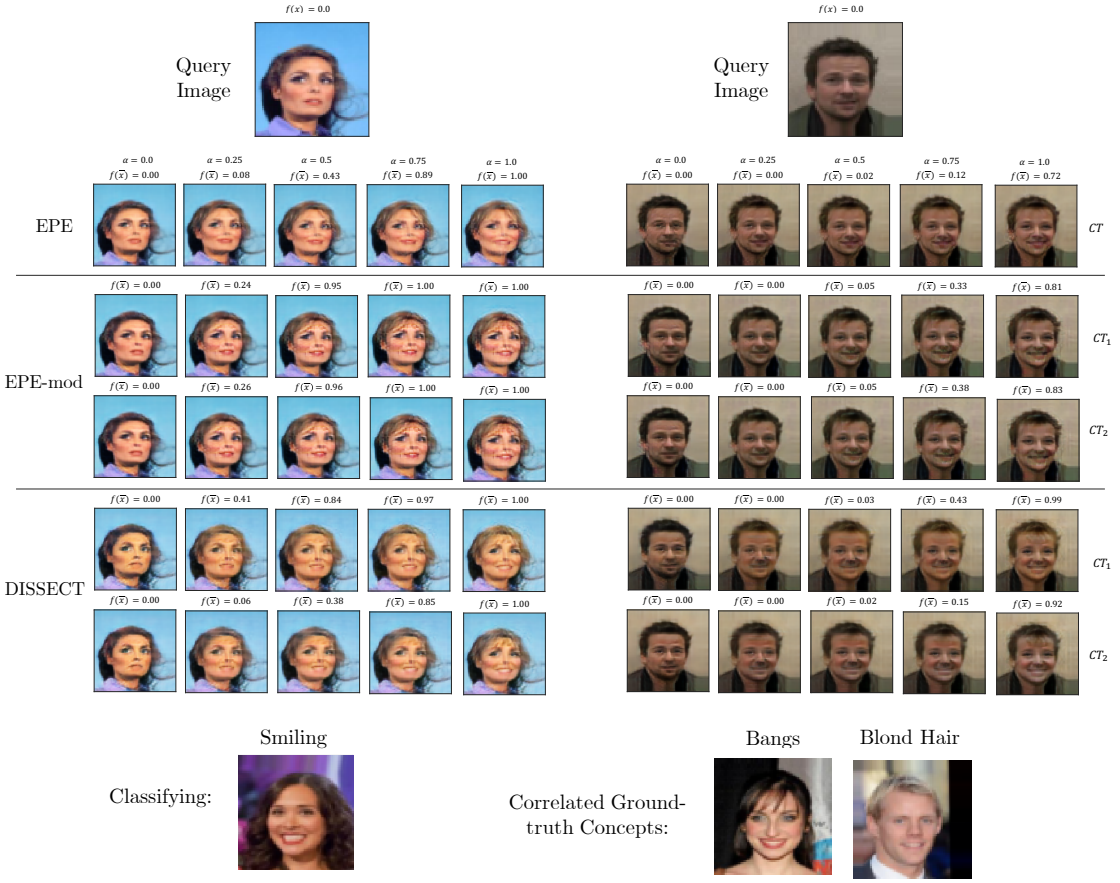


Figure 5-10: Qualitative results on `CelebA`. A biased classifier has been trained to predict smile probability, where the training dataset has been sub-sampled such that smiling co-occurs only with "bangs" and "blond hair" attributes. EPE does not support multiple CTs. We observe that EPE-mod converges to finding the same concept, despite having the ability to express several pathways to change $f(\bar{x})$ through CT_1 and CT_2. However, DISSECT can discover Distinct routes: CT_1 mainly changes hair color to blond, and CT_2 does not alter hair color but focuses more on hairstyle and tries to add bangs. Thus it identifies two otherwise hidden biases.

5.5.6 Additional Quantitative Results for Case Study I

Figure 5-11 depicts more details regarding CTs' *Importance* across different groups of samples for 3D Shapes experiments.

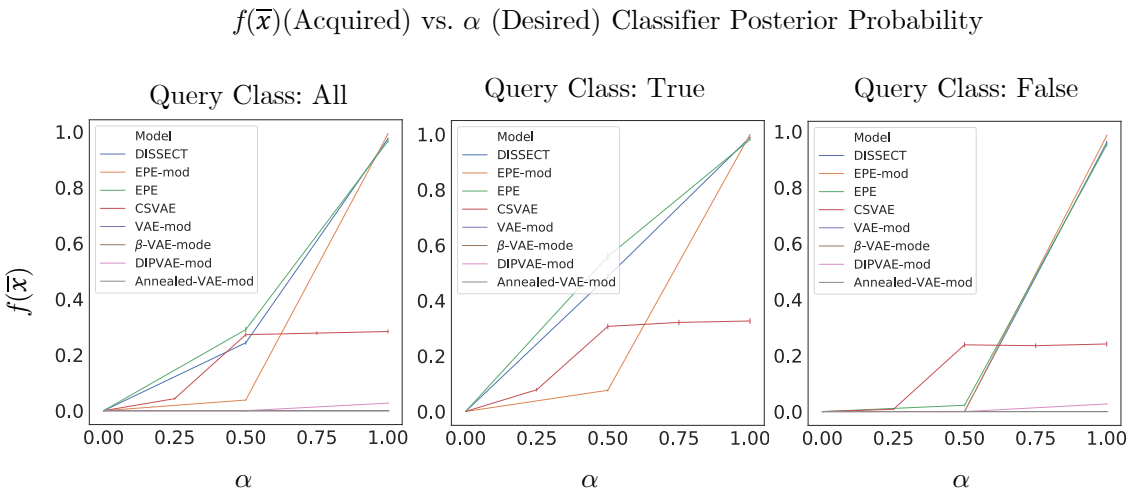


Figure 5-11: Acquired vs. desired classifier posterior probability for generated samples that constitute a CT on `3D Shapes` over 10K queries in total. The ideal would be a line of slope one. Error bars represent 95% confidence intervals. We observe that DISSECT performs similarly to EPE that has been particularly geared toward exhibiting *Influence*, and its extension, EPE-mod. VAE based methods perform poorly in terms of *Influence*. CSVAE performs significantly better than other VAE baselines but still works much worse than EPE, EPE-mod, and DISSECT. There is a significant correlation between acquired and desired posterior probabilities of generated samples for DISSECT (r=0.82, p<.0001), EPE-mod (r=0.87, p<.0001), EPE (r=0.81, p<.0001), and CSVAE (r=0.32, p<.0001). In other VAE baselines, there is very low or no correlation between acquired and desired probabilities: DIPVAE (r=0.14, p<.0001), VAE (r=0.07, p<.0001), β-VAE-mode (r=-0.01, p>.1) and Annealed-VAE-mod (r=-0.01, p>.1).

5.5.7 Additional Quantitative Results for Case Study II

Figure 5-12 provides more granular information about *Importance* scores that further confirms our qualitative results for `SynthDerm` experiments.

5.5.8 Additional Quantitative Results for Case Study III

Figure 5-13 provides further details regarding *Importance* scores on a more granular scale for celebA dataset.

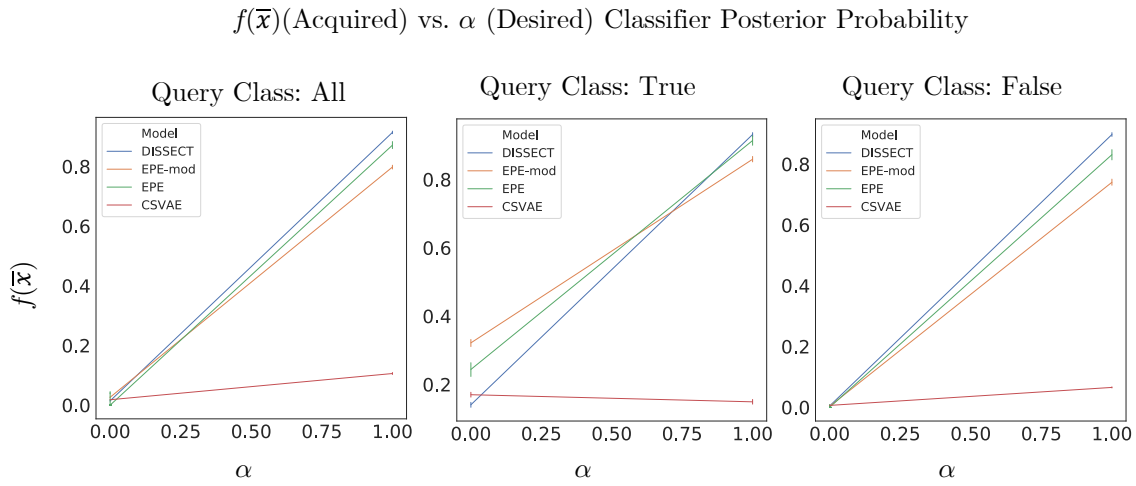


Figure 5-12: Acquired vs. desired classifier posterior probability for generated samples that constitute a CT on `SynthDerm` over 10K queries in total. The ideal would be a line of slope one. Error bars represent 95% confidence intervals. We observe that DISSECT performs similarly to EPE that has been particularly geared toward exhibiting *Influence*, and it potentially outperforms EPE-mod. Although CSVAE produces examples with acquired posterior probabilities correlated with the desired values (r=0.25, p<.0001), it performs significantly worse than EPE (r=0.87, p<.0001), EPE-mod (r=0.81, p<.0001), and DISSECT (r=0.92, p<.0001).

5.5.9 Additional Qualitative Results for Case Study III

Recall the biased `CelebA` experiment where smiling correlates with "blond hair" and "bangs" attributes. Figure 5-10 shows additional qualitative samples, suggesting that DISSECT can recover and separate the aforementioned concepts, which other techniques fail to do.

5.6 Conclusions

Unlike previous work on generative explainability that cannot disentangle important concepts effectively, produce poor quality or unrealistic examples, or fail to retain relevant information, we proposed a novel approach, DISSECT, to overcome such challenges using little supervision. We hypothesized that DISSECT could successfully find multiple distinct

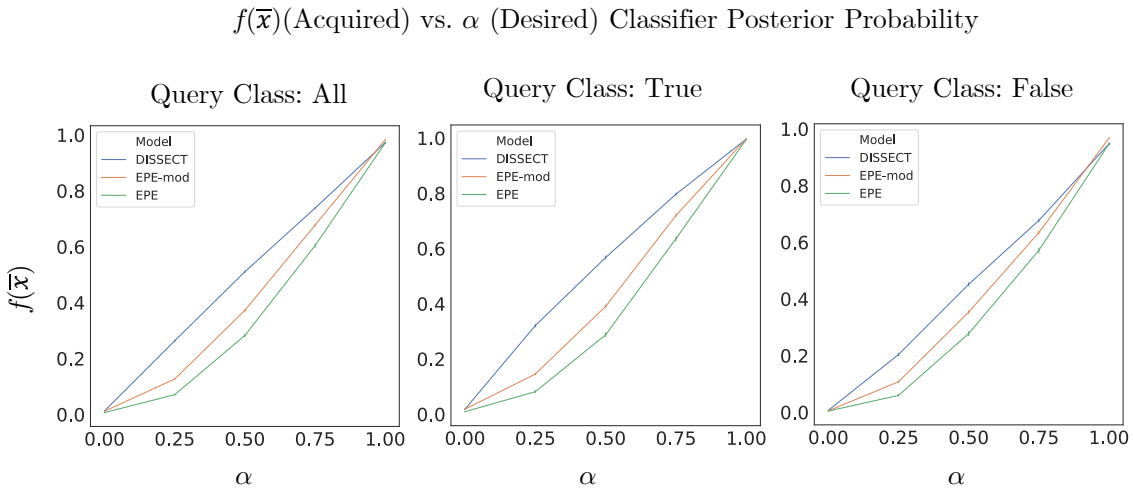


Figure 5-13: Acquired vs. desired classifier posterior probability for generated samples that constitute a CT on `CelebA` over 10K queries in total. The ideal would be a line of slope one. Error bars represent 95% confidence intervals. The results suggest that DISSECT performs on par with the three strongest baselines in terms of *Importance*. Acquired and desired probabilities of generated samples are significantly correlated for DISSECT (r=0.84, p<.0001), EPE-mod (r=0.86, p<.0001), and EPE (r=0.85, p<.0001).

concepts by generating a series of realistic-looking, counterfactually generated samples that gradually traverse a classifier's decision boundary. Our hypothesis is supported by experimental results, both quantitatively and qualitatively. This method could provide additional checks and balances for practitioners to probe how their model works before deployment, especially in high-stakes tasks such as in medical decision making. DISSECT helps identify how well the model-under-test reflects practitioners' domain knowledge and whether the model exhibits biases that might need to be rectified. Since this method does not depend on predefined user concepts, it may help discover biases that were not identified beforehand.

One avenue for future work is extending earlier theories [147, 227] to obtain theoretical guarantees for the proposed approach. While we provide an extensive list of qualitative examples in the appendix, further confirmation from human-subject studies to validate that CTs exhibit semantically meaningful attributes could strengthen our findings.

We emphasize that our proposed method does not guarantee to find all the biases of a classifier, nor does it ensure semantic meaningfulness across all found concepts. Additionally, it should not replace any current procedures for promoting accountability in model evaluation such as model cards [167]; instead, it should be used in tandem with them. Model fairness, robustness, and accountability require in-depth discussion across multiple fields, from computer science to sociology to law to psychology; we hope the new contributions in this chapter lead to better provide insights accessible to all people, not just computer and AI scientists.

5.7 Statement of Contributions

Inspired by Been Kim's work on machine learning and interpretability, I initiated this project. I started early versions of this work during an internship at Google, where Brian Eoff and Brendan Jou were my host and co-host, respectively. They provided guidance and support throughout the project with code reviews and discussions about the experiments and preliminary results. Been's role has been indispensable. She helped shape this work's framing, positioning it with respect to related work, and continued advisement on implementation and experiments. Chun-Liang Li shared his expertise in generative models that informed the modeling choices and loss function design. After returning to MIT, I continued working in this area with a new approach and conducted the analyses and implementation. I also reimplemented the methods I investigated during my time at Google as baselines for comparison. Been, Brendan, Brian, Chun-Liang, and Roz continued their advisement and guidance throughout.

Chapter 6

Conclusions and Future Work

This thesis has introduced a conceptual framework for human-machine collaboration composed of two components: interpretation *of people* by machines and interpretation of machines *by people*. Throughout this thesis, I have presented several novel tools that have made improvements across these axes. Based on the insights drawn from extended experiments, I have argued that improvements across both axes can bring us closer to human-centered (HC) optimality.

To provide a more comprehensive inference of the human state, Chapter 2 started by addressing several challenges encountered in depressive symptom estimation in outpatient clinical settings. Most of the previously proposed machine learning solutions used data gathered through constrained conditions, used data requiring active patient input daily, and rarely compared their performance against clinically validated mental health measurements. However, Chapter 2 presented a pipeline that achieved less than 8% error rate in predicting Hamilton Depression Rating Scale (HDRS) scores solely from phone and wearable sensor data. Collecting measurable real-world behavioral and physiological data allows for more scalable, accurate, and less burdensome symptom tracking and can overcome the limitations of current office-based clinical interviews and self-reports in diagnosing and treating major depressive disorder.

Making accurate inferences without increasing the human burden by requiring them to

explicitly self-report their evaluations is not unique to measuring depressive symptoms. Chapter 3 approached this challenge from another angle, text-based interactions. In the open-domain dialog, automated evaluation metrics are poorly correlated with human judgments of quality, if correlated at all. Chapter 3 showed that considering the conversation trajectory and implicit signals that capture sentiment, semantics, and user engagement that are psychologically motivated can be a powerful solution. Combined with a self-play scenario where the dialog system talks to itself and calculates our proposed novel metric based on the aforementioned qualities, an automated metric is born that significantly alleviates this gap between automated and human-rated evaluation.

For a more comprehensive interpretation of machines by people, Chapter 4 characterized the interpretability benefits of uncertainty quantification in a multi-annotator setting. Chapter 4 showed that applying Monte Carlo dropout to a classical network provides uncertainty measures and helps disambiguate annotator and data bias, inter-rater disagreement, and provides a proxy for model calibration. This disambiguation informs actionable directions for improvements, a prioritization that otherwise would not be available.

For richer explanations that allow people to ask what-if questions, Chapter 5 presented a novel explanation that jointly trains a generator, a discriminator, and a concept traversal disentangler. Compared to various strong baselines, the proposed method simultaneously satisfies several desirable qualities for interpretability across multiple realistic and synthetic datasets. Results from multiple simulated experiments confirmed that our novel method could help people reach more profound insights about the model by revealing potential biases of a classifier, investigating its alignment with expert domain knowledge, and identifying spurious artifacts that impact predictions.

6.1 Contributions

This thesis makes the following contributions that improve interpretation *of people* by machines and interpretation of machines *by people* and bring us closer towards HC opti-

mality:

- A novel pipeline using machine Learning techniques to estimate depressive symptoms from phone and wearable sensor data with low error rate;
- New data-driven insights that identify behavioral and physiological features most informative for depressive symptom estimation;
- A novel, model-agnostic, and dataset-agnostic method using self-play to evaluate open-domain dialog that approximates human evaluation more strongly than other automated metrics known today;
- New insights about Monte Carlo dropout uncertainty estimation in deep learning settings such as proxies for inter-rater disagreement, model calibration, and dataset bias;
- A novel counterfactual explanation model to translate the decision boundary of a trained model into human-understandable concepts, producing a trajectory of examples for each concept, and showing that our method outperforms other explanation methods on five HC criteria.
- Novel tools and insights for identifying potential biases of a classifier, investigating its alignment with expert domain knowledge, identifying spurious artifacts that impact predictions, and informing actionable directions for model improvement.

The above contributions have led to several peer-reviewed publications directly [62, 63, 66, 67, 71, 74] and make up major parts of several other [69, 70, 72, 73, 104, 105, 112, 113, 116, 135, 163, 186, 187, 206, 207, 211, 263]. In addition, the code supporting these projects has been open-sourced and is available at `https://github.com/asmadotgh`.

6.2 Future Work

This thesis provided an array of technical solutions in the intersection of machine learning and human-computer interaction to improve the interpretation *of people* by machines and the interpretation of machines *by people*. However, this is only part of the solution to the HC optimality puzzle. While I focus on a few application areas, a broader definition of HC optimality could be under-specified. Many variables influence the problem definition itself, let alone its solution. Articulating HC optimality criteria in a well-defined manner in different application domains requires in-depth discussion across multiple fields, from computer science to sociology to law to psychology.

While this thesis laid out the groundwork for improving the interpretations of people by machines and interpretations of machines by people, there are several ways the techniques provided can be improved or applied to other application domains. Additionally, the insights drawn from this thesis gave rise to new questions for future exploration. For example, one area of improvement is improving the accuracy of depressive symptom estimation. In this thesis, we hand-crafted daily features and assumed no time dependency between data points. One question is to what degree other ways of framing this question might help predictive power, such as using few-shot learning approaches and treating each individual as a new class.

Another area of future work is studying what characteristics naturally arise in high-quality generated dialogs using an inverse reinforcement learning approach. Can we extract the same qualities that we identified as necessary in successful human-human conversations, such as sentiment, semantics, and user engagement? Can we disentangle them successfully in the latent inferred representation?

One of the areas for future work is investigating to what degree the insights drawn from Monte Carlo dropout uncertainty estimation are transferrable to other uncertainty estimation methods in deep learning, such as ensemble methods [134] and stochastic variational Bayesian inference methods such as Bayes by Backprop [10]. Additional

studies to confirm our findings' robustness concerning hyperparameters such as dropout probability can further strengthen the utility of our takeaway messages. Another question for future work is to what degree these techniques can lead to insights about domains and high-level concepts instead of single input images? Grouping samples and providing statistics of such uncertainty measurements over a group instead of each input image can be a potential strategy to address this question.

To deepen our understanding of our proposed techniques for interpreting machines by people, further evaluation through human-subject experiments is another venue for exploration. Human-subject experiments can complement the findings from our simulation experiments by providing ratings of generated concepts and assigning semantically meaningful names to them. Additionally, experiments to test if crowd-workers achieve a learning task more accurately or quickly can further ground this work in the application. Such experiments can help investigate the contingency of the results on the problem setup's choice to confirm its effectiveness across problem settings and close the loop on one of this method's potential applications in a realistic educational setting. Additionally, extending this technique to domains beyond visual data such as time series or text-based input can broaden this work's impact.

Bibliography

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

[3] Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. *arXiv preprint arXiv:2003.09461*, 2020.

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[5] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283. ACM, 2016.

[6] Antoine Bechara, Hanna Damasio, and Antonio R Damasio. Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, 10(3):295–307, 2000.

[7] Sarah R Beck, Kevin J Riggs, and Sarah L Gorniak. Relating developments in children's counterfactual thinking and executive functions. *Thinking & reasoning*, 15(4):337–354, 2009.

[8] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[9] Isabelle Blanchette and Anne Richards. The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning. *Cognition & Emotion*, 24(4):561–595, 2010.

[10] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *ICML*, 2015.

[11] Graham D Bodie, Kellie St. Cyr, Michelle Pence, Michael Rold, and James Honeycutt. Listening competence in initial interactions i: Distinguishing between what listening is and what listeners do. *International Journal of Listening*, 26(1):1–28, 2012.

[12] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 477–486, 2014.

[13] Andrey Bogomolov, Bruno Lepri, and Fabio Pianesi. Happiness recognition from mobile phone data. In *2013 international conference on social computing*, pages 790–795. IEEE, 2013.

[14] Marko Borazio, Eugen Berlin, Nagihan Kücükyildiz, Philipp Scholl, and Kristof Van Laerhoven. Towards benchmarked sleep detection with wrist-worn sensing units. In *Healthcare Informatics (ICHI), 2014 IEEE International Conference on*, pages 125–134. IEEE, 2014.

[15] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[16] Joan E Broderick and Gregory Vikingstad. Frequent assessment of negative symptoms does not induce depressed mood. *Journal of clinical psychology in medical settings*, 15(4):296–300, 2008.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[18] Daphna Buchsbaum, Sophie Bridgers, Deena Skolnick Weisberg, and Alison Gopnik. The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2202–2212, 2012.

[19] Chris Burgess and Hyunjik Kim. 3D shapes dataset, 2018.

[20] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-VAE. *Advances in neural information processing systems: Workshop on Learning Disentangled Representations*, 2017.

[21] Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushkova, Scott Reed, Rae Jeong, Konrad Zolna, Yusuf Aytar, David Budden, Mel Vecerik, et al. Scaling data-driven robotics with reward sketching and batch reinforcement learning. *Robotics: Science and Systems Conference 2020*, 2020.

[22] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[23] Luca Canzian and Mirco Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1293–1304, 2015.

[24] Ginevra Castellano, Loic Kessous, and George Caridakis. Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in human-computer interaction*, pages 92–103. Springer, 2008.

[25] Runjin Chen, Hao Chen, Jie Ren, Ge Huang, and Quanshi Zhang. Explaining neural networks semantically and quantitatively. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9187–9196, 2019.

[26] Weixuan Chen, Javier Hernandez, and Rosalind W Picard. Estimating carotid pulse and breathing rate from near-infrared video of the neck. *Physiological measurement*, 39(10):10NT01, 2018.

[27] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[28] W Clerx, A Phillips, S Lockley, C O'Brien, E Klerman, and C Czeisler. Impact of irregularity of sleep-wake schedules on circadian phase and amplitude in college undergraduates. In *The 2014 Meeting of the Society for Research on Biological Rhythms*, 2014.

[29] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, 2017.

[30] Leda Cosmides and John Tooby. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *cognition*, 58(1):1–73, 1996.

[31] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. Assessing and addressing algorithmic bias in practice. *Interactions*, 25(6):58–63, 2018.

[32] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics, 2011.

[33] Mark A Davis. Understanding the relationship between mood and creativity: A meta-analysis. *Organizational behavior and human decision processes*, 108(1):25–38, 2009.

[34] Kalyanmoy Deb. Multi-objective optimization. In *Search methodologies*, pages 403–449. Springer, 2014.

[35] Orianna DeMasi, Konrad Kording, and Benjamin Recht. Meaningless comparisons lead to false optimism in medical machine learning. *PloS one*, 12(9):e0184604, 2017.

[36] Orianna DeMasi and Benjamin Recht. A step towards quantifying when an algorithm can and cannot predict an individual's wellbeing. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 763–771, 2017.

[37] John S Denker and Yann Lecun. Transforming neural-net output levels to probability distributions. In *Advances in neural information processing systems*, pages 853–859, 1991.

[38] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation, 2019.

[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[40] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in neural information processing systems*, pages 592–603, 2018.

[41] Hamdi Dibeklioğlu, Zakia Hammal, and Jeffrey F Cohn. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics*, 22(2):525–536, 2017.

[42] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*, 2019.

[43] Sidney D'Mello, Rosalind W Picard, and Arthur Graesser. Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, 22(4):53–61, 2007.

[44] Raymond J Dolan. Emotion, cognition, and behavior. *science*, 298(5596):1191–1194, 2002.

[45] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[46] Michael Downs, Jonathan L Chu, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Pan. Cruds: Counterfactual recourse using disentangled subspaces. *ICML Workshop on Human Interpretability in Machine Learning*, 2020.

[47] Sylvie Droit-Volet and Warren H Meck. How emotions colour our perception of time. *Trends in cognitive sciences*, 11(12):504–513, 2007.

[48] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Conference on Empirical Methods in Natural Language Processing*, pages 11–21, 2018.

[49] Empatica. Empatica E4. `https://store.empatica.com/products/e4-wristband`, 2012. Online; accessed May'17.

[50] E Epstein. Verification of forecasts expressed in terms of probability. *J. Appl. Meteorol*, 8(6):985–987, 1969.

[51] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[52] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *2017 Conference on Empirical Methods in Natural Language ProcessingConference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017.

[53] Angelos Filos, Panagiotis Tigkas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2020.

[54] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

[55] Joseph L Fleiss, Jacob Cohen, and Brian S Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323, 1969.

[56] Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. Bootstrapping dialog systems with word embeddings. In *NIPS, modern machine learning and natural language processing workshop*, volume 2, 2014.

[57] Christian Frings and Dirk Wentura. Trial-by-trial effects in the affective priming paradigm. *Acta Psychologica*, 128(2):318–323, 2008.

[58] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 1(3):4, 2016.

[59] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[60] Hugo Filipe Silveira Gamboa. *Multi-Modal Behavioral Biometrics Based on HCI and Electrophysiology*. PhD thesis, Universidade Técnica de Lisboa, 2008.

[61] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

[62] A. Ghandeharioun, B. Eoff, B. Jou, and R. W. Picard. Characterizing sources of uncertainty to proxy calibration and disambiguate annotator and data bias. In *ICCVW*, pages 4202–4206. IEEE, 2019.

[63] A. Ghandeharioun, S. Fedor, L. Sangermano, D. Ionescu, J. Alpert, Chelsea Dale, D. Sontag, and R. Picard. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In *ACII*. IEEE, 2017.

[64] A. Ghandeharioun, B. Kim, C. Li, B. Jou, B. Eoff, and R. Picard. DISSECT: Disentangled simultaneous explanations via concept traversals. 2021. In preparation.

[65] A. Ghandeharioun, B. Kim, C. Li, B. Jou, B. Eoff, and R. Picard. SynthDerm dataset, 2021.

[66] A. Ghandeharioun, L. Sangermano, R. Picard, J. Alpert, C. Dale, D. Ionescu, and S. Fedor. Objective vs. subjective reports of sleep quality in major depressive disorder: A pilot study. In *Anxiety and Depression Association of America*. ADAA, 2017.

[67] A. Ghandeharioun*, J. H. Shen*, N. Jaques*, C. Ferguson, N. Jones, A. Lapedriza, and R. Picard. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *NeurIPS*, pages 13658–13669, 2019. *Equal contribution.

[68] Asma Ghandeharioun. Dr. Tick: An android application for measuring heart rate and respiratory rate on smart phones, 2014. Bachelor's Thesis.

[69] Asma Ghandeharioun, Asaph Azaria, Sara Taylor, Pattie Maes, and Rosalind Picard. Promoting kindness and gratitude with a smartphone and triggers. *Annals of Behavioral Medicine*, 50(Supplement 1):266, 2016.

[70] Asma Ghandeharioun, Asaph Azaria, Sara Taylor, and Rosalind W Picard. "Kind and grateful": A context-sensitive smartphone app utilizing inspirational content to promote gratitude. *Psychology of well-being*, 6(1):9, 2016.

[71] Asma Ghandeharioun, Szymon Fedor, Lisa Sangermano, Jonathan Alpert, Chelsea Dale, Dawn Ionescu, and Rosalind Picard. Location variability from commodity phone sensors is negatively associated with self-reported depression score: A pilot study. In *Association for Psychological Sciences*. APS, 2017.

[72] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. EMMA: An emotion-aware wellbeing chatbot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.

[73] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. Towards understanding emotional intelligence for behavior change chatbots. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 8–14. IEEE, 2019.

[74] Asma Ghandeharioun, Lisa Sangermano, Rosalind Picard, Jonathan Alpert, Chelsea Dale, Dawn Ionescu, and Szymon Fedor. Objective vs. subjective reports of sleep quality in major depressive disorder: A pilot study. In *Anxiety and Depression Association of America*. ADAA, 2017.

[75] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

[76] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[77] Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291*, 2016.

[78] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 227–236, 2008.

[79] Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1):3–19, 2010.

[80] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[81] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[82] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[83] Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice*, pages 91–121. Springer, 1984.

[84] Isidore Gormezano and JW Moore. Classical conditioning. *Experimental methods and instrumentation in psychology*, 1:385–420, 1966.

[85] Yash Goyal, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (CaCE). *arXiv preprint arXiv:1907.07165*, 2019.

[86] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384, 2019.

[87] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE TBME*, 63(4):797–804, 2016.

[88] Agnes Grünerbl, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Oehler, Gerhard Tröster, Oscar Mayora, Christian Haring, and Paul Lukowicz. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE Journal of Biomedical and Health Informatics*, 19(1):140–148, 2014.

[89] Andrea Guidi, Sergio Salvi, Manuel Ottaviano, Claudio Gentili, Gilles Bertschy, Danilo de Rossi, Enzo Pasquale Scilingo, and Nicola Vanello. Smartphone application for the analysis of prosodic features in running speech with a focus on bipolar disorders: system performance evaluation and case study. *Sensors*, 15(11):28070–28087, 2015.

[90] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

[91] Suzanne N Haber and Scott L Rauch. Neurocircuitry: a window into the networks underlying neuropsychiatric disease. *Neuropsychopharmacology*, 35(1):1, 2010.

[92] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 890–897, 2017.

[93] Chikara Hashimoto and Manabu Sassano. Detecting absurd conversations from intelligent assistant logs by exploiting user feedback utterances. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 147–156. International World Wide Web Conferences Steering Committee, 2018.

[94] Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*, 2019.

[95] Jennifer Hay. Functions of humor in the conversations of men and women. *Journal of pragmatics*, 32(6):709–742, 2000.

[96] Javier Hernandez, Mohammed Ehsan Hoque, Will Drevo, and Rosalind W Picard. Mood meter: counting smiles in the wild. In *UbiComp 2012*, pages 301–310. ACM, 2012.

[97] Javier Hernandez, Yin Li, James M Rehg, and Rosalind W Picard. Bioglass: Physiological parameter estimation using a head-mounted wearable device. In *2014*

4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH), pages 55–58. IEEE, 2014.

[98] Javier Hernandez, Daniel McDuff, Xavier Benavides, Judith Amores, Pattie Maes, and Rosalind Picard. Autoemotive: bringing empathy to the driving experience to manage stress. In *Proceedings of the 2014 companion publication on Designing interactive systems*, pages 53–56. 2014.

[99] Javier Hernandez, Daniel McDuff, and Rosalind W Picard. Biowatch: estimation of heart and breathing rates from wrist motions. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 169–176. IEEE, 2015.

[100] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6629–6640, 2017.

[101] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*, 2(5):6, 2017.

[102] GE Hinton and Drew van Camp. Keeping neural networks simple by minimising the description length of weights. 1993. In *Proceedings of COLT-93*, pages 5–13.

[103] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[104] E. Howe, A. Ghandeharioun, P. Pedrelli, D. Mischoulon, R. Picard, and S. Fedor. Location patterns from phone sensors may help predict depressive symptoms: A longitudinal pilot study. *ABCT - Tech SIG*, 2017.

[105] E. Howe, M. Nauphal, B. Shapero, K. Bentley, D. Mischoulon, A. Ghandeharioun, S. Fedor, R. Picard, and P. Pedrelli. Depression and emotional reactivity: A closer examination of daily variations in affect. *ABCT*, 2018.

[106] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pages 263–272. Ieee, 2008.

[107] Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 49–54, 2018.

[108] Karen Huang, Michael Yeomans, Alison Wood Brooks, Julia Minson, and Francesca Gino. It doesn't hurt to ask: Question-asking increases liking. *Journal of personality and social psychology*, 113(3):430, 2017.

[109] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32, 2020.

[110] Molly E Ireland, Richard B Slatcher, Paul W Eastwick, Lauren E Scissors, Eli J Finkel, and James W Pennebaker. Language style matching predicts relationship initiation and stability. *Psychological science*, 22(1):39–44, 2011.

[111] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 3543–3556, 2019.

[112] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *NeurIPS workshop on Conversational AI*, 2019.

[113] Natasha Jaques*, Judy Hanwen Shen*, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind Picard. Human-centric dialog training via offline reinforcement learning. *EMNLP*, 2020. *Equal Contribution.

[114] Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *ACII*, pages 222–228. IEEE, 2015.

[115] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33, 2020.

[116] Noah Jones, Natasha Jaques, Pat Pataranutaporn, Asma Ghandeharioun, and Rosalind Picard. Analysis of online suicide risk with document embeddings and latent dirichlet allocation. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–5. IEEE, 2019.

[117] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

[118] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017.

[119] Elizabeth A Kensinger. Remembering the details: Effects of emotion. *Emotion review*, 1(2):99–113, 2009.

[120] Rajiv Khanna, Been Kim, Joydeep Ghosh, and Sanmi Koyejo. Interpreting black box predictions using Fisher kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 3382–3390, 2019.

[121] Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018.

[122] Been Kim, Cynthia Rudin, and Julie A Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in neural information processing systems*, pages 1952–1960, 2014.

[123] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pages 2673–2682, 2018.

[124] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2654–2663, 2018.

[125] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[126] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[127] Jack Klys, Jake Snell, and Richard Zemel. Learning latent subspaces in variational autoencoders. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6445–6455, 2018.

[128] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, volume 70, pages 1885–1894, 2017.

[129] Saskia Koldijk, Mark A Neerincx, and Wessel Kraaij. Detecting work stress in offices by combining unobtrusive sensors. *IEEE Transactions on affective computing*, 9(2):227–239, 2016.

[130] Kurt Kroenke and Robert L Spitzer. The phq-9: a new depression diagnostic and severity measure. *Psychiatric annals*, 32(9):509–515, 2002.

[131] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, pages 2539–2547, 2015.

[132] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

[133] Santosh Kumar, Gregory D Abowd, William T Abraham, Mustafa al'Absi, J Gayle Beck, Duen Horng Chau, Tyson Condie, David E Conroy, Emre Ertin, Deborah Estrin, et al. Center of excellence for mobile sensor data-to-knowledge (md2k). *Journal of the American Medical Informatics Association*, 22(6):1137–1142, 2015.

[134] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2017.

[135] Grace Leslie*, Asma Ghandeharioun*, Diane Zhou, and Rosalind W Picard. Engineering music to slow breathing and invite relaxed physiology. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019. *Equal Contribution.

[136] Penelope A Lewis and Hugo D Critchley. Mood-dependent memory. *Trends in cognitive sciences*, 7(10):431–433, 2003.

[137] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 994–1003, 2016.

[138] Jiwei Li and Dan Jurafsky. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, 2017.

[139] Jiwei Li, Will Monroe, Alan Ritter, et al. Deep reinforcement learning for dialogue generation. In *EMNLP*, pages 1192–1202, 2016.

[140] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, 2017.

[141] Sarah Lichtenstein and Baruch Fischhoff. Training for calibration. *Organizational Behavior and Human Performance*, 26(2):149–171, 1980.

[142] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 389–402, 2013.

[143] Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51, 2002.

[144] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[145] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, 2016.

[146] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE international conference on computer vision*, pages 3730–3738, 2015.

[147] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.

[148] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variations using few labels. In *International Conference on Learning Representations*, 2019.

[149] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, 2017.

[150] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 2017.

[151] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[152] David JC MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 354(1):73–80, 1995.

[153] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.

[154] Tan Margaret. Establishing mutual understanding in systems design: An empirical study. *Journal of Management Information Systems*, 10(4):159–182, 1994.

[155] Charles T Marx, Richard Lanas Phillips, Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Disentangling influence: Using disentangled representations to audit model predictions. *arXiv preprint arXiv:1906.08652*, 2019.

[156] Mark Matthews, Saeed Abdullah, Geri Gay, and Tanzeem Choudhury. Tracking mental well-being: Balancing rich sensing and patient needs. *Computer*, 47(4):36–43, 2014.

[157] Katie Matton, Melvin G McInnis, and Emily Mower Provost. Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder. In *Interspeech*, 2019.

[158] Alban Maxhuni, Angélica Muñoz-Meléndez, Venet Osmani, Humberto Perez, Oscar Mayora, and Eduardo F Morales. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing*, 31:50–66, 2016.

[159] Oscar Mayora, Mads Frost, Bert Arnrich, Franz Gravenhorst, Agnes Grunerbl, Amir Muaremi, Venet Osmani, Alessandro Puiatti, Nina Reichwaldt, Corinna Scharnweber, et al. Mobile health systems for bipolar disorder: the relevance of non-functional requirements in monarca project. In *E-Health and Telemedicine: Concepts, Methodologies, Tools, and Applications*, pages 1395–1405. IGI Global, 2016.

[160] Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, 2018.

[161] Daniel McDuff, Sarah Gontarek, and Rosalind Picard. Remote measurement of cognitive stress via heart rate variability. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2957–2960. IEEE, 2014.

[162] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pages 7775–7784, 2018.

[163] A. K. Meyer, S. Fedor, A. Ghandeharioun, D. Mischoulon, R. Picard, and P. Pedrelli. Feasibility and acceptability of the Empatica E4 sensor to passively assess physiological symptoms of depression. *ABCT*, 2020.

[164] Rhiannon Michelmore, Matthew Wicker, Luca Laurenti, Luca Cardelli, Yarin Gal, and Marta Kwiatkowska. Uncertainty quantification with statistical guarantees in end-to-end autonomous driving control. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7344–7350. IEEE, 2020.

[165] Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*, 2017.

[166] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, pages 236–244, 2008.

[167] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[168] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018.

[169] David C Mohr, Enid Montague, Colleen Stiles-Shields, Susan M Kaiser, Christopher Brenner, Eric Carty-Fickes, Hannah Palac, and Jenna Duffecy. Medlink: a mobile intervention to address failure points in the treatment of depression in general medicine. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*, pages 100–107. ICST, 2015.

[170] Pablo G Moreno, Antonio Artés-Rodríguez, Yee Whye Teh, and Fernando Perez-Cruz. Bayesian nonparametric crowdsourcing. *Journal of Machine Learning Research*, 2015.

[171] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[172] MovisensXS. eXperience Sampling for Android. `https://xs.movisens.com`, 2012. Online; accessed April'17.

[173] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019.

[174] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 1996.

[175] Barnaby Nelson, Patrick D McGorry, Marieke Wichers, Johanna TW Wigman, and Jessica A Hartmann. Moving from static to dynamic models of the onset of mental disorder: a review. *JAMA psychiatry*, 74(5):528–534, 2017.

[176] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–41, 2019.

[177] Douglas W Oard, Jinmook Kim, et al. Implicit feedback for recommender systems. In *Proceedings of the AAAI workshop on recommender systems*, volume 83. WoUongong, 1998.

[178] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Machine Learning*, volume 70, pages 2642–2651, 2017.

[179] Jose Oramas, Kaili Wang, and Tinne Tuytelaars. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In *International Conference on Learning Representations*, 2018.

[180] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[181] Alexandros Papangelis, Yi-Chia Wang, Piero Molino, and Gokhan Tur. Collaborative multi-agent dialogue model training via reinforcement learning. *arXiv preprint arXiv:1907.05507*, 2019.

[182] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[183] Yookoon Park, Jaemin Cho, and Gunhee Kim. A hierarchical latent structure for variational conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1792–1801, 2018.

[184] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[185] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*, 2018.

[186] P. Pedrelli*, S. Fedor*, A. Ghandeharioun, E. Howe, D. Ionescu, D. Bhathena, C. Dording, L. Fisher, C. Cusin, M. Nyer, A. Yeung, L. Sangermano, D. Mischoulon, J. Alpert, and R. Picard. Monitoring changes in depression severity using wearable and mobile sensors. 2020. *Equal contribution.

[187] P. Pedrelli, E. Howe, D. Mischoulon, R. Picard, A. Ghandeharioun, and S. Fedor. Integrating EMA, clinical assessment and wearable sensors to examine the association between major depressive disorder (MDD) and alcohol use. *Iproceedings*, 3(1):e51, 2017.

[188] Elizabeth A Phelps, Sam Ling, and Marisa Carrasco. Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological science*, 17(4):292–299, 2006.

[189] Rosalind Picard. Technology and Emotions, TEDxSF. `https://www.youtube.com/watch?v=ujxriwApPP4`, June 2011.

[190] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[191] Rosalind W Picard, Szymon Fedor, and Yadid Ayzenberg. Multiple arousal theory and daily-life electrodermal activity asymmetry. *Emotion Review*, 8(1):62–75, 2016.

[192] Gregory Plumb, Maruan Al-Shedivat, Ángel Alexander Cabrera, Adam Perer, Eric Xing, and Ameet Talwalkar. Regularizing black-box models for improved interpretability. *Advances in Neural Information Processing Systems*, 33, 2020.

[193] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, 2020.

[194] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *UbiComp 2011*, pages 385–394. ACM, 2011.

[195] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical

second opinions. In *International Conference on Machine Learning*, pages 5281–5290. PMLR, 2019.

[196] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, et al. Sequence level training with recurrent neural networks. *arXiv:1511.06732*, 2015.

[197] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. I know the feeling: Learning to converse with empathy. *arXiv preprint arXiv:1811.00207*, 2018.

[198] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, 2019.

[199] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 12268–12279. Curran Associates, Inc., 2019.

[200] Vikas C Raykar and Shipeng Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *The Journal of Machine Learning Research*, 13(1):491–518, 2012.

[201] Byron Reeves and Clifford Nass. *The media equation: How people treat computers, television, and new media like real people*. Cambridge university press Cambridge, UK, 1996.

[202] Darrell S Rigel, Robert J Friedman, Alfred W Kopf, and David Polsky. Abcde—an evolving concept in the early detection of melanoma. *Archives of dermatology*, 141(8):1032–1034, 2005.

[203] Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C Lipton. Decoding and diversity in machine translation. *NeurIPS Resistance AI Workshop*, 2020.

[204] Vasile Rus and Mihai Lintean. An optimal assessment of natural language student input using word-to-word similarity metrics. In *International Conference on Intelligent Tutoring Systems*, pages 675–676. Springer, 2012.

[205] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *JMIR*, 17(7):e175, 2015.

[206] Ardavan Saeedi, Matthew Hoffman, Stephen DiVerdi, Asma Ghandeharioun, Matthew Johnson, and Ryan Adams. Multimodal prediction and personalization of photo edits with deep generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 1309–1317. PMLR, 2018.

[207] Abdelrhman Saleh*, Natasha Jaques*, Asma Ghandeharioun, Judy Hanwen Shen, and Rosalind Picard. Hierarchical reinforcement learning for open-domain dialog. *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2019. *Equal Contribution.

[208] Tim Salimans, Ian J Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NIPS*, 2016.

[209] Pouya Samangouei, Ardavan Saeedi, Liam Nakagawa, and Nathan Silberman. ExplainGAN: Model explanation via decision boundary crossing transformations. In *European Conference on Computer Vision*, pages 666–681, 2018.

[210] David Sandberg. Open-source implementation of a face recognition model. `https://github.com/davidsandberg/facenet/wiki`, 2018. [Online; accessed 10-April-2019].

[211] Lisa Sangermano, Asma Ghandeharioun, Rosalind Picard, Jonathan Alpert, Chelsea Dale, Szymon Fedor, and Dawn Ionescu. Incoming cell phone data as a potential predictor of depression severity: A pilot study. In *Anxiety and Depression Association of America*. ADAA, 2017.

[212] Chinnadhurai Sankar and Sujith Ravi. Modeling non-goal oriented dialog with discrete attributes. In *NeurIPS Workshop on Conversational AI:"Today's Practice and Tomorrow's Potential*, 2018.

[213] Alberto Santamaria-Pang, James Kubricht, Aritra Chowdhury, Chitresh Bhushan, and Peter Tu. Towards emergent language symbolic semantic segmentation and model interpretability. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 326–334. Springer, 2020.

[214] Stefan Scherer, Gale M Lucas, Jonathan Gratch, Albert Skip Rizzo, and Louis-Philippe Morency. Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing*, 7(1):59–73, 2015.

[215] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[216] Peter Schulam and Suchi Saria. Can you trust this prediction? auditing pointwise reliability after learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1022–1031. PMLR, 2019.

[217] Joao Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. Chateval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, 2019.

[218] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*, 2019.

[219] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[220] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9030–9038, 2019.

[221] Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*, 2017.

[222] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[223] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[224] Sofia Serrano and Noah A Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, 2019.

[225] Pararth Shah, Dilek Hakkani-Tur, Bing Liu, and Gokhan Tur. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line

reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, 2018.

[226] Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. Improving variational encoder-decoders in dialogue generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[227] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020.

[228] Wilson Silva, Alexander Poellinger, Jaime S Cardoso, and Mauricio Reyes. Interpretability-guided content-based medical image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–314. Springer, 2020.

[229] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[230] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. In *International Conference on Learning Representations*, 2020.

[231] Sumedha Singla, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly-a counterfactual approach. *arXiv preprint arXiv:2101.04230*, 2021.

[232] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why modified bp attribution fails. *arXiv preprint arXiv:1912.09818*, 2019.

[233] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise. *ICML Workshop on Visualization for Deep Learning*, 2017.

[234] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc., 2020.

[235] Marilyn Strathern. 'improving ratings': audit in the british university system. *European review*, 5(3):305–321, 1997.

[236] Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 715–724, 2017.

[237] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, volume 70, pages 3319–3328, 2017.

[238] Michael Sung, Carl Marci, and Alex Pentland. Wearable feedback systems for rehabilitation. *Journal of neuroengineering and rehabilitation*, 2(1):17, 2005.

[239] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[240] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11244–11253, 2019.

[241] Corcoran R Rowse G. Telford C, McCarthy-Jones S. Experience sampling methodology studies of depression: the state of the art. *Psychological medicine*, 42(6):1119–1129, 2012.

[242] Hensin Tsao, Jeannette M Olazagasti, Kelly M Cordoro, Jerry D Brewer, Susan C Taylor, Jeremy S Bordeaux, Mary-Margaret Chren, Arthur J Sober, Connie Tegeler, Reva Bhushan, et al. Early detection of melanoma: reviewing the abcdes. *Journal of the American Academy of Dermatology*, 72(4):717–723, 2015.

[243] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pages 9625–9635. PMLR, 2020.

[244] Gaetano Valenza, Claudio Gentili, Antonio Lanatà, and Enzo Pasquale Scilingo. Mood recognition in bipolar patients through the psyche platform: Preliminary evaluations and perspectives. *Artificial intelligence in medicine*, 57(1):49–58, 2013.

[245] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 3–10, 2014.

[246] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. On evaluating and comparing conversational agents. *arXiv preprint arXiv:1801.03625*, 4:60–68, 2018.

[247] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[248] Nicholas G Ward, Hans O Doerr, and Michael C Storrie. Skin conductance: A potentially sensitive test for depression. *Psychiatry Research*, 10(4):295–302, 1983.

[249] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[250] Deena S Weisberg and Alison Gopnik. Pretense, counterfactuals, and bayesian causal models: Why what is not real really matters. *Cognitive science*, 37(7):1368–1381, 2013.

[251] Peter Welinder and Pietro Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 25–32. IEEE, 2010.

[252] WHO. World Health Organization news release. `http://www.who.int/mediacentre/news/releases/2017/world-health-day`, 1948. Online; accessed April'17.

[253] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, 2019.

[254] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: A roadmap for responsible machine learning for health care. *Nature medicine*, pages 1–4, 2019.

[255] Leonard J Williams. Tunnel vision induced by a foveal load manipulation. *Human factors*, 27(2):221–227, 1985.

[256] Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019.

[257] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[258] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic aware neural response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[259] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456. ACM, 2013.

[260] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in neural information processing systems*, 2018.

[261] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[262] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2017.

[263] Sebastian Zepf, Neska El Haouij, Jinmo Lee, Asma Ghandeharioun, Javier Hernandez, and Rosalind W Picard. Studying personalized just-in-time auditory breathing guides and potential safety implications during simulated driving. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 275–283, 2020.

[264] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[265] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, 2017.

[266] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[267] Xianda Zhou and William Yang Wang. Mojitalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, 2018.