The Science and Art of Human and Artificial Intelligence Collaboration

by

Matthew Groh

M.A., Massachusetts Institute of Technology (2019)
B.A. Middlebury College (2010)

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 in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Author____
Program in Media Arts and Sciences
May 8, 2023

Certified by____
Rosalind Picard
Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by____
Todd Machover
Academic Head
Program in Media Arts and Sciences

The Science and Art of Human and Artificial Intelligence Collaboration

by

Matthew Groh

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on May 8, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences

Abstract

While artificial intelligence (AI) appears to be surpassing the performance of human experts on a wide variety of games and real-world tasks, these algorithms are prone to systematic and surprising failures when deployed. In contrast to today's state-of-the-art algorithms, humans are highly capable of adapting to new contexts. The different strengths and weaknesses of humans and AI motivate a guiding research question for the emerging field of human-AI collaboration: When, where, why, and how does the combination of human problem solving and AI systems lead to a hybrid system that surpasses (or fails to surpass) the performance of either humans or the machine alone? This dissertation addresses various dimensions of this guiding question by conducting large-scale, digital experiments across three distinct tasks and domains: deepfake detection, dermatology diagnosis, and Wordle. First, the experiments in deepfake detection examine the similarities and differences between human and machine vision in identifying visual manipulations of people's faces in videos and identify important performance trade-offs between hybrid systems and human or AI only systems for deepfake detection. Second, the experiments in dermatology diagnosis reveal that non-visual information is often essential for diagnosing skin disease, diagnostic accuracy disparities across skin color exist in image-only store-and-forward teledermatology, and clinical decision support based on a fair deep learning system can significantly increase physicians' diagnostic accuracy in this experimental setting. Third, the experiment on Wordle demonstrates that digitally mediated expressions of empathy can counteract the negative effect of anger on human creative problem solving. In addition to these digital experiments, this dissertation presents two algorithmic audits on clinical dermatology images to reveal where systematic errors arise in state-of-the-art algorithms, examines how context influences automated affect recognition, and proposes methods for more effectively incorporating context in applied machine learning. Together, these contributions provide empirical evidence for why human-AI collaborations succeed and fail across a variety of tasks and domains, insights into how to design human-AI collaborations more effectively, and a framework for when and where hybrid systems should rely on human problem solving.

Thesis Supervisor: Rosalind Picard
Title: Professor of Media Arts and Sciences, Program in Media Arts and Sciences

The Science and Art of Human and Artificial Intelligence Collaboration

by

Matthew Groh

The following people served as readers for this thesis:

Thesis Reader____

Rosalind Picard
Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Thesis Reader____

David Rand
Erwin H. Schell Professor of Management Science and Brain and Cognitive Sciences
Massachusetts Institute of Technology

Thesis Reader____

Chaz Firestone
Assistant Professor of Psychological and Brain Sciences
Johns Hopkins University

Inspiration

“By this art you may contemplate the variations of the 23 letters...” – Robert Burton and Jorge Luis Borges

“The important thing is not to stop questioning. Curiosity has its own reason for existing. One cannot help but be in awe when contemplating the mysteries of eternity, of life, of the marvelous structure of reality.” – Albert Einstein

“Although there is a sense in which the camera does indeed capture reality, not just interpret it, photographs are as much an interpretation of the work as paintings and drawings are.” – Susan Sontag

“Understanding a theory means, I suggest, understanding it *as an attempt to solve a certain problem*.” – Karl Popper

“The hybrid or the meeting of two media is a moment of truth and revelation from which new form is born.” – Marshall McLuhan

Acknowledgments

Doing a PhD felt like a Type II fun adventure through the world of ideas. I am absolutely grateful for the gift of this experience, and I feel transformed as a result. And, I will be the first to admit the journey was hard. I felt quite lost at times. Everybody's PhD is unique and it takes time and effort to find one's way; my experience was no exception. For the reader considering a PhD, the reader in the midst of the quest, or the reader who is on any kind of intellectual voyage, I will share a couple insights in case they may helpful for you. First, curiosity can be both an energy source and a compass that leads you to the place you didn't even know you wanted to go. Sometimes it can be helpful to stop searching for an idea and simply follow your curiosity and let the idea find you. Second, the narrative structure of the Hero's Journey can be a useful map for recognizing one's the obstacles and opportunities along the PhD process. I did not discover the concept of the Hero's Journey until part way through my PhD, but once I did, I realized it offers a map that helped me better know myself in the context of the PhD process. Ideally, we would all minimize the "Refusal of the Call" and maximize "Apothesosis" but these are all transitory states that happen over and over again. The point of sharing this structure is to help promote reflection and recognition of the many stages that appear along the way of seeking something unique. Finally, the PhD journey is a balance of the solitary and social. One needs time and space to think independently, and one also needs help from mentors, peers, strangers, friends, and family.

"Human-human" collaboration underpins the "Human-AI" contributions in this dissertation. I feel deeply grateful for the many brilliant and kind people who inspired and encourage me along the way. Now, it's time for the many, specific acknowledgments of the many friends made along the way of writing this dissertation:

My Media Lab advisor, Rosalind Picard, for her keen intellect, emotional intelligence, engineering mindset, and brilliant imagination. Thank you for always encouraging deeper levels of analysis, nurturing interdisciplinary ideas, and genuinely caring.

My outside Media Lab advisors, Chaz Firestone and David Rand for welcoming me into their

research labs, demonstrating effective science communication, and sharing countless insights on how to make research both more timeless and timely. Chaz, thank you for introducing me to the world of vision science and constantly revealing points of inspiration, creativity, and curiosity. Dave, thank you for introducing me to social psychology and management science and always encouraging careful analysis and thoughtful framing.

Thank you Linda Peterson and Tod Machover for helping me navigate the Media Lab. Thank you Joost Bonsen for always willing to share an ear or a story (or three!). Thank you Andy Lippman for livening so many conversations with your sharp wit. Thank you Esteban Moro for always caring and engaging in so many fun science and metascience conversations.

Thank you:

Zivvy Epstein, for a partnership in mystical mischief.

Alex Berke, for always honoring curiosity and exploring ideas to their fullest.

Michiel Bakker, for relentless extraversion and late night class projects.

Morgan Frank, for sociotechnical creativity and his musical keyboard.

Aruna Sankaranarayanan, for the best emoji game and amazing thoughtfulness.

Joy Buolamwini, for showing the way on how interdisciplinary science and art can change the world.

Amazing labmates in Affective Computing, Asma Ghandeharioun, Rob Lewis, Noah Jones, Katie Matton, Sable Aragon, Neska Elhaouij, Ila Kumar, Boyu Zhang, Agata Lapedriza, Szymon Fedor, Kristy Johnson, and Vincent Chen, and fantastic labmates throughout the Media Lab, Tobin South, Robert Mahari, Nikhil Singh, Manaswi Mishra, Oceane Boulais, Ramon Weber, Micah Epstein, Sebastien Kamau, David Ramsay, Abdul Alotaibi, Irmandy Wicaksono, Pat Pataranutaporn, Vald Danry, Devora Najjar, Caroline Jaffe, Filippos Tourlomousis, Rubez Ming, Matt Carney, Samantha Gutierrez-Arango, Alessandra Davy-Falconi, Erik Strand, Dan “Noyvsan” Novy, Tara Sowrirajan, Michael Stern, Adis Ojeda, Mohammed

Alsobay, Hope Schroeder, and many more people at the Media Lab, for fantastic camaraderie.

Coauthors from outside the Media Lab, Luis Soenksen, Omar Badri, Arash Koochek, Roxana Daneshjou, and Murali Doraiswamy, for expertise and rigor turned research and friendships.

Aerin Kim, Kira Prentice, Erik Duhaime, and Joanna Molke, who provided ideas and resources for collecting large-scale datasets in this dissertation.

My economist mentors, David McKenzie and Tara Vishwanath. David, thank you for introducing me to the world of randomized experiments and trusting me as a 22-year-old with the responsibility for managing research projects across the globe. Tara, thank you for wisdom, open-mindedness, and guidance in finding my path, and I am ever grateful that you encouraged me to go back to school!

My large-scale optimization mentor Matt Kraning, for sharing insights on how to build technology that is indistinguishable from magic and how to build a business around that technology.

My undergraduate thesis advisor, Casey Rothschild, for being the first professor to suggest that I pursue a PhD. It took me awhile to take his sage advice, and I am happy I did.

My many teachers including Tucker Hiatt for inspiring a deep interest in cognitive science and science communication from his physics courses and Norm Lyons for instilling curiosity in operations research from his algebra course.

The Lavoie family, for making Boston a home away from home.

My friends, for so many good laughs and wild adventures.

Alex, for her love, support, patience, wisdom, and the joy, inspiration, sanity, conscientiousness, and kindness she brings to the world everyday.

Bros, dad, and mom for **everything**.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The guiding question of human-AI collaboration is: When, where, why, and how does the combination of human problem solving and AI systems lead to a hybrid system that surpasses (or fails to surpass) the performance of either humans or the machine alone? If a task can be solved perfectly by humans or machines (e.g. multiplying two numbers or playing tick tack toe), then human-AI collaboration is unnecessary. For any other problem solving task where both humans and machines are prone to some error, the guiding question of human-AI collaboration becomes useful for designing hybrid systems. However, this question is difficult to answer because there is often a trade-off in statistical learning about what is being optimized: explanations or predictions [83, 278]. The optimization trade-off leads to an interpretability-accuracy trade-off. For example linear regressions and logical rules are often much easier to interpret than neural networks and reinforcement learning, but the more complex algorithms are often more accurate for many perceptual tasks. In particular, it can be difficult to distinguish when humans should override or defer to machine predictions (and similarly when machines should be allowed to override or forced to defer to human decisions). Naturally, the answer depends on human and machine capabilities in a particular context. Moreover, the answer involves building mental models of how algorithms and humans perform. Ideally, human-AI collaboration combines the strengths of humans and machines to produce a system that is more effective than either alone as characterized in Figure 1-1. However, a recent literature review reveals that more often than not the reality of human-AI collaboration is it is less effective than either humans or AI alone [101] as characterized in Figure 1-2. In order to effectively combine human problem solving with artificial intelligence and develop a generalizable theory of intelligence augmentation, it is necessary to identify the strengths and weaknesses of humans and machines and how humans and machines can come to model each other's strengths and weaknesses across diverse domains, tasks, and contexts.

The introduction to this dissertation proceeds by highlighting both the successes and the systematic and surprising failures that have been documented in applications of machine learning. Next, the introduction provides a brief historical context for the emerging field of human-AI collaboration and describes the process of building a theory of mind and theory of machine. With the high level framework in mind, the introduction examines past model

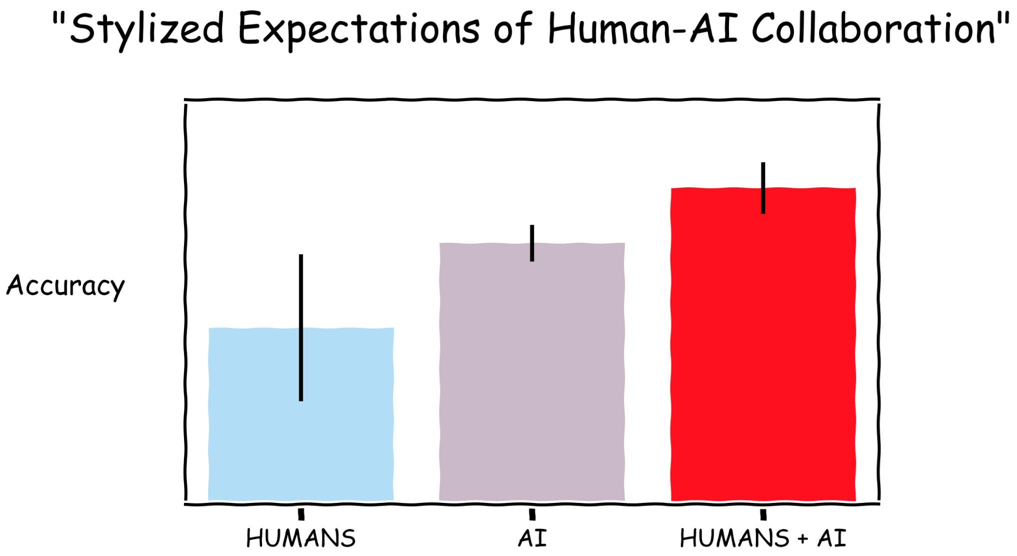


Figure 1-1: The data behind the bar plot and 95% confidence intervals are made up and intended to illustrate the common idealized expectations of integrating human problem solving with artificial intelligence.

organisms for human and machine problem solving and introduces the three domains – deepfake detection, dermatology diagnosis, and a digital game for examining digitally mediated expressions of empathy – that this dissertation examines. Finally, the introduction presents an overview of the rest of the dissertation.

1.1 Successes and Surprising yet Systematic Failures of Machine Learning

Recently, researchers have demonstrated that machine learning models can outperform human experts on a wide variety of games (e.g., Chess [100], Arimaa [594], Go [568], 57 classic Atari games [557], Poker [84], StarCraft II [622], Crosswords [625], Diplomacy [1], and more). In reinforcement learning models trained to play games with well-defined rules, the models benefit from the constraints of the known finite set of actions. Unlike these games with well-defined rules, most real-world problems are too high-dimensional and data collection is too constrained for a model to access data on the entire state space of environments

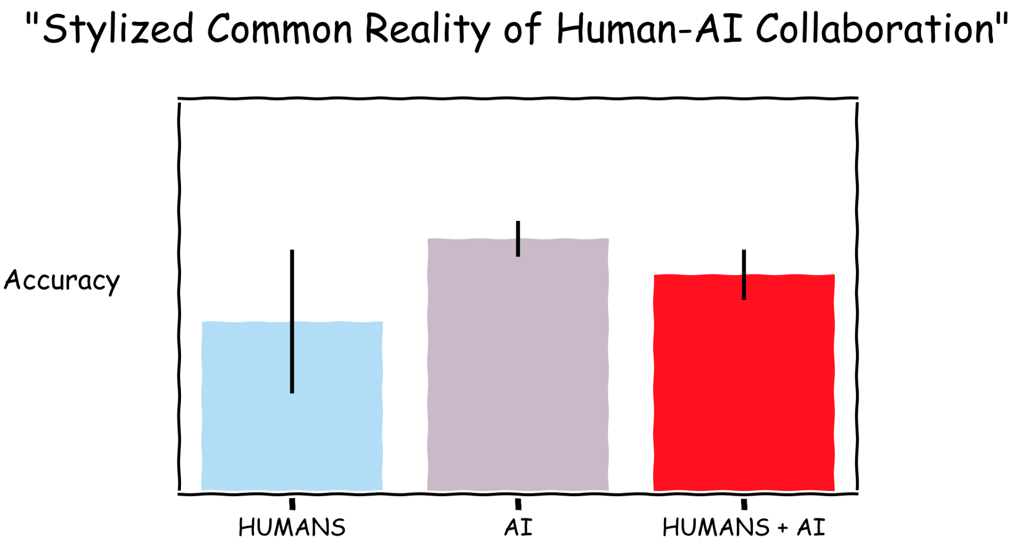


Figure 1-2: The data behind the bar plot and 95% confidence intervals are made up and intended to illustrate the common realities of integrating human problem solving with artificial intelligence as revealed in a large literature review evaluating 79 human-AI collaboration studies [101]

and agent actions. However, if a problem is sufficiently constrained to collect enough data across the problem solving state space, then machine learning models have the potential to perform well. For example, researchers have demonstrated that models can perform at the level of medical specialists on specific healthcare tasks (e.g., identifying breast cancer in mammograms [418], classifying skin lesions based on a single image [187], and predicting the diagnosis of hundreds of diverse skin conditions based on a few images and a brief patient history [388]) and outperform experts on other specific tasks (e.g. face identification to determine whether pairs of face images show the same or different person [499] and natural language understanding tasks [586, 627]). In addition, recent research – that has not yet been peer-reviewed – reveals large language models – trained on natural language containing at least a trillion words [361] – can pass professional exams such as the Uniform Bar Exam and United States Medical Licensing Examination [319, 464] and score highly in other standardized testing settings. Despite machine learning models' impressive performance on many specific tasks and standardized testing, machine learning models tend to be brittle in the face of changing contexts [242].

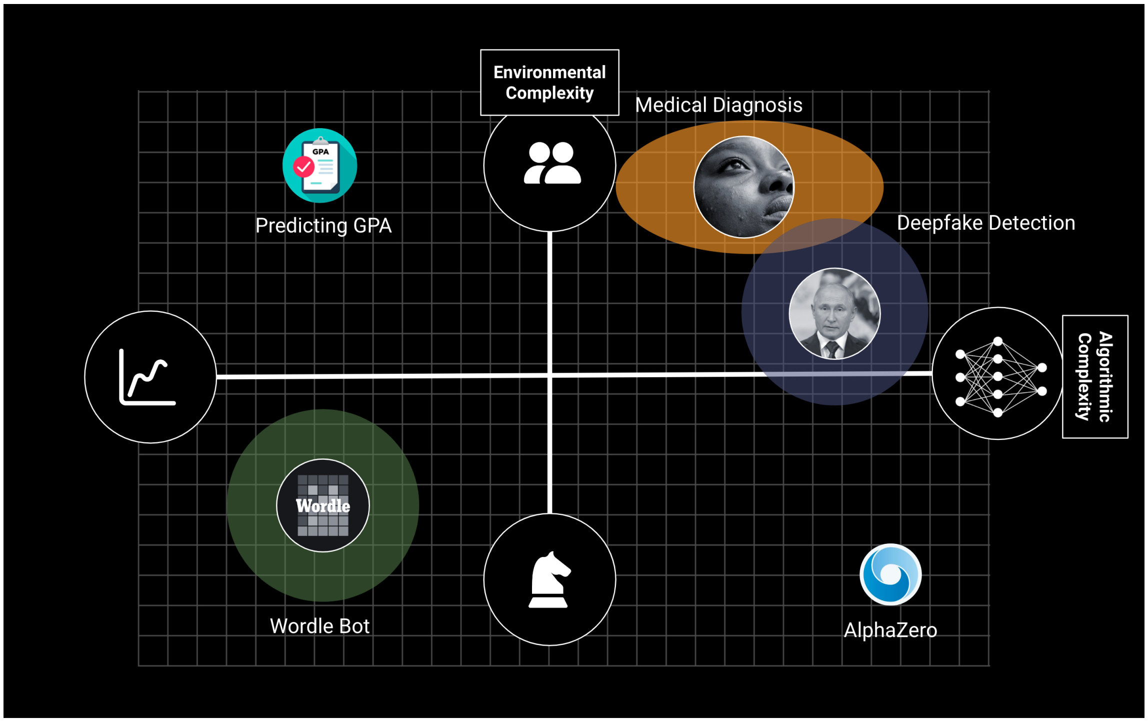


Figure 1-3: The problem solving space represented as a map with two dimensions: (1) algorithmic complexity on the x-axis where the left side represents solving problems with linear regressions, logical rules, and semantically meaningful features and the right side represents solving problems with neural networks, reinforcement learning and perceptual data (2) environmental complexity on the y-axis where the top represents social situations with changing dynamics and the bottom represents games with well-defined rules. AlphaZero is an example of low environmental, high algorithmic complexity because it is based on reinforcement learning and self-play without any guidance beyond the rules of Go (or other games like Chess and Shogi) [568, 569]. Predicting GPA represents high environmental complexity (GPA is a proxy for academic achievement, which involves complex social dynamics) and low algorithmic complexity (early research on human-AI collaboration involved researchers predicting GPA using ordinary least squares regression on semantically meaningful features) [164]. Medical diagnosis and deepfake detection are problems involving both high environmental complexity (these are high-context real-world problems involving dimensions of deception, subjectivity in annotations, and social influences) and algorithmic complexity (the visual components of these problems involve neural networks trained on visual features). The Wordle bot illustrates a game that can be solved exactly with dynamic programming based on the rules and number of possible words [64], which becomes useful for examining complex social and emotional phenomena that we address in Chapter 7.

Machine learning models trained to complete real-world tasks are prone to surprising yet systematic errors. For example, adversarial perturbations of data lead to errors that reasonable humans would not make (e.g. misclassifying a stop sign based on a small adversarial sticker [85], falling for adversarial perturbations to 3D objects [38], learning false representations from backdoor poisoning attacks [119], and falling for adversarial Go strategies such that human Go amateurs can defeat superhuman Go AIs [630]). In the healthcare domain, machine learning models make errors that appear nonsensical if made by healthcare practitioners (e.g. changing classification of skin lesions in dermoscopic images based on the presence of surgical markings [643] and mistaking radiographs with chest drains as clinically relevant pneumothorax cases [468]). The development of machine learning models for human-centered applications requires careful considerations of what is being optimized [24, 114, 444, 589], and poor model design can lead to socially undesirable outcomes (e.g., recognizing faces of women and people with dark skin less accurately than men and people with light skin [87], amplifying existing inequities [49, 470, 561], and magnifying moral hazard and error [443]).

The unexpected and unintended errors of machine learning models often arise from a model's inability to adapt to contexts in which it has not yet been sufficiently trained. In statistical learning, supervised models are traditionally developed on a test set, tuned on a validation set, and evaluated on a test set. When this model is deployed in production, the accuracy on the production sets is only expected to resemble the accuracy of the test set to the extent that the test and production data both arise from the same data generating functions i.e. the same independent and identically distributed (i.i.d) random variables. If production data deviates from test data (because the sensors to collect the data have changed, the objects of interest have changed, or anything else has changed), then it is likely that unexpected errors will arise and the performance on the production data will not match the performance on the test data.

1.2 Speculations on Human-AI Collaboration

The concept of AI and human-AI collaboration has long existed in the human imagination. Ancient speculations on human-AI collaboration can be found in the Greek myths such as Talos, a giant bronze animated statue (what we would call a robot today) who is "made not born" and built to defend the island of Crete from invaders [411]. Akin to the surprising yet systematic machine learning failures, Talos' ultimate demise came about from an unforeseen vulnerability: a single screw coming loose and the subsquent draining of his power source [411]. More recently, 20th century speculations on human-AI collaboration can be found in the scholarly work published during the emergence of artificial intelligence as a field of study. In 1958, Allen Newell, John Shaw, and Herbert Simon conceptualized the necessary elements for developing a theory of human problem solving and presented an example of how a computer could predict human behavior [455]. Just two years later, Joseph Licklider speculated that a human-machine symbiosis would "facilitate formulative thinking" and "enable [people] and computers to cooperate in making decisions and controlling complex situations without inflexible dependence on predetermined programs" [384]. Another two years later, Douglas Engelbart imagined a conceptual framework for identifying "the factors that limit the effectiveness of the individual's basic information-handling capabilities" and developing systems to enhance and augment these capabilities. In a 1971 review of their theory of human problem solving, Herbert Simon and Allen Newell present an 11-step strategy for developing a theory of human problem solving, describe human information processing as a series of "simple schemes of heuristic search," and speculate on how complex problem solving programs can develop via a computational step by step process [571]. In 1995, Pattie Maes speculated on autonomous agents that can reduce information overload and supported her ideas with a number of prototypes [398]. In the same year, Rosalind Picard published Technical Report No. 321 introducing the concept of affective computing, which imagines how computers may come to recognize and respond to affect [500].

1.3 Building a Theory of Mind and Machine

Over the last decade, an empirical science of human-AI collaboration has begun to emerge across a wide range of fields including organizational behavior [164, 389], human-computer interaction [101, 359], information systems [40, 369], medicine [445, 608], law [335, 336], and cognitive science [205, 516, 566].

In an influential study on human-AI collaboration, which focuses on two tasks in the upper left quadrant of Figure 1-3 (predicting MBA students' GPAs from the admission materials and predicting the rank of U.S. states in terms of the number of airline passengers that departed from that state in 2011), researchers identify a phenomenon they call “algorithm aversion” where people lose confidence in an algorithm after it makes an error more quickly than they otherwise would lose confidence if they thought the algorithm was a human [164]. In another study on human-AI collaboration, which also focuses on tasks (estimating the weight of an individual from a photograph, forecasting the popularity of songs, estimating a person's attractiveness based on a text description from the perspective of a person in a photograph, and forecasting economic and political events) that would generally be located in the upper left quadrant of Figure 1-3, researchers identify what they call the “algorithm appreciation” effect which demonstrates that across a variety of upper left quadrant tasks, people prefer advice from algorithms to advice from people [389]. Despite what sound like two contradictory effects, “algorithm aversion” and “algorithm appreciation” are not contradictory because the first refers to how people update and the second refers to overall advice taking. These experiments “scratch the surface of *theory of machine*” [389] and provide initial insights into what we might expect of human-AI collaboration, but many questions and boundary conditions are left unexplored for future research to examine including the degree to which the results depend on the level of algorithmic performance, familiarity of the individual with the algorithm, the interface by which the algorithm provides information, the kind of problem, people's expertise in the domain, people's understanding of how the algorithm works, whether the decision is high- or low-stakes, the evaluation metrics, and many other dimensions and interactions between these dimensions.

An understanding of the dynamics of humans' theory of machine, machine's theory of mind, and human-AI collaboration depends on building theories of human problem solving [571], theories of machine behavior [36, 516], careful consideration of the difference in agents' internal capacities versus ability to demonstrate those capacities [205], and empirical examination of hybrid human-AI systems across diverse, real-world problem to address the large research design space [20].

1.4 Model Organisms for Human-AI Collaboration

In the 20th century, researchers described the game of chess as the drosophila of both cognitive psychology and AI [412, 570]. By calling chess a drosophila of these two fields, researchers were making an analogy to model organisms studied in biology to convey that chess represents a particularly apt space for studying problem solving. So, why has chess been a great model organism for cognitive psychology and AI? Chess is a well-known, two-player, turn-taking game with simple rules, a clear objective, and a large decision-making possibility space of moves to make. From a high-level problem solving perspective, chess involves recognizing patterns (heuristic search), planning one's moves, anticipating one's opponent's moves, considering counterfactuals, and making decisions potentially under time pressure. Beyond the simple rules and complex strategy that make chess a model organism for studying problem solving in both cognitive psychology and AI, chess has cultural capital. Before artificial intelligence got its name, G.H. Hardy wrote, "chess problems are the hymn tunes of mathematics" [268]. As another example of human-machine collaboration in chess in the 18th century, Baron von Kempelen toured the *Mechanical Turk* around Europe and led his audiences to believe they were witnessing a robot playing chess when the trick was a human chess master artfully hiding inside and operating the machine.

Chess and other games with simple, well-defined rules and objectives (e.g. Go, Poker, and Diplomacy) can also serve as model organisms for human-AI collaboration. Former World Chess Champion, Gary Kasparov, who was the first World Chess Champion to ever be defeated by an AI, wrote "There will be cases where an AI will fail to detect exceptions to their rules. Therefore, we must work together, to combine our strengths. I know better than most people what it's like to compete against a machine. Instead of raging against them, it's better if we're all on the same side" [317]. Recent research on chess knowledge acquisition by AlphaZero, a reinforcement learning model that learns through self-play, reveals that AlphaZero exhibits many human concepts as it learns to play chess (e.g. material score, position, king safety, and many more concepts detailed in Table S1 of McGrath et al 2022) and poses a follow up: "Can we go beyond finding human knowledge and learn something new?" [416] Following the advent of AlphaGo, researchers have demonstrated that professional human Go performance has significantly improved and involves more novel moves than before [566]. However, open questions remain about the conceptual and qualitative differences in pre- and post-AlphaGo professional Go decision-making.

While Chess, Go, and other games may serve as model organisms for studying human-AI collaboration, these games do not represent the complexity of the real-world and expose human-AI collaboration to the ludic fallacy where insights may be limited to the "narrow world of games and dice" [596]. In particular, games typically involve well-defined rules and objective performance criteria, which can be optimized via reinforcement learning with enough computational resources. In contrast, social settings represent complex environments with ill-defined, dynamic rules, criteria for success, and dependencies on contextual information that may be difficult to collect. In Figure 1-3, I offer a map for situating model organisms in the human-AI collaboration problem solving space; the y-axis represents environmental complexity and the x-axis represents algorithmic complexity. Chess and Go, well-defined state spaces that are most appropriately traversed with relatively complex algorithms, are located in the bottom left quadrant. If Chess and Go occupy one part of one quadrant, then a natural question becomes what model organisms may be appropriate for examining human-AI collaboration along these other dimensions of problem solving?

Perhaps, the simplest form of human-AI collaboration is a human using a calculator to solve an arithmetic problem e.g. 7*857 or 6000-1? Back in first grade, I remember discovering it's quicker for a human with a calculator to solve the first kind of problem with a calculator but the second problem in one's head. Human-calculator collaboration for arithmetic is deterministic, and as such, humans can simply generally rely on arithmetic calculations (though it is wise to be careful with floating point arithmetic when you need extreme precision). Moreover, arithmetic is verifiable. If I claim 7*857=5999, then you can verify the result for yourself. This closed-system, immediate verification sets math and many games apart from the real world. In games like Wordle, salet (a variant of sallet, which means a combat helmet from Renaissance-era Europe) turns out to be the optimal word for beginning Wordle, which can be verified by dynamic programming [64]. If a problem can be verifiably solved perfectly by a set of rules from the given information, then it's likely not a relevant problem for examining how to optimize hybrid system performance. However, the bottom left quadrant in Figure 1-3 may still be interesting and useful because the problem solving state space enables systematic examination of human problem solving.

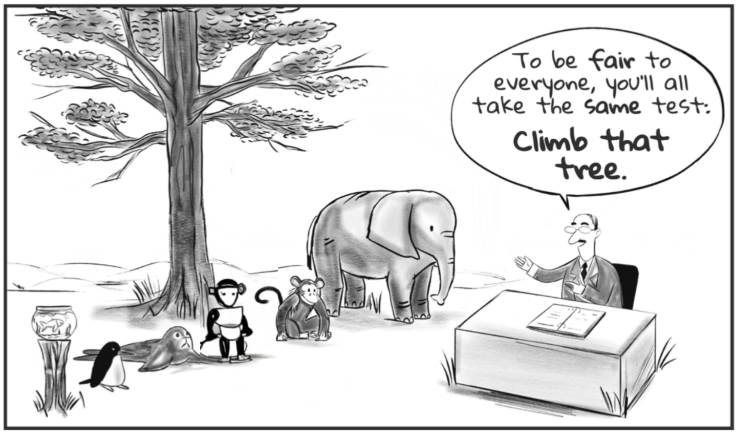
Two recent literature reviews reveal a high degree of diversity in the kinds of tasks evaluated for human-AI collaboration. In one literature review evaluating 79 experimental results on the performance of human computer systems, tasks ranged from answering natural language questions, teaching math concepts, detecting objects in images (general object detection, car scratches, wildlife conservation, objects in endoscopic images), labeling data, detecting deepfakes, predicting survival on the Titanic, playing games, predicting recidivism, distinguishing truthful and deceptive statements, predicting a person's profession, predicting fraud, diagnosing dermatologic conditions, classifying recyclables, prescribing drugs, and editing food ingredients [101]. In another literature review synthesizing 80 papers primarily drawn from the field of human-computer interaction, tasks ranged from legal (predicting recidivism, bail outcomes and child maltreatment), healthcare (diagnosing disease, classifying cancer, annotating clinical notes, and assessing stroke rehabilitation), business (predicting income, loan risk, sales, property prices, money laundering, stock prices, marketing, and insurance prices), education (predicting student performance, student admission, student dropout, and LSAT answers), leisure (recommending movies, music, dates, news, and trivia, playing chess, and classifying plants), professional (predicting job promotion, scheduling meetings, classifying email topics, monitoring cybersecurity, planning military missions), and other (classifying images, analyzing sentiment, answering natural language questions, detection deception, predicting forest cover, nutrition, toxicity, person weight, attractiveness, and religion) [359].

Insights on human-collaboration may come from surprising spaces; for example, researchers revealed that human experts with access to a deep learning model for restoring and dating ancient texts surpass the time-restricted accuracy of experts or the model alone [37]. In particular, the model trained on 78,608 inscriptions of ancient text can offer predictions of the most likely missing characters in a given ancient tablet, but it lacks historical and material context for what predictions make most sense. As a result, experts can rapidly consider many reasonable options, which leads to higher concordance with established specialists' restorations under time constraints than either experts or the model alone.

This dissertation aims to build towards a theory of intelligence augmentation by considering three model organisms: deepfake detection, medical diagnosis, and digital games. I pursued questions across these three domains because the first two domains are model organisms for complex, visually-oriented problems with high-stakes and important societal implications (upper right quadrant of Figure 1-3) and the third domain is a simple model organism for evaluating how expressions of empathy can influence human problem solving (bottom left quadrant of Figure 1-3). The choice of these three domains draws on solution-oriented computational social science [364, 366, 633] where a researcher starts with a practical problem and asks what theories and methods are required to solve the problem. Part of the art of human-AI collaboration is identifying the relevant contexts for comprehensively and effectively building mental and statistical models of human and AI performance across domains. For example, professional decision makers (e.g. physicians, judges, managers, content moderators, digital forensics experts, and others) often have access to private information unavailable to an algorithm, and it becomes important to characterize how this private information influences problem solving in order to understand when and why human and machine performance deviates. The inclusion of a diversity of complex domains and tasks is intended to allow for embracing the hard challenges of AI [434], uncovering generalizable insights, and working towards a theory that avoids reductionary comparisons like the one featured in the caricature in Figure 1-4.

1.5 Overview

This dissertation is divided into an introduction, 4 sections on deepfake detection, dermatology diagnosis, affective computing, and robustness in applied machine learning that make up 8 body chapters, a conclusion, and a final chapter with citations. Each body chapter is written to stand on its own. Half of the body chapters have been peer-reviewed and published in journals or conferences (Chapter 2 [243], 4 [248], 5 [246], and 7 [245]) and the other half are under review or drafts to be submitted for peer review (Chapter 3, 6, 8, and


To be fair to everyone, you'll all take the same test: Climb that tree.

"Intelligent systems may fail to express their knowledge in tests that don't accommodate their particular performance constraints." Firestone 2020

Figure 1-4: The figure is adapted from *Performance vs. competence in human-machine comparisons* [205]. Image credit: Victoria Dimitrova (artist); inspired by Hans Traxler, and edited by the author using DALL-E.

9).

1.5.1 Deepfake Detection

The first section – chapters 2 and 3 – presents large digital experiments on deepfake detection. Chapter 2 examines the similarities and differences between human and machine vision in identifying visual manipulations of people's faces in videos and identifies important performance trade-offs between hybrid systems and human or AI only systems for deepfake detection. Chapter 3 further examines human performance in deepfake detection by experimentally evaluating the degree to which humans can distinguish authentic and fabricated political speeches relies on the content of what is said versus the audio-visual cues of how it is said.

1.5.2 Dermatology Diagnosis

The second section – chapters 4, 5, and 6 – presents two algorithmic audits of and a large digital experiment on dermatology diagnosis. Chapter 4 investigates algorithmic accuracy disparities in machine learning applied to open-source data with over 15,000 clinical dermatology images and crowdsourced Fitzpatrick skin type labels. Chapter 5 further investigates the subjectivity of estimating Fitzpatrick skin type labels from clinical dermatology images by evaluating the inter-rater reliability of experts, crowds, and an algorithm. Chapter 6 examines diagnostic accuracy of specialists, generalists, and physician-machine partnerships in a store-and-forward dermatology diagnosis experiment with a large number of images of skin conditions across a diverse range of skin colors.

1.5.3 Affective Computing

The third section – chapters 7 and 8 – presents a large experiment on digitally mediated expressions of empathy and a systematic review and analysis of how context influences automated affect recognition. Chapter 7 examines how digitally mediated expressions of empathy interacts with incidental emotions to influence human problem solving in Wordle. Chapter 8 reviews research in affective computing, affective science, and artificial intelligence to build an initial framework for systematically identifying the roles of context in automated affect recognition.

1.5.4 Robustness in Applied Machine Learning

The fourth section – chapter 9 – proposes methods for more effectively incorporating context in applied machine learning. Specifically, chapter 9 reveals the conceptual problems with distribution shift, introduces the concept of context shift, and offers three approaches for building robustness in applied machine learning: integrating human intuition and expertise in identifying potential context shifts, evaluating models with dynamic benchmarks, and articulating the limitations of machine learning models.

Finally, chapter 10 concludes with a discussion, future research questions, and an example of human-AI collaboration for transforming prose to verse.

Chapter 2

Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds

Abstract

The recent emergence of machine-manipulated media raises an important societal question: how can we know if a video that we watch is real or fake? In two online studies with 15,016 participants, we present authentic videos and deepfakes and ask participants to identify which is which. We compare the performance of ordinary human observers against the leading computer vision deepfake detection model and find them similarly accurate while making different kinds of mistakes. Together, participants with access to the model's prediction are more accurate than either alone, but inaccurate model predictions often decrease participants' accuracy. To probe the relative strengths and weaknesses of humans and machines as detectors of deepfakes, we examine human and machine performance across video-level features, and we evaluate the impact of pre-registered randomized interventions on deepfake detection. We find that manipulations designed to disrupt visual processing of faces hinder human participants' performance while mostly not affecting the model's performance, suggesting a role for specialized cognitive capacities in explaining human deepfake detection performance.[1]

[1] This chapter, which is co-authored by Ziv Epstein, Chaz Firestone, and Rosalind Picard, appeared as a research article on December 28, 2021 in the Proceedings of the National Academy of Science [243].

2.1 Motivation

How do we tell the difference between the genuine and the artificial? The emergence of deepfakes – videos that have been manipulated by neural network models to either swap one individual's face for another, or alter the individual's face to make them appear to say something they have not said – presents challenges both for individuals and for society at large. Whereas a video of an individual performing an action or making a statement has long been one of the strongest pieces of evidence that the relevant event actually occurred, deepfakes undermine this gold standard, with potentially alarming consequences [120, 365, 371, 484].

How should we best meet this new challenge of evaluating the authenticity of a video? One approach is to build automated deepfake detection systems that analyze videos and attempt to classify their authenticity. Recent advances in training neural networks for computer vision reveal that algorithms are capable of surpassing the performance of human experts in some complex strategy games [567, 568] and medical diagnoses [187, 418], so we might expect algorithms to be similarly capable of outperforming people at deepfake detection. Indeed, such computational methods often surpass human performance in detecting physical implausibility cues [191], such as geometric inconsistencies of shadows, reflections, and distortions of perspective images [318, 460, 461]. Similarly, face recognition algorithms often outperform forensic examiners (who are significantly better than ordinary people) at identifying whether pairs of face images show the same or different people [499]. This focus on automating the analysis of visual content has advantages over certain methods from traditional digital media forensics, which often rely on image metadata [190] that are not available for many of today's most concerning deepfakes, which typically appear first on social media platforms stripped of such metadata [251, 397]. Moreover, metadata from an individual's decision to share on social media may not be a reliable predictor of media's veracity because social media tends to focus people's attention on factors other than truth and accuracy [495, 497].

The artificial intelligence (AI) approach to classifying videos as real or fake focuses on developing large datasets and training computer vision algorithms on these datasets [11, 166, 168, 303, 348, 383, 403, 432, 534, 535, 604, 620, 656]. The largest open-source dataset is the Deepfake Detection Challenge (DFDC) dataset, which consists of 23,654 original videos showing 960 consenting individuals and 104,500 corresponding deepfake videos produced from the original videos. The first frames of both a deepfake and original video from this dataset appear in Figure 2-1. The deepfakes examined here contain only visual manipulations produced using seven synthetic techniques: two deepfake autoencoders, a neural network face swap model [284], the NTH talking heads model [661], the FSGAN method for reenactment and inpainting [462], StyleGAN [314], and sharpening refinement on blended faces [168]. Unlike viral deepfake videos of politicians and other famous people, the videos from the competition have minimal context: they are all 10 second videos depicting unknown actors making uncontroversial statements in nondescript locations. As such, the cues for discerning real from fake can be based only on visual cues and not auditory cues or background knowledge of an individual or the topic they are discussing. In a contest run from 2019 to 2020, The Partnership for AI, in collaboration with large companies including Facebook, Microsoft, and Amazon, offered $1,000,000 in prize money to the most accurate deepfake detection models on the DFDC holdout set via Kaggle, a machine learning competition website. A total of 2,116 teams submitted computer vision models to the competition, and the leading model achieved an accuracy score of 65% on the 4,000 videos in the holdout data, which consisted of half deepfake and half real videos [167, 168]. While there are many proposed techniques for algorithmically detecting fakes (including affective computing approaches like examining heart rate and breathing rate [513] and looking for emotion-congruent speech and facial expressions) [12, 435], the most accurate computer vision model in the contest [560] focused on locating faces in a sample of static frames using multitask cascaded convolutional neural networks [663], conducting feature encoding based on EfficientNet B-7 [597], and training the model with a variety of transformations inspired by albumentations [88] and grid-mask [115]. Based on this model outperforming 2,115 other models to win significant prize money in a widely publicized competition on the largest dataset of deepfakes ever produced, we refer to this winning model as the “leading model” for detecting deepfakes to date.

The rules of the competition strictly forbid human-in-the-loop approaches, which leaves open questions surrounding how well human-AI collaborative systems would perform at discerning between manipulated and authentic videos. In this paper, we address the following questions: How accurately do individuals detect deepfakes? Is there a "wisdom of the crowds" [219, 590] effect when averaging participants' responses for each video? How does individual performance compare with the wisdom of the crowds, and how do these performances compare to the leading model's performance? Does access to the model's predictions and certainty levels help or hinder participants' discernment? And, what explains variation in human and machine performance; specifically, what is the role of video-level characteristics, can emotional priming influence participants' performance at detecting deepfakes, and does specialized processing of faces play a role in human and machine deepfake detection?

Crowdsourcing and averaging individuals' responses are promising and practical solutions for handling the scale of misinformation that would be otherwise overwhelming for an individual expert. Recent empirical research finds that averaged responses of ordinary people are on-par with third-party fact checkers for both factual claims in articles [18] and overall accuracy of content from URL domain names [185, 493]. In order to comprehensively compare humans to the leading AI model and evaluate collective intelligence against its artificial counterpart, we need to conduct two comparisons: How often do individuals outperform the model and how often does the the crowd wisdom outperform the model's prediction?

While a machine will consistently predict the same result for the same input, human judgment depends on a range of factors including emotion. Recent research in social psychology suggests that negative emotions can reduce gullibility [81, 211], which could perhaps improve individual's discernment of videos. In particular, anger has been shown to reduce depth of thought by promoting reliance on stereotypes and previously held beliefs [121]. Moreover, priming people with emotion has been demonstrated to both increase and decrease people's gullibility depending on the category of emotion [210] and hinder people's ability to discern real from fake news [404]. The role of emotion in deepfake detection is of practical concern because people share misinformation, especially political misinformation, because of its novelty and emotional content [624]. While a detailed examination of emotions as potential mechanisms to explain deepfake detection performance is outside the scope of this paper, we have included a pre-registered randomized experiment to evaluate whether experimentally elicited anger impairs participants' performance in detecting deepfakes.

Based on research demonstrating human's expert visual processing of faces, we may expect humans to perform quite well at identifying the synthetic face manipulations in deepfake videos. Research in visual neuroscience and perceptual psychology has shown that the human visual system is equipped with mechanisms dedicated for face perception [574]. For example, there is a region of the brain specialized for processing faces [312]. Human infants show sensitivity to faces even before being exposed to them [236, 526], and adults are less accurate at recognizing faces when images are inverted or contain misaligned parts [529, 531, 659]. The human visual system is faster and more efficient at locating human faces than other objects including objects with illusory faces [325]. Whether human visual recognition of faces is an innate ability or a learned expertise through experience, visual processing of faces appears to proceed holistically for the vast majority of people [532, 660]. In order to examine specialized processing of faces as a potential mechanism explaining deepfake detection performance, we include a randomized experiment where we obstruct specialized face processing by inverting, misaligning, and occluding videos.

In order to answer questions about human and machine performance at deepfake detection, we designed and developed a website called Detect Fakes where anyone on the internet could view deepfake videos sampled from the DFDC and see for themselves how difficult (or easy) it is to discern deepfakes from real videos. On this website, we conducted two randomized experiments to evaluate participants' ability to discern real videos from deepfakes and examine cognitive mechanisms explaining human and machine performance at detecting fake videos. We present a screenshot of the user interface of these two experiments in Figure S4 in the Supporting Information. In the first experiment, we present a two-alternative forced choice design where a deepfake video is presented alongside its corresponding real video. In the second experiment, we presented participants with a single video design and asked them to share how confident (from 50% to 100% in 1 percentage point increments) they are that the video is a deepfake (or is not a deepfake). In this single video framework, we present participants with the option to update their confidence after seeing the model's predicted likelihood that a video is a deepfake. By doing so, we evaluate how machine predictions affect human decision-making. In both experiments, we embedded randomized interventions to evaluate whether incidental emotion (emotion unrelated to the task at hand) or obstruction of specialized processing of faces influence participants' performance.

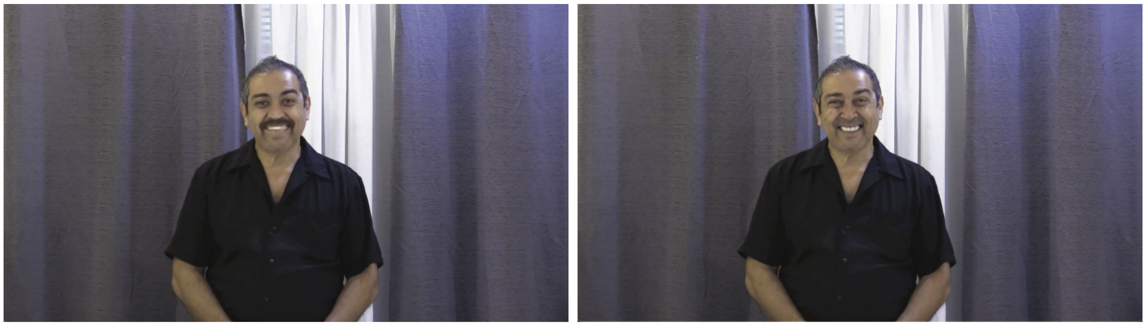


Figure 2-1: One of these two images is the first frame of a deepfake from Experiment 1; the other is the first frame of the original, authentic video from which the deepfake was created. Experiment 1 asked whether participants can tell which is which, using a two alternative forced-choice paradigm (i.e., selecting which of two video clips is a deepfake). Experiment 2 presented a single video and asked participants for their confidence the video is a deepfake or not. In this figure, the left panel is the deepfake; the man was not mustachioed at the time of filming.

2.2 Results

2.2.1 Experiment 1: Two-Alternative Forced Choice (N=5,524)

In Experiment 1, 5,524 individuals found our website organically and participated in 26,820 trials. The 56 pairs of videos in Experiment 1 were sampled from the DFDC training dataset because the experiment was conducted before the holdout videos for the DFDC dataset were publicly released. As such, we compare participants' performance in Experiment 1 to the overall performance of the leading model. We leave a direct comparison of participant and model performance for Experiment 2 which focuses on performance across holdout videos.

Individual vs. Machine

As stated in our pre-analysis plan[2] for Experiment 1, we examined the accuracy of all participants who saw at least 10 pairs of videos, for a total of 882 participants. 82% of participants outperform the leading model, which achieves 65% accuracy on the holdout dataset [167]. Half of the stimuli set (28 of 56 pairs of videos) was identified correctly by over 83% of participants, 16 pairs of videos were identified correctly by between 65% and 83% of participants, and 12 pairs of videos were identified correctly by less than 65% of participants. Out of these 12 pairs of videos, 3 pairs of videos were identified correctly by less than 50% of participants. Figure 2-2a presents the distribution of participants' performance in Experiment 1 (in blue in the second column) next to the model's overall performance (in black in the first column).

We do not find any evidence that participants improve in their ability to detect these videos within the first ten videos seen ($p = 0.112$) (all p-values reported in this paper are generated by linear regression with robust standard errors clustered on participants unless otherwise stated). On average, participants took 42 seconds to respond to each pair of videos, and we find that for every additional ten seconds participants take to respond, participants' accuracy decreased by 1.1 percentage points ($p < 0.001$). We embedded three randomized experiments in Experiment 1 to evaluate the roles of specialized processing of faces, time for reflection, and emotion elicitation. We find participants are 5.6 percentage points less accurate ($p = 0.004$) at detecting pairs of inverted videos than pairs of upright videos. In contrast, we do not find statistically significant effects of the additional time for reflection intervention or this particular emotion elicitation intervention. The custom emotion elicitation intervention in this first experiment did not have a statistically significant influence on participants' self-reported emotions, which suggests the custom emotion elicitation experiment did not work here. We provide additional details on the interventions in Experiment 1 in the Supplementary Information section.

[2] Pre-analysis plan for non-recruited participants in Experiment 1: https://aspredicted.org/blind.php?x=wg84ic

2.2.2 Experiment 2: Single Video Design (N=9,492)

In Experiment 2, 9,492 individuals participated: 304 individuals were recruited from Prolific and completed 6,390 trials; 9,188 individuals found our website organically and completed 67,647 trials.[3] In the recruited cohort, all but 3 participants viewed 20 videos. In the non-recruited cohort, over half of participants viewed 7 videos and the 90th percentile participant viewed 17 videos. The website instructed participants about videos that "Half are deepfakes, half are not." After viewing each video, participants move a slider to report their response ranging from "100% confidence this is NOT a DeepFake" to "100% confidence this is a DeepFake" in one percent increments with "just as likely a DeepFake as not" in the middle (see Figure S4 in the Supporting Information for a screenshot of the user interface). Participants can never make a selection with less than 50% confidence; the slider's default position is in the center (at the "just as likely a DeepFake as not" position and one increment to the right becomes "51% confidence this is a DeepFake" and one increment to the left becomes "51% confidence this is NOT a DeepFake." The stimuli in Experiment 2 consist of 50 videos randomly sampled from the competition holdout dataset (half deepfake and half non-manipulated), 4 videos of Kim Jung-un and Vladimir Putin including both one deepfake and non-manipulated video of each leader, and a deepfake attention check video.

In Experiment 2, we define the accuracy score as the participant's response between 0 and 1 normalized for correctness, which is the participant's response if correct or 1 minus the participant's response if incorrect. For example, if a participant responded "82% confidence this is a DeepFake" and the participant is correct, then the participant is assigned an accuracy score of 0.82. If the participant is incorrect, then the participant would be assigned an accuracy score of 0.18. We define accurate identification as an accuracy score greater than 0.5.

Participants' and the leading model's performance on deepfake detection depends on the population of participants, the population of videos, how performance is measured at the individual or collective level, and whether videos are presented side by side or by themselves.

[3] Pre-analysis plan for recruited individuals participating in Experiment 2: https://aspredicted.org/blind.php?x=mp6yg9

In some cases, we find a machine advantage and in others, we find a human advantage. The rest of the results section examines individual participant performance compared with the leading model, participants' collective performance compared with the leading model, participants' collective performance when participants have access to the model's predictions, variations in human and machine performance across videos, and randomized experiments designed to evaluate the role of emotional priming and specialized visual processing of faces.

Individual vs. Machine

For participants who pass the attention check, recruited participants accurately identified deepfakes from the randomly sampled holdout videos in 66% of attempts while the non-recruited participants accurately identified videos in 69% of trials (or 72% of attempts when limiting the analysis to non-recruited participants who saw at least 10 videos). In comparison, the leading model accurately identified deepfakes on 80% of the sampled videos, which is significantly better than the 65% accuracy rate this model achieves on the full holdout dataset of 4,000 videos [167].

In a direct comparison of performance, 13% of recruited participants, 27% of non-recruited participants who saw at least 10 videos, and 37% of non-recruited participants who saw fewer than 10 videos outperform the model. Figure 2-2a presents the distribution of participants' accuracy on the sampled holdout videos (in teal for recruited participants and gold for non-recruited participants). Relative to the leading model, participants are less accurate at identifying deepfakes than they are at identifying real videos. Recruited participants accurately identify deepfakes as deepfakes in 57% of attempts compared to the leading model identifying deepfakes as deepfakes in 84% of videos while both recruited participants and the leading model identify real videos as real videos at nearly same rate (75% of participants' observations and 76% of videos). Recruited participants predicted the sampled holdout videos were real (57% of observation) considerably more often than fake (38% of observations) while the computer vision model predicted videos were real (44% of observations) barely more frequently than fake (42% of observations). In 5% of recruited participant observations and 14% of computer vision model observations, the prediction was a 50-50 split between real and fake. We report confusion matrices for each treatment condition in Tables 3-7 in the Supplemental Information.

On the additional set of videos of political leaders, participants outperform the leading model. Specifically, 60% of recruited participants and 68% of non-recruited participants who saw at least ten videos outperform the model on these videos. For the deepfake videos of Kim Jong-un, Vladimir Putin, and the attention check, the state-of-the-art computer vision model outputs a 2%, 8%, and 1% probability score that each respective video is a deepfake, which is both confident and inaccurate.

We do not find any evidence that participants' overall accuracy changes as participants view more videos ($p = 0.433$). However, we find that for every additional video seen by recruited participants, they are 0.9% ($p < 0.001$) more likely to report any video as a deepfake. This corresponds to performing about 18% better at detecting deepfakes and 18% worse at identifying real videos by the last video.

Recruited participants spent a median duration of 22 seconds before submitting their initial guess and a median duration of 3 seconds adjusting (or not adjusting) their initial guess when prompted with the model's predicted likelihood. Non-recruited participants spend a similar amount of time. For both sets of participants, we find that for every ten additional seconds of participant response time, participants' accuracy decreases by 1 percentage point ($p < 0.001$).

Crowd Wisdom vs. Machine

The crowd mean, participants' responses averaged per video, is on par with the leading model performance on the sampled holdout videos. For recruited participants, the crowd mean accurately identifies 74% of videos. For non-recruited participants, the crowd mean accurately identifies 80% of videos. For the 1,879 non-recruited participants who saw at least 10 videos, the crowd mean is 86% accurate. In comparison, the leading model accurately identifies 80% of videos.

In Figure 2-2b, we compare statistics on participants' accuracy (the mean and interquartile range) with the model's predictions for each video. In Table 2 in the Appendix, we present the mean accuracy of recruited and non-recruited participants and the computer vision model for all videos. There are 2 videos (both deepfakes) on which both the crowd mean and the leading model are at or below the 50% threshold. One of these videos (video 7837) is extremely blurry, while the other video (video 4555) is filmed from a low angle and the actress's glasses show significant glare.

There are 8 videos on which the crowd mean is accurate but the model is at or below the 50% threshold and another 5-13 videos on which the model is accurate but the crowd mean (depending on how the population selected) is below the 50% threshold.

Human-AI Collaboration

In addition to comparing individual and collective performance to the leading model's performance, we examined how an AI model could complement human performance. After participants' submitted their initial response for how confident they are that a video is or is not a deepfake in Experiment 2, we revealed the likelihood that the video is a deepfake – as predicted by the leading model – and gave participants a chance to update their response. After taking into account the model's prediction, participants updated their confidence in 24% of trials (and crossing the 50% threshold for accurate identification in 12% of trials). By updating their responses, recruited participant's accurate identification increased from 66% to 73% of observations ($p < 0.001$ based on a Student's t-test). Figure 2-2c presents the distribution of changes in overall participant accuracy for the 50 videos sampled from the DFDC. For the 40 videos upon which the model accurately identifies the video as a deepfake or not, participants updated their responses to be on average 10.4% more accurate at identification than before seeing the model's prediction. For the remaining 10 videos on which the model made an incorrect or equivocal prediction, participants updated their responses to be on average 2.7% less accurate at identification than before seeing the model's predictions. In the most extreme example of incorrect updating, the model predicted a 28% likelihood the video was a deepfake when it was indeed a deepfake and participants updated their responses to be on average 18% less accurate at identifying the deepfake. This particular video (video 7837) is quite blurry, and perhaps, participants changed their responses because it's very difficult to discern manipulations in low quality video.

For the additional deepfake videos of Kim Jung-un and Vladimir Putin that are not included in the overall analysis, the model predicted a 2% and 8% likelihood the video was a deepfake, respectively. This prediction is not only incorrect but confidently so, which led participants to update their responses such that participant's accurate identification dropped from 56% to 34% on the Kim Jung-un deepfake and 70% to 55% on the Vladimir Putin deepfake.

In Figure 2-2d, the receiver operating characteristic (ROC) curve of the leading model is plotted alongside the ROC curves of the crowd mean and crowd-mean responses where participants have access to the model's prediction for each video. While the model has a slightly higher AUC score of 0.957 relative to the crowd means's AUC score of 0.936, either the model or the crowd mean could be considered to perform better depending on the acceptable false positive rate. However, the crowd mean response after seeing the model's predictions strictly outperforms both the performance of the crowd mean and leading model. When we condition the ROC analysis on confidence following methods for estimating the reliability of eyewitness identifications [646], we find that medium and high confidence responses outperform low confidence responses by a large degree. We define low confidence as responses between 33.5 and 66.5, medium confidence as responses between 17 and 33.5 or 66.5 and 83, and high confidence as responses between 0 and 17 or 83 and 100. Figure S2 in the Supporting Information ROC curves presents a visual comparison of model performance to low, medium, and high confidence responses from participants, which reveals medium and high (but not low) confidence responses can outperform the model's predictions depending on the acceptable false positive rate.

Video Features Correlated with Accuracy

Given the heterogeneity in both participants' and the leading model's performance on videos, we extend the analysis of performance across seven video-level features: graininess, blurriness, darkness, presence of a flickering face, presence of two people, presence of a floating distraction, and the presence of an individual with dark skin. These video-level characteristics were hand-labeled by the research team. On the 14 videos that are either grainy, blurry, or dark, the crowd wisdom of recruited participants is correct on 8 videos while the crowd wisdom of non-recruited participants and the model is correct on 10 videos. When we examine the 36 videos that are neither grainy, blurry, nor extremely dark, the crowd wisdom of recruited participants is correct on 29 out of 36 videos, the crowd wisdom of non-recruited participants is correct on 32 out of 36 videos, and the model is correct on 30 of 36 videos. The presence of a flickering face is associated with an increase in recruited participants' accuracy rates by 24.2 percentage points ($p < 0.001$) and an increase in the model's accuracy rates by 21.7 percentage points ($p = 0.170$) in detecting a deepfake. The presence of two people in a video instead of a single person is associated with an overall increase in recruited participants' accuracy rates by 7.6% ($p < 0.001$) and a 21.9% decrease ($p = 0.023$) by the model in identifying real videos. The presence of a floating distraction is associated with a decrease in recruited participants' accuracy rates on real videos of 3.5% ($p = 0.034$) and an increase in recruited participants' accuracy rates on fake videos of 11.3% ($p < 0.001$). In 12 of 50 videos, at least one person in the video has dark skin (precisely defined as skin classified as type 5 or 6 on the Fitzpatrick scoring system, which is a classification system developed for dermatology that computer vision researchers have used to examine the context of skin color) [87]. We find that the presence of an individual with dark skin in the video is associated with a decrease in recruited participants' accuracy by 8.8% ($p < 0.001$) and a decrease in the model's accuracy by 12.0% ($p = 0.192$). In order to control for these seven comparisons conducted simultaneously, we can apply a Bonferroni correction of 1/7 to the standard statistical significance thresholds (e.g., a p-value threshold of 0.01 becomes 0.0014). Based on this correction, the influence of a flickering face, two people in the same video, floating distractions, and the presence of an individual with dark skin continue to be statistically significant for participants if the original p-value threshold is chosen as 0.01.

Randomized Experiments for Evaluating Emotion Priming and Specialized Face Processing

Within Experiment 2, we embedded two randomized experiments to examine potential cognitive mechanisms underpinning how humans discern between real and fake videos. Specifically, we examine an affective intervention designed to elicit anger based on a well-established intervention [576] (see Figure S3 in the Supporting Information) and a perceptual intervention designed to obstruct specialized processing of faces via inversion (videos presented upside down), misalignment (videos presented with actors' faces horizontally split), and occlusion (videos presented with a black bar over the actors' eyes).

We present results of the anger elicitation intervention in columns 1-3 in Table 2.1. We do not find statistically significant effects ($p = 0.280$) of the anger elicitation intervention on overall accuracy. However, in our pre-registered follow-up analysis limiting the dataset to real videos, we find that participants who were assigned to the anger elicitation treatment underperformed control participants by 5.2 percentage points ($p = 0.032$). In other words, participants in the anger elicitation treatment are more 5.2 percentage points more likely than participants in the control group to make a false positive identification that a real video is a deepfake. Notice here that the floor is not 0% accuracy but rather 50% accuracy (i.e., chance responding); the maximum effect of the anger elicitation treatment is 21.6 percentage points (71.6 from the constant term in column 2 of Table 2.1 minus 50), so a 5.2 percentage point reduction represents an effect that is 24.1% of the maximum possible effects under these conditions.

Figure S3 in the Supporting Information presents accuracy and confidence scores by treatment assignment to visually reveal the heterogeneous effect of anger elicitation on how participants discern between real and fake videos. When we examine the relationship between assignment to the anger elicitation group and how confident participants guess, we do not find a statistically significant effect ($p = 0.347$). When we examine real videos and the relationship between anger elicitation and updating predictions after seeing what the model would predict, we find that participants assigned to the anger elicitation group are 3.7% ($p = 0.100$) more likely to change their guess to a correct answer than participants assigned to the control group. As a result, we do not find statistically significant effects of anger elicitation on accuracy after participants update their response ($p = 0.246$).

We present results of the perceptual obstruction intervention in columns 1-6 in Table 2.1. We find statistically significant effects of all three specialized processing obstructions on participants' ability to accurately identify deepfakes from authentic videos. The overall effects – reported in column 1 of Table 2.1 – are all statistically significant and range from a decrease of 4.3 percentage points in accuracy for the inversion treatment ($p = 0.002$) to a decrease of 4.4 percentage points in accuracy for the eye occlusion treatment ($p = 0.004$) to a decrease of 6.3 percentage points for the misalignment treatment ($p < 0.001$) on a base rate of 65.5% accuracy when controlling for video fixed effects. In addition, we find that inverting the videos decreases participants' reported confidence scores (absolute distance in guesses from the 50-50 selection) by 2 percentage points ($p = 0.002$), but we do not find similar decreases in reported confidence on videos with misalignment or occlusion transformations.

In the sample of recruited participants, the specialized face processing obstructions have different effects depending on whether the videos are manipulated or not. When we limit the analysis to the algorithmically manipulated deepfakes (column 3 of Table 2.1), we do not find statistically significant effects on the inversion treatment ($p = 0.638$) but we do find that the misalignment and eye occlusion treatments show a decrease in 7.7 ($p = 0.002$) and 6.3 ($p = 0.008$) percentage points, respectively, relative to the control videos. In contrast, when we limit the analysis to the other half of videos that have not been manipulated (column 2 of Table 2.1), we do not find statistically significant effects for misalignment ($p = 0.075$) or eye occlusion interventions ($p = 0.263$), but we find participants' accuracy on inverted authentic videos is 9.1 percentage points lower than when the videos are upright ($p < 0.001$).

The experimental results on non-recruited participants provide a replication and robustness check for the results on the recruited participants. The results from the non-recruited participants were not pre-registered because we weren't expecting many people to continue visiting our website organically. In fact, 9,188 visitors participated in the single video design between November 2020 and January 2021. In columns 4 through 6 in Table 2.1, we present the linear regressions results of the specialized face processing obstruction interventions on non-recruited participants' accuracy. Similar to the results for the recruited sample, we find statistically significant effects ($p < 0.001$) of all three obstruction interventions on ability to accurately discern deepfakes from authentic videos. The number of observations in the non-recruited sample is over 16 times larger than the number of observations in the recruited sample. Likewise, the number of participants is 33 times larger. These numbers differ because the number of videos seen by participants in the non-recruited sample varied depending on participants' interest. With a larger sample size, we see statistically significant and negative effects of obstructions on all videos ranging from 4% percentage point drop in accuracy from the eye occlusion intervention ($p < 0.001$) to a 7% drop on accuracy from the misalignment intervention ($p < 0.001$). We also find all three treatments decrease participants' confidence scores by one half to one percentage point ($p < 0.001$).

In the non-recruited sample, each of the 50 videos were viewed by between 945 to 1168 participants. We run separate linear regressions for each video and find statistically significant and negative effects at the 1% significance level for inversion in 24 videos, misalignment in 15 videos, and occlusion in 20 videos. Furthermore, we find at least one of these specialized processing obstructions is negative and statistically significant at the 1% significance for 29 of the 50 videos.

In the sample of videos of political leaders, the specialized face processing obstructions had a significant effect on participants' ability to accurately identify the Vladimir Putin deepfake as a deepfake. The misalignment obstruction leads to a drop in accuracy of 20.7 percentage points ($p = 0.001$). Likewise, the occlusion obstruction leads to a drop of 10.3 percentage points ($p = 0.002$) and the inversion obstruction leads to a drop of 5.2 percentage points ($p = 0.072$).

In columns 7 through 9 in Table 2.1, we present the linear regression results of the specialized face processing obstruction interventions on the model's predictions and find the computer vision model is affected by one specialized face processing obstruction but not the other two. We find the computer vision model's predictive accuracy drops by 12.1 percentage points on the inverted videos ($p = 0.005$). We do not find a statistically significant difference in accuracy between either the control and misalignment sets of videos ($p = 0.800$) or the control and occlusion sets of videos ($p = 0.944$).

	Dependent variable: Accuracy								
	Recruited			Non-recruited			Computer		
	All	Real	Fake	All	Real	Fake	All	Real	Fake
Constant	0.655***	0.716***	0.567***	0.679***	0.700***	0.632***	0.813***	0.786***	0.841***
	(0.009)	(0.014)	(0.015)	(0.002)	(0.003)	(0.003)	(0.030)	(0.040)	(0.044)
Inversion	-0.043***	-0.091***	0.010	-0.053***	-0.080***	-0.027***	-0.121***	-0.110*	-0.132**
	(0.014)	(0.021)	(0.021)	(0.004)	(0.006)	(0.006)	(0.042)	(0.056)	(0.063)
Misalignment	-0.061***	-0.042*	-0.077***	-0.070***	-0.056***	-0.084***	0.011	0.000	0.021
	(0.016)	(0.024)	(0.025)	(0.005)	(0.007)	(0.007)	(0.042)	(0.056)	(0.063)
Eye Occlusion	-0.044***	-0.023	-0.063***	-0.040***	-0.035***	-0.043***	-0.003	-0.007	0.001
	(0.015)	(0.021)	(0.024)	(0.004)	(0.006)	(0.006)	(0.042)	(0.056)	(0.063)
Anger	-0.020	-0.052**	0.012						
	(0.014)	(0.024)	(0.021)						
Number of Participants	229	229	229	7563	6368	6670	0	0	0
Number of Guesses (Real)	2349	1514	835	27446	18524	8922	81	76	5
Number of Guesses (Deepfake)	1707	549	1158	22766	6316	16450	87	7	80
Number of Guesses (50-50)	180	68	112	3713	1726	1987	32	17	15
Number of Unique Videos	50	25	25	50	25	25	50	25	25
Observations	4,236	2,131	2,105	53,925	26,566	27,359	200	100	100
R^2	0.180	0.069	0.225	0.185	0.057	0.273	0.062	0.054	0.073
Adjusted R^2	0.170	0.056	0.215	0.184	0.057	0.272	0.048	0.025	0.044
Residual Std. Error	0.340	0.329	0.350	0.349	0.350	0.346	0.210	0.198	0.222
F Statistic	288.686***	164.804***	169.388***	3687.143***	2150.874***	4525.903***	4.337***	1.841	2.514*

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2.1: Treatment effects of interventions on accuracy. Linear regressions on participant data includes video fixed effects with Eicker-Huber-White standard errors clustered at the participant level.

2.3 Discussion

How do ordinary human observers compare with the leading deepfake detection models? Our results are at odds with the commonly held view in media forensics that ordinary people have extremely limited ability to detect media manipulations. Past work in the cognitive science of media forensics has demonstrated that people are not good at perceiving and reasoning about shadow, reflection, and other physical implausibility cues [191, 318, 460, 461]. On first glance, deepfakes and other algorithmically generated images of people (e.g., images generated by StyleGAN) look quite realistic [314]. But, we show that deepfake algorithms generate artifacts that are perceptible to ordinary people, which may be partially explained by human's specialized visual processing of faces. In contrast to recent research showing ordinary people quickly learn to detect AI generated absences in photos [244], we do not find evidence that participants improve in their ability to detect deepfakes.

By showing participants videos of unknown individuals making uncontroversial statements, we focused the truth discernment task specifically on visual perception. The lack of additional context creates a level playing field for a reasonably fair comparison of human and machine vision because humans cannot also reason about contextual, conceptual clues in these videos [206]. In the two alternative forced-choice paradigm of Experiment 1, 82% of participants respond with higher accuracy than the leading model. In the more challenging single video framework in Experiment 2, participants still perform really well, and we find that between 13% and 37% of ordinary people outperform the leading deepfake detection model. When we aggregate participants' responses in Experiment 2, we find that collective intelligence, as measured by the crowd mean, is just as accurate as the model's prediction.

In the extension of the experiment to videos of well-known political leaders (Vladimir Putin and Kim Jong-un), participants significantly outperform the leading model, which is likely explained by participants' ability to go beyond visual perception of faces. Unlike the 50 sample holdout videos, participants could critically contemplate the authenticity of the video of the political leader. For example, participants might have considered whether Vladimir Putin or Kim Jong-un speak English, whether they actually sound like they do in the video, and whether such a well-known political figure would say such a thing. Not only do the majority of participants identify the deepfake status of videos of political leaders correctly, but the computer vision model is confident in its wrong predictions. Perhaps, the model failed because it was trained on face swapping deepfake manipulations as opposed to synthetic lip syncing manipulations. What the evidence shows is that today's leading model does not generalize well to stylistically different videos than the videos on which it has been trained, whereas human deepfake detection abilities do generalize across these different contexts.

The model's predictions helped participants improve their accuracy overall, but whether a participant's accuracy increased depended on whether the model accurately identified the video as a deepfake or not. Participants often made significant adjustments based on the model's predictions, and inaccurate or equivocal model predictions led participants astray in 8 of 10 instances. Moreover, the model's incorrect assessment of the political leader deepfake videos is associated with a decrease in participant accuracy, which is in line with recent empirical research that shows deepfake warnings do not improve discernment of political videos [601]. Likewise, these results mirror other recent research revealing human-AI collaborative decision making does not necessarily lead to more accurate results than either humans or AI alone [5, 221, 299, 608, 615].

Videos are heterogeneous, high-dimensional media, and as a result, participants were accurate on some videos on which the leading model failed and vice versa. In line with recent research examining perceptual differences between authentic and deepfake videos [647], we identified 7 salient dimensions across the 50 sampled holdout videos to evaluate differences in how participants and the leading model discern authenticity: We find that the leading model performs slightly better than participants on low-quality videos that were categorized as grainy, blurry, and very dark. This differential performance suggests that the model is picking up on low-level details that participants appear to ignore. On the other hand, we find both recruited and non-recruited participants attain similar accuracies as the model on standard quality videos. Both participants and the model are quite adept at picking up on flickering faces. The model has trouble discerning between real and deepfake videos when two actors appear in the video while participants have no trouble in this context. This suggests that the model may be vulnerable to changes in context whereas participants are more robust to varying context. With respect to visual distractions, we find distractions are associated with participants identifying videos more often as deepfakes. While we showed recruited participants examples of distraction videos that should not be reported as deepfakes and we explicitly described these distractions in the instructions as not necessarily characteristic of deepfakes, we imagine the results concerning the distraction videos may possibly reflect confusion by the participants. Nonetheless, all reported results are robust to the exclusion of distraction videos. In light of recent research showing intersectional disparities in accuracy of commercial facial recognition software [87] and the impact of race on credibility with deepfakes [271], we examine accuracy on the videos with dark skin actors. Participants and the leading model are both less accurate on videos with dark skin actors, but as we reported in the results section, we only find a statistically significant difference in participants' performance not the model's performance.

In Experiment 2, we find some evidence for our pre-registered hypothesis that anger would impair participants' ability to identify manipulated media. When we elicit incidental anger (i.e., anger unrelated to the task at hand), participants' accuracy at identifying real videos decreases, a pattern that held across almost all videos (see Figure S3 in the Supporting Information where participants assigned to the anger elicitation under perform participants assigned to the control in 22 out of 25 real videos, and see the Limitations section below). The negative and heterogeneous effect of incidental anger on the discernment of real (but not fake) videos may be related to the negative and heterogeneous effect of emotion priming on accuracy ratings of fake (but not real) news headlines [404]. Drawing on Martel et al 2020, one potential explanation for the negative effects of anger elicitation on the discernment of authentic but not deepfake videos is emotion leading to an over-reliance on intuition; in this experiment, if a participant sees something that looks like a deepfake manipulation, then she is unlikely to think the video is real, but if a participant does not see something that looks like a deepfake manipulation, then he might think he's simply unable to spot the detailed manipulation and may respond based on his intuition that a video is fake rather than whether he clearly saw a manipulation or not.

Both Experiments 1 and 2 provide support for the claim that specialized processing of faces helps people discern authenticity in visual media. In particular, we show that three visual obstructions designed to hinder specialized processing – inversion, misalignment, and partial occlusion – decrease participants' accuracy. In contrast to human visual processing, we find only inversion and not misalignment or partial occlusion change the model's performance. While the computer vision model is robust to misalignment and occlusion, this robustness may be a bug – the model overfitting to the training data – rather than a feature. Future research should explore whether specialized processing in computer vision models for deepfake detection enables better generalization to new contexts.

2.4 Limitations

We evaluated human and machine performance on 167 videos (84 deepfake and 83 authentic videos) across Experiments 1 and 2. While these videos represent a balanced group of individuals across demographic dimensions and a variety of deepfake models, only the two political deepfake videos include lip syncing manipulations, which are some of the most commonly used models for producing political deepfakes [13, 168, 510, 592]. Moreover, we do not specifically recruit expert fact-checkers or expert media forensic analysts, and as such, our results only generalize to the performance of ordinary people. Our comparison of untrained participants' predictions to the predictions of the leading computer vision model is limited to the best performance in 2020. If current trends continue as we expect they would, computer vision detection models will continue to improve (and possibly incorporate more human-like specialized processing of faces to better generalize across contexts) just as the realism of synthetic media generation algorithms will continue to improve. As a consequence, society will require more than just visual-based classification algorithms to protect against the potentially harmful threats that deepfakes pose [432].

The minimal context videos used here may not resemble the most problematic deepfakes because the videos here show unknown people saying non-controversial things in nondescript settings. On one hand, this minimal context makes the human participants' performance all the more impressive because such videos are missing many of the contextual cues they might normally use to discern authentic videos from deepfakes. On the other hand, perhaps videos designed to deceive are stylistically very different than the videos from the sampled holdout. As such, persuasive, manipulated video is important to consider in future research. The role of persuasion in synthetic media is beginning to be explored across varying media modalities [165, 645], but it is not the central focus of this paper. Instead, we ask how well the human visual processing system can detect the visual manipulations characteristic of deepfakes. We limit the bulk of our evaluation to uncontroversial videos of unknown actors to focus on the visual component of truth discernment. We begin to examine more realistic examples based on four videos of political leaders, but a larger sample size and further experimentation is necessary before making conclusions about how people judge the authenticity of political deepfakes. Furthermore, there is still much to learn about how AI systems and ordinary people can incorporate all the other information beyond facial features to make accurate judgments about a video's authenticity.

In this experiment, half of the videos were real and half deepfakes. This is useful for comparing human and machine performance, but this base rate of deepfakes does not reflect the base rate of misinformation in today's media ecosystems [19]. In 2021, less than a fraction of a percent of news was misinformation [634]. Future experiments might consider examining people's ability to identify deepfakes when they do not have foreknowledge of the base rate of deepfakes. Moreover, an experiment embedded in a social media ecosystem could further identify how well people identify deepfakes within an ecologically valid context where people have access to contextual information such as who shared the video and how many others have shared or commented on the video. Ultimately, there are many ways to discern between real and fake videos, and visual perception should be considered as one tool in a user's toolkit for truth discernment.

We also considered how incidental emotions (i.e., emotions unrelated to the task at hand) affect participants' discernment of real and fake videos. Here, our two experiments found different results, and so we do not draw firm conclusions about the role of emotion on deepfake detection. In Experiment 1, the custom emotion elicitation interventions did not significantly alter deepfake detection performance — though it also did not significantly alter self-reported emotions, making it unclear how much to read into the lack of effects on performance. The results from Experiment 2, though statistically significant by conventional standards, were near the cutoff for statistical significance for authentic videos and not statistically significant for deepfake videos. As such, future research could further explore the role of emotions in deepfake detection by running experiments with larger samples, examining additional emotions, ensuring effective elicitation, and focusing on integral emotions (emotions elicited directly from the stimuli). Recent research shows that inferences from feelings are context-sensitive and incidental emotions may be more likely to lead individuals astray in judgment tasks than integral emotions [559].

2.5 Implications

Relative to today's leading computer vision model, groups of individuals are just as accurate or more accurate depending on which videos are considered. Participants and the model perform equally well on standard resolution, visual-only deepfake manipulations. Participants perform better on the four political videos and attention check video while the computer vision model performs slightly better on blurry, grainy, and very dark videos. The model's poor performance on both deepfakes of world leaders and videos with two people instead of one suggests that the model may not generalize well to stylistically different videos than the videos on which it has been trained. Humans have no problem with this kind of generalization, and as a consequence, social media content moderation of video-based misinformation is likely to be more accurate when performed by teams of people than today's leading algorithm. As such, future research in crowd-based deepfake detection may consider how to most effectively aggregate wisdom of the crowds to improve discernment accuracy beyond the crowd mean (e.g., using algorithms such as the surprisingly popular answer [511] and revealed confidence [665]).

Sociotechnical systems may benefit from the combination of artificial intelligence and crowd-wisdom, but decision support tools for content moderation must be carefully designed to appropriately weigh human and model predictions. The confidently wrong predictions of the model on out-of-sample videos reveals the leading model is not ready to replace humans in detecting real-world deepfakes. Moreover, decision support tools can be counter-productive to accurate identification as evidenced by the many instances in which participants saw incorrect predictions from the model and subsequently adjusted their predictions to be less accurate. Instead of solely informing people on the likelihood that a model is a deepfake, crowd-wisdom could likely benefit from more explainable AI. Given that the leading model was more accurate at detecting certain classes of videos while humans were better at other classes, a future human-AI collaborative system might include additional information on video sub-types and how humans and machines perform across these sub-types. For example, video-level qualities (e.g., blurry, grainy, dark, specialized obstruction, stylistic similarities to training set, or other components upon which human and machine performance tends to diverge) and individual-level qualities could be factored into the interface and information presented by a human-AI collaborative system. By presenting model predictions alongside this information, it is possible humans could develop a better sense for confronting conflicting model predictions and deciding between second-guessing their own judgments and overriding the model's prediction. Machine-informed crowd-wisdom can be a promising approach to deepfake detection and other classification tasks more generally where human and machine classification performance is heterogeneous on sub-types of the data.

Specialized visual processing of faces helps humans discern between real and deepfake videos. In future instances when humans are tasked with deepfake detection, it is important to consider whether a video has been manipulated in such a way as to reduce specialized processing. Moreover, given the usefulness of specialized processing of faces for humans in detecting deepfakes, it is possible that computer vision models for deepfake detection may find use in incorporating (and/or learning) such specialized processing [193].

Visual cues will continue to be helpful in deepfake detection, but ultimately, identifying authentic video can involve much more than visual processing. When attempting to discern the truth from a lie, people rely on the available context, their knowledge of the world, their ability to critically reason, and their capacity to learn and update their beliefs. Similarly, the future of deepfake detection by both humans and machines should consider not only the perceptual clues but the greater context of a video and whether its message resembles an ordinary lie.

2.6 Methods

This research complies with all relevant ethical regulations and the Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects determined this study to fall under Exempt Category 3 – Benign Behavioral Intervention. This study's exemption identification number is E-2070. All participants are informed that "Detect Fakes is an MIT research project. All guesses will be collected for research purposes. All data for research is collected anonymously. For questions, please contact detectfakes@mit.edu.

If you are under 18 years old, you need consent from your parents to use Deep Fakes." Most participants arrived at the website via organic links on the Internet. For recruited participants, we compensated each individual at a rate of $7.28 an hour and provided bonus payments of 20% to the top 10% of participants. Before beginning the experiment, all recruited participants were also provided a research statement, "The findings of this study are being used to shape science. It is very important that you honestly follow the instructions requested of you on this task, which should take a total of 15 minutes. Check the box below based on your promise:" with two options "I promise to do the tasks with honesty and integrity, trying to do them uninterrupted with focus for the next 15 minutes." or "I cannot promise this at this time." Participants who responded that they could not do this at this time were re-directed to the end of the experiment.

We hosted the experiment on a website called Detect Fakes at `https://detectfakes.media.mit.edu/`. Figure S4 in the Supporting Information presents a screenshot of the user interface for both Experiments 1 and 2. The rest of the methods are described in the Supplementary Information section.

2.7 Data and Code Availability

The datasets and code generated and analyzed during the current study are available in our public Github repository, `https://github.com/mattgroh/cognitive-science-detecting-deepfakes`. All DFDC videos are available at `https://dfdc.ai/` [168] and the 5 non-DFDC videos are available in our public Github repository.

Acknowledgements

Authors would like to acknowledge funding from MIT Media Lab member companies, thank Alicia Guo for excellent research assistance, and thank the following communities for helpful feedback: the Affective Computing lab, the Perception and Mind lab, Human Cooperation

lab, and the moderator and participants at the Human and AI Decision-Making panel at the CODE2020 conference.

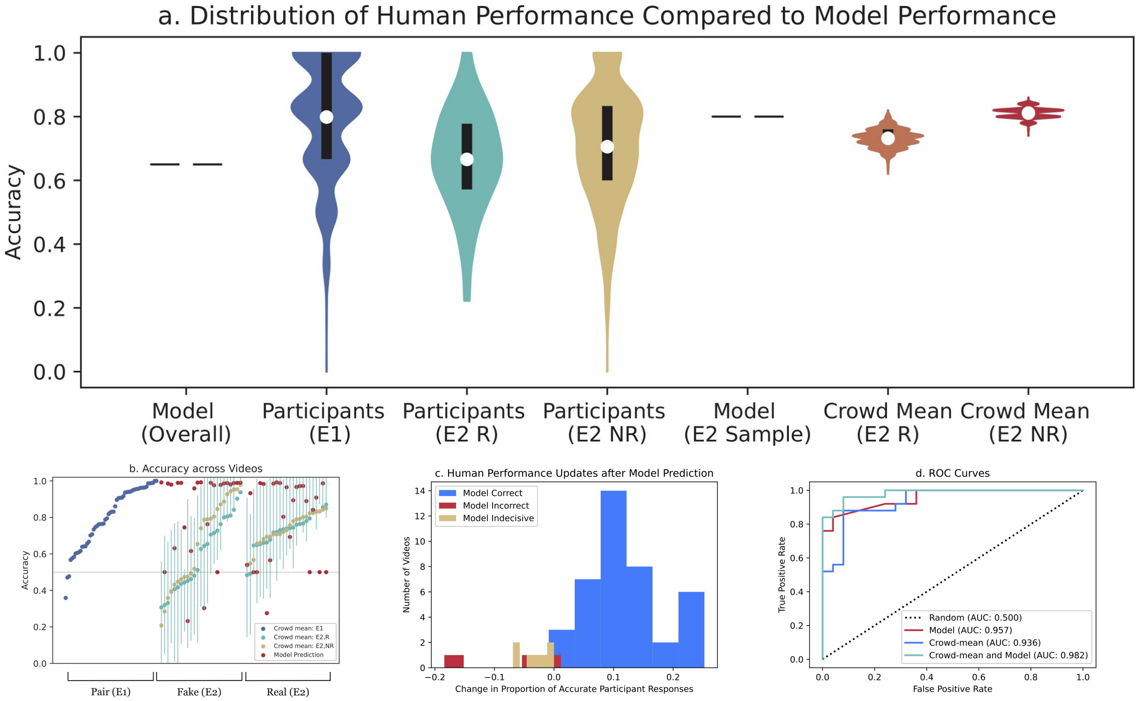


Figure 2-2: Figure 2a presents the distribution of participant performance across experiments compared to the model's performance via violin plots where the white dots indicate the mean and the black bars indicate the interquartile range. R refers to recruited participants, NR refers to non-recruited participants, E1 refers to Experiment 1, and E2 refers to Experiment 2. In the Experiment 1 (two-alternative forced choice), accuracy is defined as identifying a deepfake from a pair of videos correctly. In Experiment 2 (single video design), accurate identification is defined as responding with the correct answer with more than 50% confidence. The model's performance represents a single observation in each instance, and as such, we present the model's performance as a horizontal black line with a white dot in the middle. The crowd mean distributions are obtained by bootstrapping confidence intervals based on 1000 randomly drawn samples that are each half of the total observations. Figure 2b presents a scatter plot of the model's accuracy and the distribution of participants' accuracy scores for each video. The x-axis of Figure 2b is an index of the videos, and it is ordered by experiment, true class of each video, and participant's average accuracy. The teal lines in Figure 2b represent the interquartile range of recruited participants' responses. Figure 2c presents the distribution of changes in recruited participants' accuracy after updating their response based on whether the model's prediction is correct, incorrect, or indecisive. Figure 2d presents the receiver operator characteristic curves of computer performance, recruited participants' collective performance, and recruited participants' collective performance with the model's decision support across the 50 DFDC videos in Experiment 2.

Chapter 3

Human Detection of Political Deepfakes across Transcripts, Audio, and Video

Abstract

Recent advances in technology for hyper-realistic visual effects provoke the concern that deepfake videos of political speeches will soon be visually indistinguishable from authentic video recordings. The conventional wisdom in communications research predicts people will fall for fake news more often when the same version of a story is presented as a video rather than text. Here, we evaluate how accurately 41,822 participants distinguish real political speeches from fabrications in an experiment where speeches are randomized to appear as permutations of text, audio, and video. We find access to audio and visual communication modalities improve participants' accuracy. Here, human judgment relies more on how something is said, the audio-visual cues, than what is said, the speech content. However, we find that reflective reasoning moderates the degree to which participants consider visual information: low performance on the Cognitive Reflection Test is associated with an over-reliance on what is said.[1]

[1] This chapter, which is co-authored by Aruna Sankaranarayanan, Andrew Lippman, and Rosalind Picard, is currently under review and available as a pre-print [247].

3.1 Motivation

Recent advances in technology for algorithmically applying hyper-realistic manipulations to video are simultaneously enabling new forms of interpersonal communication and posing a threat to traditional standards of evidence and trust in media [13, 120, 250, 264, 371, 484, 486]. In the last few years, computer scientists have trained machine learning models to generate photorealistic images of people who do not exist [314, 315, 458], inpaint people out of images [244, 591], clone voices based on a few samples [31, 396], modulate the lip movements of people in videos to make them appear to say something they have not said [358, 510], and create fake videos based on simple text prompts [281]. The synthetic videos' false appearance of indexicality – the presence of a direct relationship between the photographed scene and reality [425, 489] – has the potential to lead people to believe video-based messages that they otherwise would not have believed if the messages were communicated via text. This potential influence is particularly concerning because research demonstrates that videos, especially videos of an injustice, elicit more engagement and emotional reactions (e.g., anger, sympathy) than text descriptions displaying the same information [27, 230, 653] (although, see ref. [509]). Moreover, visual misinformation is common on social media [220] and the emotional and motivational influences of visual communication have been attributed to why fake, viral videos have provoked mob-violence [231, 588]. While people are more likely to believe a real event occurred after watching a video of the event than reading a description of the event [645], an open question remains: Does visual communication relative to text increase the believability of *fabricated* events?

The realism heuristic [587, 588] predicts "people are more likely to trust audiovisual modality [relative to text] because its content has a higher resemblance to the real world." This prediction is relevant for many deepfake videos [265] and suggests fabricated video would be more believable than fabricated text conditional on the absence of obvious perceptual distortions. Yet there exists little direct empirical evidence for this heuristic applied to algorithmically manipulated video. In an experiment using three fake videos as stimuli, researchers found that stories presented as videos are perceived as more credible than stories presented as text or read aloud in audio form [588]. In contrast, in an experiment showing 6 political deepfake videos (videos manipulated by artificial intelligence to make someone say something they did not say) and 9 non-manipulated videos, researchers did not find differences between truth discernment rates in video, audio, and text [46]. Perhaps some of the experiments' participants did not take the videos' "indexicality" as evidence of authenticity because participants were aware of how easily such videos could be manipulated. Alternatively, some participants may have noticed perceptual distortions in the videos, which would naturally lead one to believe the video has been manipulated. The mixed evidence on how communication modalities mediate people's ability to discern fabricated content may be due to the small samples of stimuli in media effects research [524]. In related work on how fake images can be persuasive and difficult to distinguish from real images: research finds people rarely question the authenticity of images even when primed [318], images can increase the credibility of disinformation [258], and images of synthetic faces produced by StyleGAN2 [315] are indistinguishable from the original photos on which the StyleGAN2 algorithm was trained [459]. Moreover, research shows that non-probative and uninformative photos can lead people to believe false claims [103], lead people to believe they know more than they actually know [104], promote "truthiness" by creating illusory truth effects [456, 457], which can lead people to believe falsehoods they previously knew to be falsehoods [180, 194]. When it comes to ostensibly probative videos of political speeches, the question whether people are more likely to believe an event occurred because they saw it as opposed to only read about it remains open.

In fact, today's algorithmically generated deepfakes are not yet consistently indistinguishable from real videos. On a sample of 166 videos from the largest publicly available dataset of deepfake videos to date [168], people are significantly better than chance but far from perfect at discerning whether an unknown actor's face has been visually manipulated by a deepfake algorithm [243]. This finding is significant because it demonstrates that people can identify deepfake videos from real videos based solely on visual cues. However, some videos are more difficult than others to distinguish due to blurry, dark, or grainy visual features. On a subset of 11 of the 166 videos, researchers do not find that people can detect deepfakes better than chance [340]. In another experiment with 25 deepfake videos and 4 real videos but only 94 participants, researchers found that the overall discernment accuracy is 51% and a media literacy training increases discernment accuracy by 24 percentage points for participants assigned to the training relative to the control group [595]. In experiments examining how people react to deepfake videos of politicians, researchers find people are more likely to feel uncertain than misled after viewing a deepfake of Barack Obama [614] and people consider a deepfake of a Dutch politician significantly less credible than the real video from which it was adapted [165] and the deepfake video is not more persuasive than the text alone [259]. In the experiment examining the fabricated video of a Dutch politician, some respondents explained their credibility judgements by indicating audio-visual cues of how the message was communicated (e.g., unnatural mouth movements); others indicated inconsistency in the content of the message itself (e.g., contextually unrealistic speeches) [165].

People's capacity to identify multimedia manipulations raises questions: how do various kinds of fabricated media (e.g., audio and video of fake political speeches) alter the perceived credibility of misinformation, how do audience characteristics (e.g., reflective reasoning) moderate media effects, and how does the source and content of a message interact with the fabricated media and audience characteristics [370]? A growing field of misinformation science is beginning to address these questions. Research on news source quality demonstrates that people in the United States are generally accurate at identifying high and low-quality publishers [493] and the salience of source information does not appear to change how accurately people identify fabricated news stories [41], manipulated images [563], or fake news headlines [163, 302] although evidence on fake news headlines is mixed [330, 451]. Research on political fake news content suggests an individual's tendency to rely on intuition instead of analytic thinking is a stronger factor than motivated reasoning in explaining why people fall for fake news [496], and similarly, people with more analytic cognitive styles worldwide are more accurate at discerning between authentic and fabricated political videos [26] and true and false headlines related to COVID-19 [29]. In fact, people tend to be better at discerning truth from falsehood when evaluating news headlines that are concordant with their political partisanship relative to when evaluating news headlines that are discordant [495]. While the science of fake news has generally focused on the messengers (the source credibility of publishers) [365] and the message of what is said (the media credibility of written articles and headlines) [495], the relevance of audio-visual communication channels to the psychology of misinformation has received less attention [140] and is important for addressing the problem of misinformation [97].

In this paper, we evaluate discernment across 32 political speeches by two well-known politicians. We present these speeches to participants via the 7 possible permutations of 3 digital media communication modalities: text, audio, and video. Based on 46,713 responses from 3,317 individuals who participated in a pre-registered (and an additional 387,274 responses from 38,510 participants who participated after the pre-registration window) [2] cross-randomized experiment, we examine ordinary people's performance at discerning political speeches randomized to appear in each of the following seven conditions: a transcript, an audio clip, a silent video, audio with subtitles, silent video with subtitles, video with audio, and video with audio and subtitles. By randomly assigning political speeches to these permutations of text, audio, and video modalities and asking participants to discern truth from falsehood, this experiment is designed to disentangle the degree to which participants attend to and consider the content of what is said and the audio-visual cues as to how it is said. In addition, we evaluate these disentangled components across message types (speeches that are either concordant or discordant with the general public's perception of a speaker's political identity) and audience characteristics (reflective reasoning as measured by the Cognitive Reflection Test (CRT) [214]), which is a robust test for measuring an individual's tendency towards reflecting on questions before answering [66, 492, 583] that helps explain why people fall for fakes news [494, 496] and is strongly associated with the reliability of news sources people engage with and share on social media [441].

[2] The pre-registered analysis is available at `https://aspredicted.org/VFZ_6HK`. We continued collecting responses from participants who found our experiment organically after the pre-registered cut-off date for data collection passed, and our final sample includes 432,987 responses from 41,822 participants who passed the attention check. Our results are robust to include or exclusion of participants who participate after the pre-registered cut-off date.

3.2 Results

3.2.1 Participants (N=41,822)

A total of 73,236 individuals participated in the experiment. We used the Prolific platform [480] to recruit 554 individuals from the United States who completed 16,699 trials. In addition to the recruited participants, 5,106 individuals (76% of whom visited from outside the United States) participated in the experiment during the pre-registration window from March 4, 2021 to June 1, 2021. These participants found the website organically and completed 44,461 trials. Between June 1, 2021 and July 1, 2022, an additional 67,576 individuals (70% of whom visited from outside the United States) completed 566,343 trials. We focus our analysis on 41,822 participants: the 509 of 554 recruited participants and 41,313 of 72,682 non-recruited participants who passed the attention check where we presented an obvious deepfake and explicitly instructed participants to respond that the video is a deepfake with 100% confidence.

The sample of 509 recruited participants is balanced across political identities (in this experiment failure on the attention check does not correlate with political identities [63]); 257 recruited participants self-report as Democrats, and the other 252 recruited participants self-report as Republicans. We do not find demographic differences in recruited participants who passed the attention check. We did not collect data for recruited participants who failed the attention check, but we did collect data for the non-recruited participants who failed the attention check. In the Supplementary Information, we demonstrate that the main results are robust to including participants who failed the attention check and robust to including or excluding participants who participated after June 1, 2021.

Many but not all participants responded to all 32 speeches; 482 recruited participants and 6,374 non-recruited participants viewed all 32 speeches. Before the experiment began, participants in the recruited cohort (but not the non-recruited cohort) responded to a baseline survey that included questions on political preferences, trust in media and politics, and the three questions from the CRT.

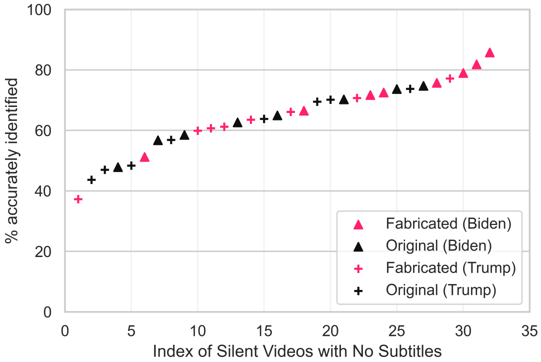


Figure 3-1: Mean identification accuracy across the 32 silent videos (with no subtitles) to illustrate the heterogeneity in how difficult the visual deepfake manipulations are to detect. There are 8 fabricated videos and 10 non-fabricated videos out of the 32 on which participants identify less than 67% of the time. There are 2 fabricated videos accurately identified in more than 80% of observations. The 95% confidence interval range is less than 1% for all silent videos.

3.2.2 Discernment Performance across Communication Modalities

We begin by examining how frequently participants correctly identified the stimuli as fabricated or not. Across all 224 stimuli, recruited and non-recruited participants correctly identified the stimuli in 75% and 70% of observations, respectively.

We find the fabricated political speech transcripts and visual deepfake manipulations are difficult for participants to discern. The proportion of people who accurately identified fabrications from authentic text varied by stimuli. Across the 32 text transcripts, the least accurately identified transcript is identified correctly in 26% of trials, the most accurately identified one is identified correctly in 67% of trials, and the median accurately identified one is identified correctly in 43% of trials. Similarly, the range for accurate identification across the 32 silent videos (silent videos refers to only silent videos and not silent videos with subtitles) is 37% to 86% with a median of 66%. There are 8 out of 16 fabricated silent videos and 10 out of 16 non-fabricated silent videos that participants accurately identify less than than 67% of the time. Figure 3-1 illustrates the proportion of participants who accurately distinguish between authentic and fabricated for the 32 silent videos.

In contrast, we find audio clips are easier to discern than text transcripts or silent videos.

On the audio clips with no subtitles, the accurate identification ranges from 56% to 86% with a median of 79%.

Figure 3-2 presents participants' weighted accuracy, confidence, perceived fabrications in fabricated speeches, perceived fabrications in non-fabricated speeches, and response duration across modality conditions. Weighted accuracy indicates participants' accuracy weighted by confidence (e.g., if a participant responded "82% confidence this is fabricated" and the participant is correct, then the participant is assigned a weighted accuracy score of 82, and otherwise, if the participant is incorrect, then the participant would be assigned a weighted accuracy score of 18). Confidence indicates participants' self-reported level of confidence which ranges from 50 (just as likely as chance) to 100 (full certainty). Perceived fabrications in fabricated and non-fabricated speeches is defined as a participant indicating a 51% or higher confidence that a stimulus is fabricated. Response time is measured in seconds and windsorized at the 99th percentile to control for time response outliers, which are an artifact of participants who return to the experiment after an extended time.

We evaluate the marginal effect of each condition on participants' weighted accuracy (and additional outcomes) via an ordinary least squares regression with standard robust errors clustered at the participant level following Abadie et al (2017) [2]. We find both recruited and non-recruited participants' accuracy increase as political speeches are presented with video and audio modalities. In this regression, which is also presented in column 1 of Table 3.1 in the Appendix, the dependent variable is weighted accuracy and the independent variables are binary indicators for assignment to communication modalities. Recruited participants' accuracy is 58% ($p < 0.001$) on transcripts, 7% ($p < 0.001$) higher on silent videos, 9% ($p < 0.001$) higher on silent videos with subtitles, 19% ($p < 0.001$) higher on audio clips and audio clips with subtitles, and 25% ($p < 0.001$) higher on videos with audio and videos with audio and subtitles.[3] Similarly in column 3 of Table 3.1, non-recruited participants' accuracy is 53% on transcripts, 13% ($p < 0.001$) higher on silent videos and silent videos with subtitles, 21% ($p < 0.001$) higher on audio clips and audio clips with subtitles, and 28-29% ($p < 0.001$) higher on videos with audio and videos with audio and subtitles. Overall,

[3] All p-values reported in this paper are generated by linear regression with robust standard errors clustered at the participant level unless otherwise noted.

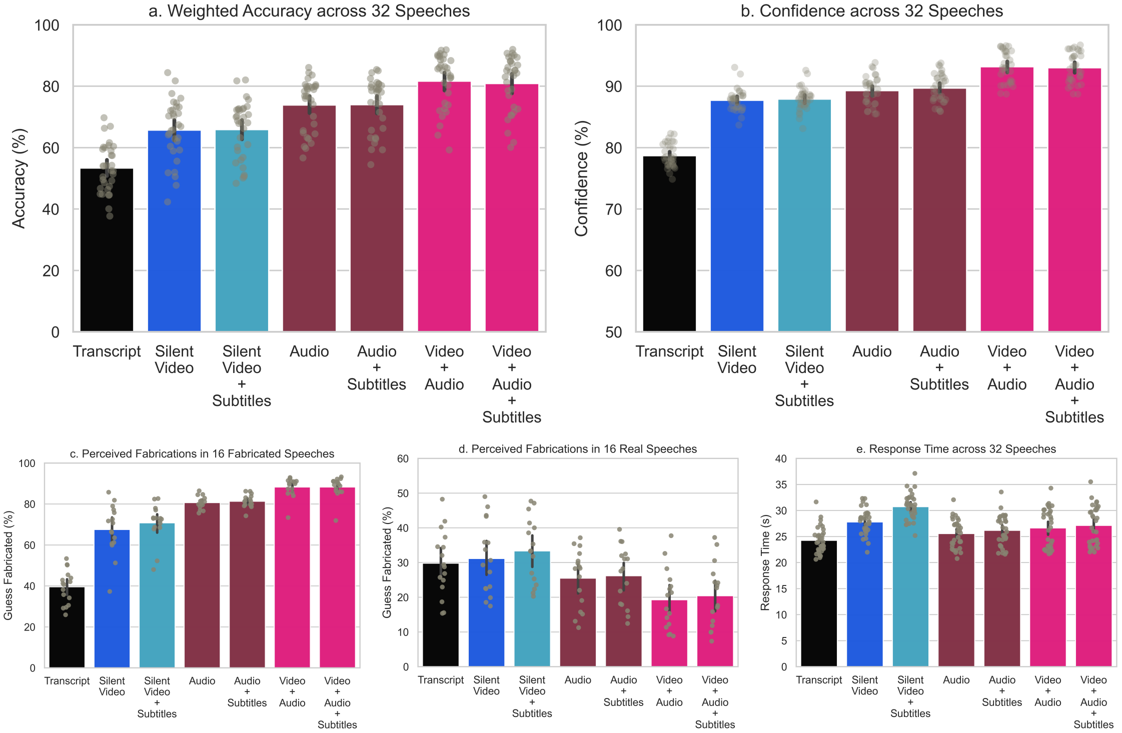


Figure 3-2: The mean and distribution of (a) weighted accuracy, (b) confidence, (c) perceived fabrications in fabricated speeches, (d) perceived fabrications in real speeches, and (e) response time are plotted for each of the seven modality conditions. The black lines indicate the 95% confidence interval of the true mean and the gray dots indicate each of the 32 speeches. Figure 3-2b plots confidence on a scale that ranges from a minimum of 50% confidence (just as likely as chance) to 100% confidence (full confidence). Figure 3-2e plots response time windsorized at the 99th percentile to control for time response outliers, which are an artifact of participants who return to the experiment after an extended time.

participants are better at identifying whether an event actually happened when watching videos or listening to audio than reading transcripts.

In contrast to the high variability in participants' accuracy across speeches and modality conditions, participants' confidence is less variable. On text transcripts, participants' mean confidence is 79%. Speeches presented via video and audio increase participants' confidence relative to text transcripts by 9% and 11% ($p < 0.001$) independently, respectively, and 15% together ($p < 0.001$). We find small effects of learning over time; for every stimuli seen, participants' accuracy increases by 0.27% ($p = 0.001$) and participants' confidence increases by 0.03% ($p = 0.006$), which means that on average accuracy increased by 8.64%

and confidence increase by 0.96% from the first stimulus seen to the last one seen.

As participants have access to additional communication modalities, participants' weighted accuracy, confidence, discernment of fabricated speeches, and discernment of real speeches increase on average. However, we do not find any significant, marginal effects of subtitles on any of the dependent variables for modality conditions that already include audio. The median response time across all stimuli was 27 seconds, which is 6 seconds longer than the average video length. The median response time for the silent, subtitled videos is 34 seconds, which is slightly longer than the response time for all other modality conditions. Across all 7 modality conditions, the median response time for fabricated stimuli is shorter than the median response time for non-fabricated stimuli; fabricated text, video, and audio have 3.8 seconds ($p < 0.001$), 5.6 seconds ($p < 0.001$), and 3.5 seconds ($p < 0.001$) shorter response times than their non-fabricated counterparts.

Based on this experiment's large sample size of 432,987 observations by participants who passed the attention check, the 224 stimuli in this experiment have a mean of 1,932 observations each. This large sample size per stimuli provides high statistical power to individually evaluate whether participants are discerning stimuli more accurately than chance. Specifically, using 1,933 observations provides over 99% statistical power to detect a 15 percentage point increase beyond chance at the $p < 0.05$ threshold. We evaluate the degree to which participants' discernment surpasses random chance by running a binomial test on responses to each stimuli within a modality condition and applying a Bonferonni correction [76], which means multiplying each p-value by 32 (the number of speeches per modality condition) to correct for multiple hypothesis testing.

After applying this correction for multiple hypothesis testing, we find participants' discernment is statistically significantly better than chance ($p < 0.05$) on 5 of 32 text transcripts and 26 of the 32 silent videos. In particular, participants are no better than chance ($p < 0.05$) on 4 of the 16 non-fabricated, silent videos and 2 of the 16 fabricated, silent videos. In other words, we have high statistical power, and we do not find evidence that participants are better than chance on 6 of the silent videos and 27 of the 32 text transcripts.

When the information from the political speech transcript and video are combined in the silent, subtitled videos, we find participants discern better than chance ($p < 0.05$) on all 16 of 16 fabricated, silent videos with subtitles and 9 of 16 non-fabricated, silent videos with subtitles. Likewise, the addition of audio significantly increases discernment rates; in all modality conditions with audio, participants discern better than chance ($p < 0.05$) on between 31 to 32 of the 32 political speeches.

Figure 3-2c and Figure 3-2d show the distributions of discernment rates across modality conditions for fabricated and real videos. Similarly to Figure 3-2a and Figure 3-2b, these plots show that regardless of whether the stimuli are fabricated or not, the addition of audio or video is associated with an increase in participants' discernment. However, we find slight differences in response bias: participants tend to identify text transcripts as real and the rest of the modalities as fabricated more often than random chance would suggest. For participants who did not select "Just as likely real or fabricated," participants respond that text transcripts and silent videos are fabricated in 44% ($p < 0.001$) and 53% ($p = 0.002$) of trials, respectively, while participants respond that the other 5 modality conditions are fabricated in 55% to 57% of trials ($p < 0.001$) (see Figure 3-8 for the percent of participants guessing a video is fabricated over the number of speeches a participant has seen). Participants selected "Just as likely real or fabricated" in 21% of text transcripts, 7% of silent videos, 6% of silent videos with subtitles, 6% of audio, 5% of audio with subtitles, and 3% of video and audio with or without subtitles.

In Figure 3-3, we present participants' marginal accuracy on transcripts, silent videos, and video with audio relative to silent, subtitled videos for each of the 32 speeches. Figure 3-3a reveals that participants are mostly less accurate on text transcripts than silent, subtitled videos. Likewise, Figure 3-3c shows participants are consistently more accurate on videos with audio than silent, subtitled videos. In contrast, Figure 3-3b illustrates heterogeneity in participants' performance with and without subtitles. In the following section, we examine this heterogeneity along two dimensions: whether the video is fabricated or not and whether the speech content is considered discordant with the politician's identity or not.

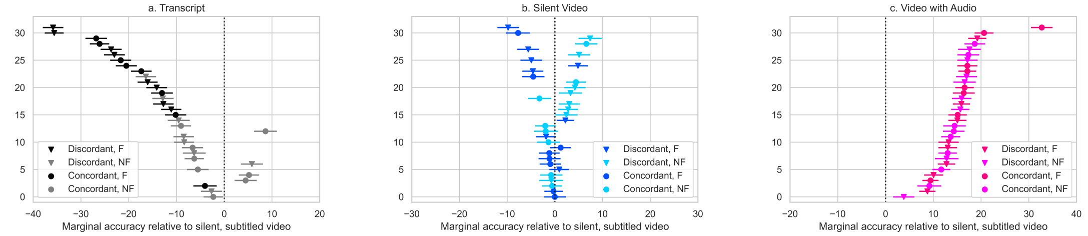


Figure 3-3: Participants' accuracy on silent, subtitled videos is compared against accuracy on transcripts, silent videos, and videos with audio for each of the 32 speeches. The error bars represent 95% confidence intervals. The 32 speeches are ordered by the absolute value of the difference in accuracy between the silent, subtitled video and the modality condition to which it is being compared. The legend indicates whether the video shows the politician expressing political views concordant or discordant with his expected political ideology and "F" and "NF" refer to fabricated and not fabricated, respectively.

3.2.3 Heterogeneous Moderating Effects of Discordant Messages

We evaluate how discordant messages influence participants' discernment by examining the interactions between discordance and modality conditions in the linear regressions on participants' weighted accuracy presented in Table 3.2 and Table 3.3 in the Appendix. We limit this analysis to recruited participants for two reasons: first, recruited participants are all from the United States while the majority of non-recruited participants visited the website from outside the United States and it is unclear how familiar non-recruited participants are with United States politicians' viewpoints; second, we also evaluate these effects with respect to CRT performance, which we only collected for recruited participants.

When considering all 32 fabricated and real speeches together (see column 1 of Table 3.2 in the Appendix), we find participants are 4.7 percentage points ($p = 0.002$) more accurate on silent, subtitled videos than the same videos without subtitles. However, we find participants are 5.0 percentage points ($p = 0.018$) less accurate on the discordant silent, subtitled videos than the same silent videos without subtitles. In other words, the addition of subtitles reduces discernment accuracy for political speeches that are discordant with the general public's perception of what politicians would say.

In order to further evaluate this effect, we consider fabricated videos and non-fabricated videos separately in columns 2 and 3 in Table 3.2 in the Appendix. We find the negative effect of discordance on subtitled videos is driven by participants' discernment of non-fabricated videos. We find participants are 6.8 percentage points ($p = 0.021$) less accurate on discordant silent, subtitled videos that are not fabricated compared to the same silent videos without subtitles. In contrast, we do not find a statistically significant difference ($p = 0.341$) between participants' performance on discordant silent, subtitled videos that have been fabricated and the same silent videos without subtitles. The negative effects of subtitles on non-fabricated yet discordant silent videos indicates the content of a message can change how participants weigh visual information.

The heterogeneous effects of subtitles on the discernment of silent videos is robust to our specification of discordance. In Table 3.3 in the Appendix, we consider the same regressions as Table 3.2 in the Appendix except we replace the binary variable indicating discordance with a continuous variable for how discordant the speech is with the speaker based on the independent survey with 84 participants on how well the political speeches match either politicians' political views. The regressions in columns 2 and 3 of Table 3.3 in the Appendix present qualitatively similar results as Table 3.2 in the Appendix. When we consider discordance based on the public's perceived discordance, we find participants are 4.2 percentage points ($p = 0.003$) less accurate on discordant silent, subtitled videos that are not fabricated compared to the same silent videos without subtitles. Likewise, we do not find a statistically significant difference ($p = .751$) between participants' performance on discordant silent, subtitled videos that have been fabricated and the same silent videos without subtitles.

3.2.4 Heterogeneous Moderating Effects of the Cognitive Reflection Test (CRT)

We find that participants' performance on the CRT moderates participants' discernment accuracy. In this analysis, the CRT score is a continuous variable ranging from 0 to 3 with 124 participants answering none correctly and 109, 122, and 154 participants answering 1, 2, and 3 questions correctly, respectively. For every question that participants answer correctly on the CRT, participants are 2.9 percentage points ($p = 0.002$) more accurate (see column 4 in Table 3.2 in the Appendix). Likewise, participants who respond correctly to all 3 items on the CRT are 8.7 percentage points ($p = 0.002$) more accurate than participants who respond incorrectly to all 3 items. In Figure 3-7 in the Appendix, we present the distribution of media truth discernment scores following Pennycook and Rand (2019) [496] for “intuitive” participants who incorrectly answered all 3 CRT items and “deliberative” participants who correctly answered all 3 CRT items.

We also find that participants’ performance on the CRT moderates the influence of subtitles on the discernment accuracy of discordant messages in silent videos. In columns 4-6 of Table 3.2 in the Appendix, we report regressions that include the same independent variables as columns 1-3 plus interaction effects of these independent variables with participants’ scores on the CRT. As a visual aid, we present these results in Figure 3-4. In column 6 where we consider only non-fabricated videos, we find the coefficient on the interaction between “Discordant” and “Silent Subtitled Video” is negative 17.5 percentage points ($p < 0.001$), which means that participants are that much less accurate on non-fabricated discordant silent, subtitled videos than the same silent videos without subtitles while holding all else constant. The interaction between “CRT Score,” “Discordant,” and “Silent Subtitled Video” is 6.3 percentage points ($p = 0.011$), which means for each correct response to the CRT, participants are 6.3 percentage points more accurate at identifying discordant silent, subtitled videos while holding all else constant. This means that participants who answered all 3 CRT items correctly would be 18.9 percentage points ($p = 0.011$) more accurate on discordant silent, subtitled videos than participants who failed to answer any CRT item correctly. This improvement by 18.9 percentage points for answering all CRT items correctly cancels out the 17.5 percentage point decrease associated with discordant silent, subtitled videos compared to the same silent videos without subtitles. In other words, perfect performance on the CRT moderates the negative effects of discordant content such that participants are considering visual information and discerning just as accurately on silent subtitled videos as the same silent videos without subtitles. These results are qualitatively similar when we replace the binary variable for discordance with the continuous variable for discordance in Table 3.3.

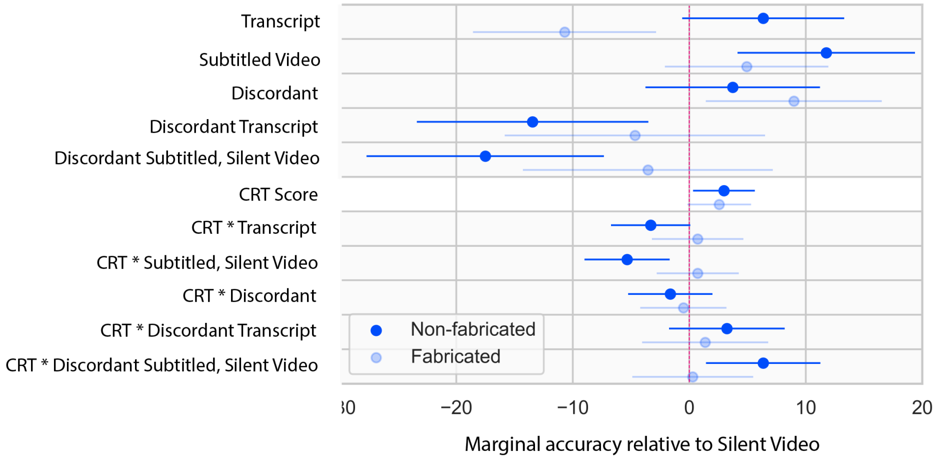


Figure 3-4: Average treatment effect of assignment to modality conditions and their interaction with discordant speeches and participants' performance on the Cognitive Reflection Test. The error bars represent 95% confidence intervals.

3.3 Discussion

This work provides evidence, via a randomized experiment with 224 authentic and fabricated stimuli and 41,822 participants, that visual and auditory communication modalities increase people's ability to distinguish authentic political speeches from fabricated political speeches. In particular, we provide corroborating evidence to the conventional wisdom around the "seeing is believing" narrative (the realism heuristic that suggests people will tend to trust video over text[587] and recent results showing people "are more likely to believe an event occurred when it is presented in video versus textual form" [645]) in the context of authentic speeches: people are significantly more accurate at identifying authentic speeches as authentic when the speeches include audio and visual modalities as opposed to only text. However, these results add considerable nuance to the seeing is believing narrative when considering fabricated speeches: people are significantly more accurate at identifying fabricated speeches as fabricated when the speeches include audio and visual modalities as opposed to only text. In other words, we find participants are significantly more accurate at distinguishing between authentic and fabricated political videos than transcripts.

These results are based on an experiment with a stimuli set that is much larger than most stimuli sets for the psychology of media effects research [524] and deepfake detection [46, 165, 614], but it is important to add a caveat that we focused on a single context, political speeches, and algorithm, the deepfake lip-syncing wav2lip algorithm, which is very effective at manipulating a person who is facing forward and already speaking into a convincing fake video. While we present evidence that adds considerable nuance to the media effects literature on communication modalities, future work may consider additional nuances by exploring heterogeneity based on other kinds of deepfake manipulations like face swapping and head puppetry [397], contexts that require more sophistication to produce a convincing fake (e.g. where a person is moving, turning their head, and interacting with other people), and who is being manipulated [86].

These results cannot simply be explained by the deepfake manipulations being too obvious or unrealistic. On silent videos without subtitles, we find participants are only 64% accurate at identifying manipulations (see Figure 3-2c and Figure 3-2d for the distribution of people guessing stimuli are fabricated across the seven modalities). Moreover, we find participants do not perform better than chance in nearly half of the silent videos. Participants' performance on the silent videos is relevant to the quality of the deepfake manipulations because it avoids the confounding from the speech content and the audio. The participants' low performance on silent videos offers evidence that visual artefacts and inconsistencies created by the lip syncing deepfake manipulations are not readily apparent to most people, and as such, these videos represent a reasonable stimuli set for examining how well people can distinguish real from fake videos.

People distinguish authentic from fabricated videos based on perceptual cues from video and audio and considerations about the content (e.g., the degree to which what is said matches participants' expectations of what the speaker would say, which is known as the expectancy violation heuristic [426]). With the message content alone, participants are only slightly better than random guessing at 57% accuracy on average. With perceptual information from video and the message content via subtitles, participants are slightly more accurate (and more confident) at 66% accuracy on average, and with information from both video and audio, participants are even more accurate (and more confident) at 82%

accuracy on average. Our finding that participants are more accurate at distinguishing between real and fabricated on audio than silent video with subtitles aligns with the social psychology literature demonstrating people tend to rely on auditory information more than visual information for both discerning sincerity [47] and ascribing authorship of a script to a human (as opposed to a computer) [558]. Overall, the experiment's results show that as participants have access to more information via audio and video, they are better able to distinguish whether a political speech has been fabricated.

However, we find one notable exception to the result that more information leads to higher accuracy in distinguishing fabricated speeches from authentic ones: political speeches that conflict with the public's perspective of what a politician would say are harder to discern in silent, subtitled videos than the same silent videos without subtitles. This effect on discordant speeches (but not concordant speeches) is not driven by subtitles distracting participants. We do not find any evidence of any effect on subtitles when audio is included. Instead, the heterogeneous effects of concordant and discordant speech content are a consequence of how participants handle cognitive dissonance and balance the consideration of perceptual and content-based information. We find that these effects are driven by responses to non-fabricated videos and are moderated by deliberative, reflective thinking as measured by the CRT.

Fabricated videos differ from non-fabricated videos in how people can discern their authenticity. Fabricated videos involve visual manipulations, which can sometimes be explicitly identified (e.g., a glitch, a flicker, or mechanical and otherwise out of place lip movement). If someone finds a suspicious visual artefact, then that individual can be quite confident the video has been fabricated. In contrast, non-fabricated videos have not been visually manipulated. As a result, there is no single bit of information to signify fabrication or authenticity. Furthermore, we find people take on average 2.5s to 3.8s longer to provide a response to non-fabricated speeches than fabricated speeches. If someone cannot find a visual distortion, then that individual cannot be perfectly certain that the video has or has not been fabricated; for example, the video may have been fabricated without any perceptible distortion, or perhaps, the individual has yet to find the subtle visual distortion. This asymmetry between assessing fabricated and non-fabricated speeches exacerbates the “liar’s dividend” where the general possibility that speeches can be fabricated calls into question whether any speech is fabricated and thus enables “liars to avoid accountability for things that are in fact true.” [120, 601] Clear articulation of the precise state-of-the-art algorithms and associated contexts in which audio-visual content can be fabricated to be indistinguishable from the real thing can help inform how people assess the content they consume and reduce the effects of the “liar’s dividend.”

We find that participants’ performance on the CRT moderates the effects of subtitles on the discernment accuracy of silent videos. In particular, participants who correctly answered all three CRT items show no difference in discernment rates of discordant silent, subtitled videos relative to the same silent videos without subtitles. But, for every CRT item that participants incorrectly answer, participants are 6.3 percentage points less accurate on real discordant silent, subtitled videos than the same silent videos without subtitles. In other words, reflective thinking moderates how participants balance what is said (the content of the speech) with how it is said (visual information). Our results show that the least reflective participants tend to rely on the expectancy violation heuristic and discount visual information more than the most reflective participants.

Unlike for videos and transcripts, we cannot disentangle the content and perceptual information for audio modalities. Nevertheless, we find that the interaction between discordant speeches and any audio condition is negative after controlling for the level effects of discordance and any audio. This suggests that discordant media not only impair the incorporation of visual cues but may also impair attention to and incorporation of auditory cues as well.

The danger of fabricated videos may not be the average algorithmically produced deepfake but rather a single, highly polished, and extremely convincing video. For example, hyper-realistic deepfakes like the Tom Cruise deepfakes on Tiktok (see `https://www.tiktok.com/@deeptomcruise`) are produced by visual effects artists using both artificial intelligence algorithms and video editing software. While these hyper-realistic deepfakes may still contain manipulation artifacts (e.g., unattached earlobes that do not match Tom Cruise’s attached earlobes [10]), future work on the psychology of multimedia misinformation may consider hyper-realistic videos produced by visual effects studios in addition to algorithmically manipulated videos.

Political deepfakes are most dangerous when people are least expecting information to be manipulated, and this experiment on multimedia truth discernment does not match the ecological realities that people typically face when confronted with fake news. In this experiment, 50% of content is fake, and we explicitly inform participants of this base rate. In today's media ecosystem, fake news is relatively rare: less than a fraction of a percent [19, 634] of news is fake news. As such, this experiment is useful to study how people discern multimedia information when attending to questions of accuracy, but it is less useful in understanding how people will share misinformation they read on social media. People are generally highly accurate in discerning the veracity of news headlines yet share fake news headlines because their attention is not focused on accuracy [497]. On social media, video-based misinformation will often be designed to incorporate characteristics (e.g., fear, disgust, surprise, novelty) that divert people's focus from accuracy and make content go viral [61, 624]. Given that multimedia misinformation may be both easier to discern and more frequently shared on social media than text-based media, more research needs to be done to understand how people allocate attention while browsing the Internet [363]. Recent research shows that educational material on common misinformation techniques can improve people's ability to discern trustworthy from untrustworthy videos [533]. Finally, discernment – how accurately people discern misinformation – is different than belief – how much people report they believe misinformation. It is possible (though quite peculiar) that someone could be highly accurate at discerning truth from falsehood while also tending to believe the fabricated content and not believe the true content. For example, research on fake news headlines and articles finds that people are better at discerning news concordant with their political leanings than discordant news while also believing concordant news more often than discordant news [495].

The finding that videos of political speeches are easier to distinguish as authentic or fabricated than text transcripts highlights the need to re-introduce and explain the oft-forgotten second half of the "seeing is believing" adage. In 1732, the old English adage appears as:

"Seeing is believing but feeling is the truth." [218] Here, "feeling" does not refer to emotion but rather direct experience. Since the advent of photography, society has generally understood that what we see in a photograph is not always the truth and further assessment is often necessary [189, 331, 424].

In this paper, we examined a bounded question – how well can ordinary people discern whether or not a short soundbite of a political speech by a well-known politician in text, audio, or video has been fabricated – and we find that more information via communication modalities – text transcripts vs. silent, subtitled video vs. video with audio – enables people to more accurately discern fabricated and real political speeches. These results are particularly relevant for the design of content moderation systems for flagging misinformation on social media. In particular, we suggest content moderation flags include explanations that address which component part of a video appears to be fabricated. These explanations could allow people to appropriately allocate attention to the content [360] or perceptual cues (e.g., low-level pixel features, high-level semantic features, and biometric-based features [14]) when trying to assess the content's authenticity.

Finally, these findings offer insights into political communication and communication theory more generally; there is more to how humans form beliefs than the "seeing is believing" narrative would suggest because people are paying attention to both what is said and how something is said.

3.4 Methods

3.4.1 Virtual Experiment Website

We hosted multimedia stimuli – transcripts, audio, and video of fabricated and authentic political speeches – on a custom designed website called Detect Fakes[4]. In the experiment, we asked participants to identify fabricated and non-fabricated stimuli. After collecting informed consent and presenting participants with instructions, we show participants a short

[4] Detect Fakes is currently hosted at `https://detectfakes.media.mit.edu/`.

political speech and ask “Did [Joseph Biden/Donald Trump] say that?” followed by “Please [read/listen/watch] this [transcript/audio clip/video] from [Joseph Biden/Donald Trump] and share how confident you are that it is fabricated. Remember half the media snippets we show are real and half are fabricated.” Figure 3-5 in the Supplementary Information section presents a screenshot of the user interface, which shows participants were instructed to move a slider to report their confidence from 50% to 100% that a stimulus is fabricated (or 50% to 100% that a stimulus is not fabricated). After each response, we informed participants whether the stimulus was actually fabricated and then presented participants with another stimulus selected at random until participants viewed all 32 stimuli or decided to leave the experiment. Each participant began the experiment with an attention check stimulus.

3.4.2 Multimedia Stimuli

The multimedia stimuli are drawn from the Presidential Deepfake Dataset (PDD)[550], which is made up of 32 videos showing two United States presidents making political speeches. Half the videos are authentic videos that have not been altered by a deepfake algorithm. The other half have been fabricated to make the politicians appear to say something that they have not said. The fabricated videos were produced by writing a fabricated script, recording professional voice-actors reading the script, and applying a deepfake lip-syncing algorithm [510] to real videos of Joseph Biden and Donald Trump to make it appear as if the politicians actually gave such a fabricated speech. The mean duration of the videos is 21 seconds and all videos are recorded at 30 frames per second and have a resolution of 854 by 480 pixels. The PDD is balanced across three dimensions: (1) videos that have and have not been fabricated, (2) videos of Joseph Biden and Donald Trump, and (3) videos of the two politicians making concordant and discordant speeches with what the general public believes are the politicians’ political views.

In this experiment, we transform each of the original videos from the PDD into 7 different forms of media: a transcript, an audio clip, a silent video, audio with subtitles, silent video with subtitles, video with audio, and video with audio and subtitles. As a result, there are 7 modality conditions, 32 unique speeches, and 224 unique stimuli. On the experiment website, the transcript appears as HTML text and the six other forms of media content appear in a video player. The audio clip shows a black screen in the video player and the audio clip with subtitles shows a black screen with subtitles at the bottom. Each participant encounters each political speech in only one modality.

3.4.3 Concordance and Discordance Validation

In order to validate the concordance and discordance of speeches, we conducted an independent survey where 84 participants who passed an attention check rated each of the 32 transcripts for how well the political speeches match either politicians' political views. Participants were instructed "For each statement, we want you to rank how closely the statement matches your understanding of President Joseph Biden or President Donald Trump's political views" and asked to provide a judgment on a 5-point Likert scale from "Strongly Disagree" (-2) to "Strongly Agree" (2) that "This statement matches President [Joseph Biden's/Donald Trump's] political viewpoint: [statement]." Participants' responses confirm that speeches designed to be concordant and discordant with the two politicians views were indeed concordant and discordant with the average participants' perception of the politicians' views. The Z-values of participants responses to concordant and discordant speeches are -0.25 and 0.21, respectively, and this difference is statistically significant with $p < 0.001$ based on a T-Test. In Table 3.2, the "Discordant (Binary)" variable refers to the categories as outlined in the PDD, and in Table 3.3 the "Discordant (Continuous)" variable refers to these Z-values.

3.4.4 Randomization

We randomly assigned the order in which the 32 unique political speeches are presented to participants and each political speech is randomly assigned to one of the seven conditions. By randomly assigning the order of political speeches and the modality condition in which speeches were presented, we can identify the causal impact of media modality on participants' ability to discern misinformation.

3.4.5 Consent and Ethics

This research complies with all relevant ethical regulations and the Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects determined this study to fall under Exempt Category 3 – Benign Behavioral Intervention. This study's exemption identification number is E-3105. All participants are informed that "Detect Fakes is an MIT research project. All guesses will be collected for research purposes. All data for research were collected anonymously. For questions, please contact detectfakes@mit.edu. If you are under 18 years old, you need consent from your parents to use Detect Fakes." Most participants arrived at the website via organic links on the Internet. For participants recruited from Prolific, we compensated participants at a rate of $9.78 an hour and provided bonus payments of $5 to the top 1% of participants. Before beginning the experiment, all participants from Prolific were also provided a research statement, "The findings of this study are being used to shape science. It is very important that you honestly follow the instructions requested of you on this task, which should take a total of 15 minutes. Check the box below based on your promise:" with two options, "I promise to do the tasks with honesty and integrity, trying to do them uninterrupted with focus for the next 15 minutes." or "I cannot promise this at this time." Participants who responded that they could not do this at this time were re-directed to the end of the experiment.

3.5 Data and Code Availability

The datasets and code generated and analyzed during the current study are available in our public Github repository, `https://github.com/mattgroh/fabricated-political-speeches` (the Github repository will be set to public upon peer-reviewed publication). All PDD videos are available on Youtube with links provided in the Presidential Deepfakes Dataset paper [550].

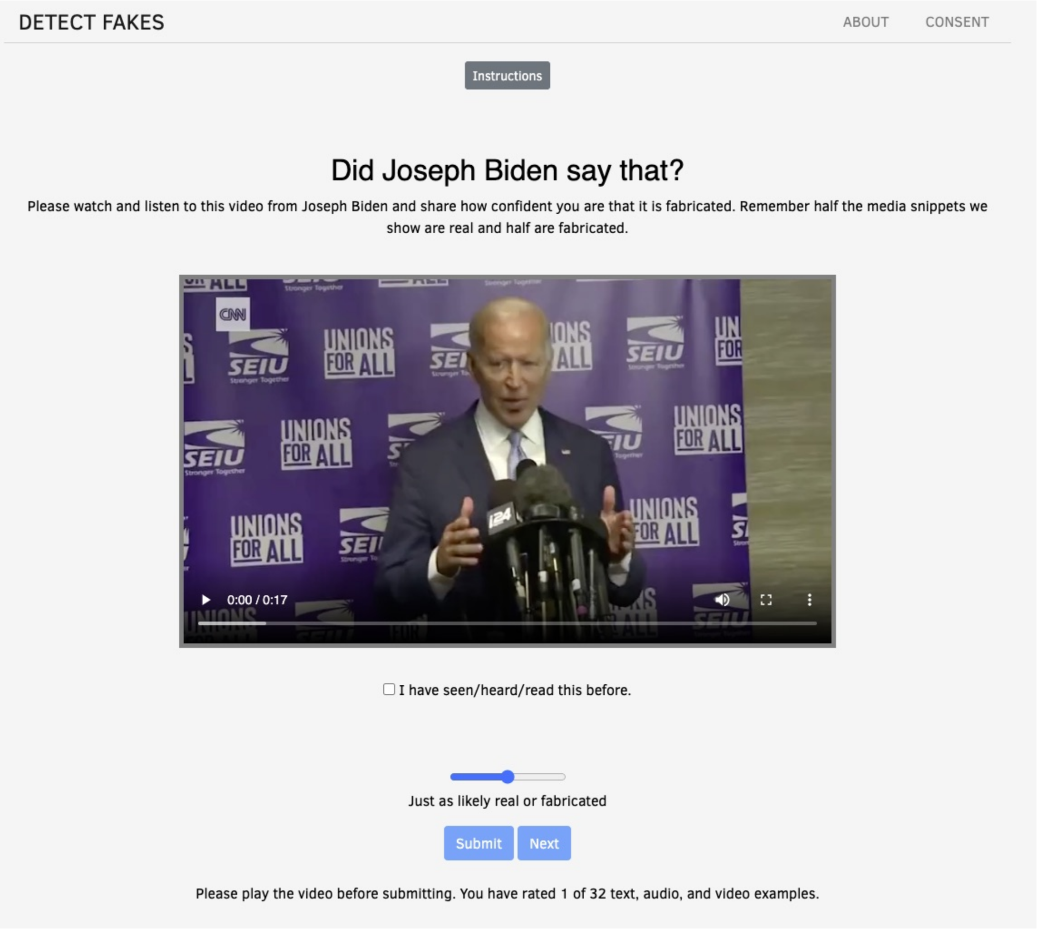
Acknowledgements

The authors would like to acknowledge funding from MIT Media Lab member companies, thank JL Cauvin and Austin Nasso (@jlcauvin and @austinnasso on Tiktok) for providing voice impressions, and thank the following individuals for helpful feedback: David Rand, Gordon Pennycook, Rahul Bhui, Yunhao (Jerry) Zhang, Ziv Epstein, and members of the Affective Computing lab at the MIT Media Lab and the Human Cooperation lab at MIT Sloan School of Management.

Author Contributions

M.G. conceived the experiments, A.S. developed the synthetic media and conducted the experiments, A.S. and M.G. analyzed the results, M.G. wrote the manuscript, and A.S., A.L., M.G., and R.P. reviewed and edited the manuscript.

3.6 Appendix



Did Joseph Biden say that?

Please watch and listen to this video from Joseph Biden and share how confident you are that it is fabricated. Remember half the media snippets we show are real and half are fabricated.

□ I have seen/heard/read this before.

Just as likely real or fabricated

Please play the video before submitting. You have rated 1 of 32 text, audio, and video examples.

Figure 3-5: Screenshot of experimental user interface.

Figure 3-6: Image from Sankaranarayan et al. (2021) showing the first frame at the 10 second mark of each of the 32 videos in the Presidential Deepfake Dataset, which is where the stimuli from this experiment are drawn.

	Dependent variable: Weighted Accuracy							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	57.61***	56.19***	52.70***	53.14***	53.11***	52.09***	52.77***	51.76***
	(0.83)	(1.11)	(0.58)	(0.15)	(0.15)	(0.19)	(0.13)	(0.17)
Silent Video	6.58***	4.52**	11.70***	12.82***	12.74***	6.83***	11.97***	6.45***
	(1.19)	(1.55)	(0.86)	(0.23)	(0.22)	(0.30)	(0.20)	(0.26)
Silent Video with Subtitles	8.80***	5.80***	12.45***	12.85***	12.82***	8.16***	11.95***	7.57***
	(1.11)	(1.47)	(0.87)	(0.23)	(0.22)	(0.29)	(0.19)	(0.25)
Audio	19.48***	18.52***	20.45***	20.66***	20.65***	19.32***	19.30***	18.07***
	(1.17)	(1.50)	(0.84)	(0.23)	(0.22)	(0.29)	(0.19)	(0.25)
Audio with Subtitles	19.41***	18.76***	19.71***	20.89***	20.80***	19.56***	19.25***	18.13***
	(1.05)	(1.46)	(0.89)	(0.23)	(0.22)	(0.29)	(0.19)	(0.25)
Video with Audio	25.12***	24.27***	27.97***	28.54***	28.50***	25.57***	26.81***	24.10***
	(1.09)	(1.46)	(0.80)	(0.21)	(0.21)	(0.28)	(0.18)	(0.24)
Video with Audio and Subtitles	24.75***	22.87***	27.42***	27.82***	27.79***	24.96***	26.18***	23.50***
	(1.08)	(1.46)	(0.81)	(0.22)	(0.21)	(0.28)	(0.19)	(0.25)
Number of Individuals	509	509	2807	38510	41313	37688	58344	52428
Number of Speeches	32	18	32	32	32	18	32	18
Observations	16,086	9,053	29,627	387,274	416,901	234,899	537,936	303,264
R^2	0.07	0.07	0.06	0.07	0.06	0.06	0.06	0.05
Note:								*p<0.05; **p<0.01; ***p<0.001

Table 3.1: Ordinary least squares regressions with robust standard errors clustered on participants. Weighted accuracy is the dependent variable. The "Transcript" condition is held out and represented by the constant term. Column (1) shows recruited participants, column (2) shows recruited participants on "difficult" videos with lower than 67% accurate identification, column (3) shows pre-registered non-recruited participants, column (4) shows non-recruited participants after the pre-registration window, column (5) shows all non-recruited participants, column (6) shows non-recruited participants on "difficult" videos, column (7) shows non-recruited participants including participants who fail the attention check, column (8) shows non-recruited participants including participants who fail the attention check on "difficult videos."

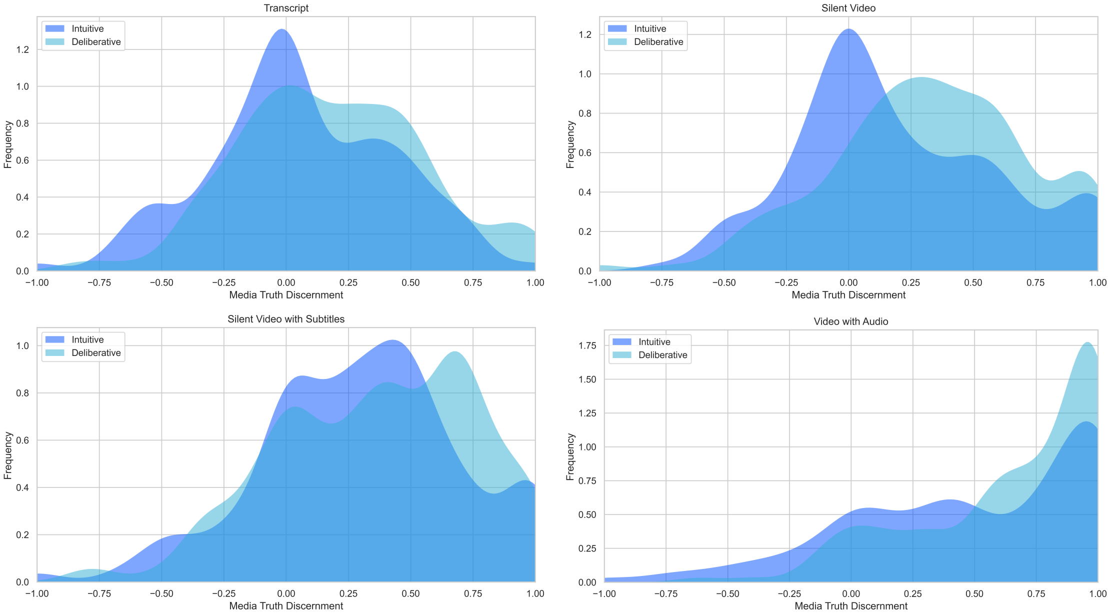


Figure 3-7: Distribution of media truth discernment scores following Pennycook and Rand (2019) where the score is positively associated with accuracy at distinguishing fabricated media from authentic media.

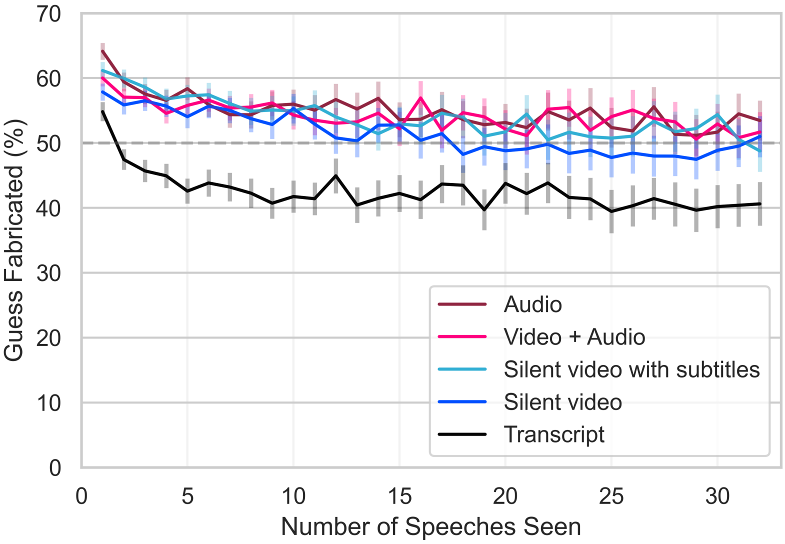


Figure 3-8: Percent of participants who respond that the speech is fabricated across modalities and the number of videos seen.

	Dependent variable: Weighted Accuracy					
	All	Fabricated	Not Fabricated	All	Fabricated	Not Fabricated
	(1)	(2)	(3)	(4)	(5)	(6)
Constant (Silent Video)	61.97***	57.97***	66.17***	57.28***	53.94***	61.22***
	(1.14)	(1.66)	(1.55)	(1.92)	(2.79)	(2.90)
Transcript	-3.92*	-9.24***	0.84	-0.85	-10.71**	6.34
	(1.58)	(2.43)	(2.01)	(2.68)	(4.01)	(3.55)
Subtitled Silent Video	4.67**	6.20**	2.88	8.67**	4.92	11.75**
	(1.53)	(2.14)	(2.13)	(2.69)	(3.58)	(3.89)
Any Audio	18.22***	23.73***	12.49***	19.06***	22.23***	15.31***
	(1.19)	(1.83)	(1.57)	(2.06)	(3.19)	(3.02)
Discordant (Binary)	4.57**	8.09***	0.85	6.64*	8.97*	3.72
	(1.54)	(2.25)	(2.17)	(2.71)	(3.86)	(3.82)
Discordant (Binary) * Transcript	-5.46*	-2.68	-7.97**	-10.18**	-4.67	-13.45**
	(2.21)	(3.28)	(3.01)	(3.90)	(5.70)	(5.07)
Discordant (Binary) * Subtitled Silent Video	-5.02*	-2.91	-6.81*	-10.86**	-3.56	-17.52***
	(2.12)	(3.06)	(2.96)	(3.75)	(5.47)	(5.20)
Discordant (Binary) * Any Audio	-5.42**	-6.54**	-4.20	-7.30*	-7.71	-6.30
	(1.68)	(2.43)	(2.36)	(3.01)	(4.15)	(4.26)
CRT Score				2.90**	2.56	2.97*
				(0.95)	(1.39)	(1.35)
CRT Score * Transcript				-1.90	0.71	-3.32
				(1.32)	(2.00)	(1.73)
CRT Score * Subtitled Silent Video				-2.48	0.72	-5.34**
				(1.32)	(1.79)	(1.86)
CRT Score * Any Audio				-0.48	0.90	-1.63
				(1.01)	(1.57)	(1.40)
CRT Score * Discordant (Binary)				-1.20	-0.51	-1.63
				(1.31)	(1.88)	(1.85)
CRT Score * Discordant (Binary) * Transcript				2.83	1.36	3.23
				(1.84)	(2.76)	(2.53)
CRT Score * Discordant (Binary) * Subtitled Silent Video				3.46	0.29	6.34*
				(1.86)	(2.64)	(2.51)
CRT Score * Discordant (Binary) * Any Audio				1.08	0.67	1.15
				(1.43)	(2.03)	(2.10)
Number of Individuals	509	507	509	509	507	509
Number of Speeches	32	16	16	32	16	16
Observations	16,086	8,042	8,044	16,086	8,042	8,044
R^2	0.07	0.12	0.04	0.07	0.13	0.04
Note:	*p<0.05; **p<0.01; ***p<0.001					

Table 3.2: Ordinary least squares regressions with robust standard errors clustered on participants. Weighted accuracy is the dependent variable. The “Silent Video” condition is held out and represented by the constant term. The "Discordance (Binary)" variable is defined in Sankaranarayanan et al. (2021) by whether the speaker’s political views are discordant with the speech content.

	Dependent variable: Weighted Accuracy					
	All	Fabricated	Not Fabricated	All	Fabricated	Not Fabricated
	(1)	(2)	(3)	(4)	(5)	(6)
Constant (Silent Video)	64.19***	61.64***	66.76***	60.60***	58.03***	63.48***
	(0.88)	(1.29)	(1.19)	(1.55)	(2.24)	(2.27)
Transcript	-6.62***	-12.21***	-5.47***	-6.02**	-14.99***	-3.19
	(1.19)	(1.84)	(1.54)	(2.11)	(3.19)	(2.81)
Subtitled Silent Video	2.22*	4.78**	-1.52	3.26	3.05	1.15
	(1.05)	(1.61)	(1.61)	(1.87)	(2.76)	(3.10)
Any Audio	15.58***	20.49***	9.35***	15.40***	18.34***	10.95***
	(0.89)	(1.42)	(1.21)	(1.53)	(2.54)	(2.27)
Discordance (Continuous)	0.29	0.99	0.73	0.93	1.30	1.74
	(0.76)	(1.25)	(0.97)	(1.32)	(2.11)	(1.74)
Discordance (Continuous) * Transcript	-2.80**	7.07***	-8.99***	-4.43*	6.79*	-9.92***
	(1.07)	(1.61)	(1.41)	(2.00)	(2.92)	(2.52)
Discordance (Continuous) * Subtitled Silent Video	-1.36	0.56	-4.19**	-4.13*	0.73	-8.51***
	(1.10)	(1.77)	(1.39)	(1.97)	(3.22)	(2.44)
Discordance (Continuous) * Any Audio	-0.94	0.46	-4.55***	-1.59	0.52	-4.91**
	(0.80)	(1.29)	(1.10)	(1.41)	(2.20)	(1.89)
CRT Score				2.26**	2.32*	2.03
				(0.76)	(1.11)	(1.06)
CRT Score * Transcript				-0.43	1.55	-1.42
				(1.03)	(1.55)	(1.35)
CRT Score * Subtitled Silent Video				-0.71	0.96	-1.66
				(0.91)	(1.35)	(1.46)
CRT Score * Any Audio				0.09	1.28	-0.98
				(0.75)	(1.22)	(1.05)
CRT Score * Discordance (Continuous)				-0.36	-0.17	-0.58
				(0.65)	(1.06)	(0.87)
CRT Score * Discordance (Continuous) * Transcript				0.96	0.28	0.51
				(0.96)	(1.39)	(1.26)
CRT Score * Discordance (Continuous) * Subtitled Silent Video				1.62	-0.16	2.51*
				(0.97)	(1.59)	(1.23)
CRT Score * Discordance (Continuous) * Any Audio				0.34	-0.07	0.16
				(0.69)	(1.10)	(0.96)
Number of Individuals	509	507	509	509	507	509
Number of Speeches	32	16	16	32	16	16
Observations	16,086	8,042	8,044	16,086	8,042	8,044
R^2	0.07	0.12	0.04	0.07	0.13	0.04
Note:						*p<0.05; **p<0.01; ***p<0.001

Table 3.3: Ordinary least squares regressions with robust standard errors clustered on participants. Weighted accuracy is the dependent variable. The “Silent Video” condition is held out and represented by the constant term. The "Discordance (Continuous)" variable is computed by calculating the z-transformation of participants’ mean response on a 5-point Likert scale for how well a speech aligns with the public’s perception of the politicians’ viewpoints.

Chapter 4

Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset

Abstract

How does the accuracy of deep neural network models trained to classify clinical images of skin conditions vary across skin color? While recent studies demonstrate computer vision models can serve as a useful decision support tool in healthcare and provide dermatologist-level classification on a number of specific tasks, darker skin is underrepresented in the data. Most publicly available data sets do not include Fitzpatrick skin type labels. We annotate 16,577 clinical images sourced from two dermatology atlases with Fitzpatrick skin type labels and open-source these annotations. Based on these labels, we find that there are significantly more images of light skin types than dark skin types in this dataset. We train a deep neural network model to classify 114 skin conditions and find that the model is most accurate on skin types similar to those it was trained on. In addition, we evaluate how an algorithmic approach to identifying skin tones, individual typology angle, compares with Fitzpatrick skin type labels annotated by a team of human labelers.[1]

[1] This chapter, which is co-authored by Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri, appeared in the proceedings for the Computer Vision and Pattern Recognition (CVPR) International Skin Imaging Collaboration (ISIC) workshop in 2021 [248].

4.1 Motivation

How does the accuracy of deep neural network models trained to classify clinical images of skin conditions vary across skin color? The emergence of deep neural network models that can accurately classify images of skin conditions presents an opportunity to improve dermatology and healthcare at large [107, 187, 388, 580, 608]. But, the data upon which these models are trained are mostly made up of images of people with light skin. In the United States, dark skin is underrepresented in dermatology residency programs [380], textbooks [9, 23], dermatology research [378], and dermatology diagnoses [253, 481]. With the exception of PAD-UFES-20 [478], none of the publicly available data sets identified by the Sixth ISIC Skin Image Analysis Workshop at CVPR 2021 (Derm7pt [321], Dermofit Image Library, ISIC 2018 [123, 607], ISIC 2019 [122, 129, 607], ISIC 2020[128, 536], MED-NODE [229], PH2 [422], SD-128 [585], SD-198, SD-260) include skin type or skin color labels or any other information related to race and ethnicity. The only dataset with such skin type labels, PAD-UFES-20, contains Fitzpatrick skin type labels for 579 out of 1,373 patients in the dataset. The lack of consideration of subgroups within a population has been shown to lead deep neural networks to produce large accuracy disparities across gender and skin color for facial recognition [87], across images with and without surgical markings in dermatology [69, 643], and across treated and untreated conditions in radiology [468]. These inaccuracies arise from dataset biases, and these underlying data biases can unexpectedly lead to systematic bias against groups of people [3, 49]. If these dataset biases are left unexamined in dermatology images, machine learning models have the potential to increase healthcare disparities in dermatology [7].

By creating transparency and explicitly identifying likely sources of bias, it is possible to develop machine learning models that are not only accurate but also serve as discrimination detectors [133, 337, 470]. By rigorously examining potentials for discrimination across the entire pipeline for machine learning model development in healthcare [114], we can identify opportunities to address discrimination such as collecting additional data from underrepresented groups [112] or disentangling the source of the disparities [503]. In this paper, we present the *Fitzpatrick 17k* dataset which is a collection of images from two online dermatology atlases annotated with Fitzpatrick skin types by a team of humans. We train a deep neural network to classify skin conditions solely from images, and we evaluate accuracy across skin types.

We also use the *Fitzpatrick 17k* dataset to compare Fitzpatrick skin type labels to a computational method for estimating skin tone: individual typology angle (ITA). ITA is promising because it can be computed directly from images, but its performance varies with lighting conditions and may not always be effective for accurately annotating clinical images with skin types [332, 344, 640].

4.2 Fitzpatrick 17k Dataset

The *Fitzpatrick 17k* dataset contains 16,577 clinical images with skin condition labels and skin type labels based on the Fitzpatrick scoring system [209]. The dataset is accessible at https://github.com/mattgroh/fitzpatrick17k.

The images are sourced from two online open-source dermatology atlases: 12,672 images from DermaAmin and 3,905 images from Atlas Dermatologico [16, 215]. These sources include images and their corresponding skin condition label. While these labels are not known to be confirmed by a biopsy, these images and their skin condition labels have been used and cited in dermatology and computer vision literature a number of times [30, 70, 187, 262, 521, 593, 618]. As a data quality check, we asked a board-certified dermatologist to evaluate the diagnostic accuracy of 3% of the dataset. Based on a random sample of 504 images, a board-certified dermatologist identified 69.0% of images as diagnostic of the labeled condition, 19.2% of images as potentially diagnostic (not clearly diagnostic but not necessarily mislabeled, further testing would be required), 6.3% as characteristic (resembling the appearance of such a condition but not clearly diagnostic), 3.4% are considered wrongly labeled, and 2.0% are labeled as other. A second board-certified dermatologist also examined this sample of images and confirmed the error rate. This error rate is consistent with the 3.4% average error rate in the most commonly used test datasets for computer vision, natural language processing, and audio processing [465].

We selected images to annotate based on the most common dermatology conditions across these two data sources excluding the following 22 categories of skin conditions: (1) viral diseases, HPV, herpes, molluscum, exanthems, and others (2) fungal infections, (3) bacterial infections, (4) acquired autoimmune bullous disease, (5) mycobacterial infection (6) benign vascular lesions (7) scarring alopecia, (8) non-scarring alopecia (9) keratoderma (10) ichthyosis, (11) vasculitis, (12) pellagra like eruption (13) reiters disease (14) epidermolysis bullosa pruriginosa (15) amyloidosis, (16) pernio and mimics (17) skin metastases of tumours of internal organs (18) erythrokeratodermia progressive symmetric, (19) epidermolytic hyperkeratosis, (20) infections, (21) generalized eruptive histiocytoma, (21) dry skin eczema. We excluded these categories because they were either too broad, the images were of poor quality, or the categories represented a rare genodermatosis. The final sample includes 114 conditions with at least 53 images (and a maximum of 653 images) per skin condition.

This dataset also includes two additional aggregated levels of skin condition classification based on the skin lesion taxonomy developed by Esteva et al. 2017, which can be helpful to improve the explainability of a deep learning system in dermatology [48, 187]. At the highest level, skin conditions are split into three categories: 2,234 benign lesions, 2,263 malignant lesions, and 12,080 non-neoplastic lesions. At a slightly more granular level, images of skin conditions are split into nine categories: 10,886 images labeled inflammatory, 1,352 malignant epidermal, 1,194 genodermatoses, 1,067 benign dermal, 931 benign epidermal, 573 malignant melanoma, 236 benign melanocyte, 182 malignant cutaneous lymphoma, and 156 malignant dermal. At the most granular level, images are labeled by skin condition.

The images are annotated with Fitzpatrick skin type labels by a team of human annotators from Scale AI. The Fitzpatrick labeling system is a six-point scale originally developed for classifying sun reactivity of skin and adjusting clinical medicine according to skin phenotype [209]. Recently, the Fitzpatrick scale has been used in computer vision for evaluating algorithmic fairness and model accuracy across skin type [87, 174, 388]. Fitzpatrick labels allow us to begin assessing algorithmic fairness, but we note that the Fitzpatrick scale does not capture the full diversity of skin types [632]. Each image is annotated with a Fitzpatrick skin type label by two to five annotators based on Scale AI's dynamic consensus process.

The number of annotators per image is determined by a minimal threshold for agreement, which takes into account an annotator's historical accuracy evaluated against a gold standard dataset, which consists of 312 images with Fitzpatrick skin type annotations provided by a board-certified dermatologist. This annotation process resulted in 72,277 annotations in total.

In the *Fitzpatrick 17k* dataset, there are significantly more images of light skin types than dark skin. There are 7,755 images of the lightest skin types (1 & 2), 6,089 images of the middle skin types (3 & 4), and 2,168 images of the darkest skin types (5 & 6). Table 4.1 presents the distribution of images by skin type for each of the three highest level categorizations of skin conditions. A small portion of the dataset (565 images) are labeled as unknown, which indicates that the team of annotators could not reasonably identify the skin type within the image.

The imbalance of skin types across images is paired with an imbalance of skin types across skin condition labels. The *Fitzpatrick 17k* dataset has at least one image of all 114 skin conditions for Fitzpatrick skin types 1 through 3. For the remaining Fitpatrick skin types, there are 113 skin conditions represented in type 4, 112 represented in type 5, and 89 represented in type 6. In other words, 25 of the 114 skin conditions in this dataset have no examples in Fitzparick type 6 skin. The mean Fitzpatrick skin types across these skin condition labels ranges from 1.77 for basal cell carcionma morpheaform to 4.25 for pityriasis rubra pilaris. Only 10 skin conditions have a mean Fitzpatrick skin type above 3.5, which is the expected mean for a balanced dataset across Fitzpatrick skin types. These 10 conditions include: pityriasis rubra pilaris, xeroderma pigmentosum, vitiligo, neurofibromatosis, lichen amyloidosis, confluent and reticulated papillomatosis, acanthosis nigricans, prurigo nodularis, lichen simplex, and erythema elevatum diutinum.

	Non-Neoplastic	Benign	Malignant
# Images	12,080	2,234	2,263
Type 1	17.0%	19.9%	20.2%
Type 2	28.1%	30.0%	32.8%
Type 3	19.7%	21.2%	20.2%
Type 4	17.5%	16.4%	13.3%
Type 5	10.1%	7.1%	6.5%
Type 6	4.4%	2.0%	2.7%
Unknown	3.2%	3.3%	4.6%

Table 4.1: Distribution of skin conditions in *Fitzpatrick 17k* by Fitzpatrick skin type and high level skin condition categorization.

	Accuracy	Accuracy (off-by-one)	# of Images
Type 1	49%	79%	10
Type 2	38%	84%	100
Type 3	25%	71%	98
Type 4	26%	71%	47
Type 5	34%	85%	44
Type 6	59%	83%	13

Table 4.2: Accuracy of human annotators relative to the gold standard dataset of 312 Fitzpatrick skin type annotations provided by a board-certified dermatologist.

4.3 Classifying Skin Conditions with a Deep Neural Network

4.3.1 Methodology

We train a transfer learning model based on a VGG-16 deep neural network architecture [572] pre-trained on ImageNet [153]. We replace the last fully connected 1000 unit layer with the following sequence of layers: a fully connected 256 unit layer, a ReLU layer, dropout layer with a 40% change of dropping, a layer with the number of predicted categories, and finally a softmax layer. As a result, the model has 135,335,076 total parameters of which 1,074,532 are trainable. We train the model by using the Adam optimization algorithm to minimize negative log likelihood loss. We address class imbalance by using a weighted random sampler where the weights are determined by each skin condition's inverse frequency in the dataset. We perform a number of transformations to images before training the model which include: randomly resizing images to 256 pixels by 256 pixels, randomly rotating images 0 to 15 degrees, randomly altering the brightness, contrast, saturation, and hue of each image, randomly flipping the image horizontally or not, center cropping the image to be 224 pixels by 224 pixels, and normalizing the image arrays by the ImageNet means and standard deviations.

We evaluate the classifier's performance via 5 approaches: (1) testing on the subset of images labeled by a board-certified dermatologist as diagnostic of the labeled condition and training on the rest of the data (2) testing on a randomly selected 20% of the images where the random selection was stratified on skin conditions and training on the rest of the data (3) testing on images from Atlas Dermatologico and training on images from Derma Amin (4) testing on images from Derma Amin and training on images from Atlas Dermatologico (5) training on images labeled as Fitzpatrick skin types 1-2 (or 3-4 or 5-6) and testing on the rest of the data. The accuracy on the validation set begins to flatten after 10 to 20 epochs for each validation fold. We trained the same architecture on each fold and report accuracy scores for the epoch with the lowest loss on the validation set.

Holdout Set	Verified	Random	Source A	Source B	Fitz 3-6	Fitz 1-2 & 5-6	Fitz 1-4
# Train Images	16,229	12,751	12,672	3,905	7,755	6,089	2,168
# Test Images	348	3,826	3,905	12,672	8,257	10,488	14,409
Overall	26.7%	20.2%	27.4%	11.4%	13.8%	13.4%	7.7%
Type 1	15.1%	15.8%	40.1%	6.6%	-	10.0%	4.4%
Type 2	23.9%	16.9%	27.7%	8.6%	-	13.0%	5.5%
Type 3	27.9%	22.2%	25.3%	13.7%	15.9%	-	9.1%
Type 4	30.9%	24.1%	26.2%	17.1%	14.2%	-	12.9%
Type 5	37.2%	28.9%	28.4%	17.6%	10.1%	21.1%	-
Type 6	28.2%	15.5%	25.7%	14.9%	9.0%	12.1%	-

Table 4.3: Accuracy rates classifying 114 skin conditions across skin types on six selections of holdout sets. The verified holdout set is a subset of a randomly sampled set of images verified by a board-certified dermatologist as diagnostic of the labeled condition. The random holdout set is a randomly sampled set of images. The source A holdout set are all images from Atlas Dermatologico. The source B holdout set are all images from Derma Amin. The 3 Fitzpatrick holdout sets are selected according to Fitzpatrick labels. In all cases, the training data are the remaining non-held out images from the *Fitzpatrick 17k* dataset.

		Predicted Class		
		Benign	Malignant	Non-neoplastic
Actual Class	Benign	275	52	54
	Malignant	106	487	109
	Non-neoplastic	788	448	1586

Table 4.4: Confusion matrix for deep neural network performance on predicting the high-level skin condition categories in the holdout set of images from Atlas Dermatologico.

4.3.2 Results

We report results of training the model on all 114 skin conditions across 7 different selections of holdout sets in Table 4.3.

In the random holdout, the model produces a 20.2% overall accuracy on exactly identifying the labeled skin condition present in the image. The top-2 accuracy (the rate at which the first or second prediction of the model is the same as the image's label) is 29.0% and the top-3 accuracy is 35.4%. These numbers can be evaluated against random guessing, which would be 1/114 or 0.9% accuracy. Across the 114 skin conditions, the median accuracy is 20.0% and ranges from a minimum of 0% accuracy on 10 conditions (433 images in the random holdout) and a maximum of 93.3% accuracy on 1 condition (30 images).

When we train the model on the 3 category partition of non-neoplastic, benign, and malignant, the model produces an accuracy of 62.4% on the random holdout (random guessing would produce 33.3% accuracy). Likewise, the model trained on the 9 category partition produces an accuracy of 36.1% on the random holdout (random guessing would produce 11.1% accuracy). Another benchmark for this 3 partition and 9 partition comes from Esteva et al. which trained a model on a dataset 7.5 times larger to produce 72.1% accuracy on the 3 category task and 55.4% accuracy on the 9 category task [187].

Depending on each holdout selection, the accuracy rates produced by the model vary across skin types. For the first four holdout selections in Table 4.3 – the verified selection, the random holdout, the source A holdout based on images from Atlas Dermatologico, and the source B holdout based on images from Derma Amin – we do not find a systematic pattern in accuracy scores across skin type. For the second three holdout selections where the model is trained on images from two Fitzpatrick types and evaluated on images in the other four Fitzpatrick types, we find the model is most accurate on the images with the closest Fitzpatrick skin types to the training images. Specifically, the model trained on images labeled as Fitzpatrick skin types 1 and 2 performed better on types 3 and 4 than types 5 and 6. Likewise, the model trained on types 3 and 4 performed better on types 2 and 5 than 1 and 6. Finally, the model trained on types 5 and 6 performed better on types 3 and 4 than types 1 and 2.

4.4 Evaluating Individual Typology Angle against Fitzpatrick Skin Type Labels

4.4.1 Methodology

An alternative approach to annotating images with Fitzpatrick labels is estimating skin tone via individual typology angle (ITA), which is calculated based on statistical features of image pixels and is negatively correlated with the melanin index [640]. Ideally, ITA is calculated over pixels in a segmented region highlighting only non-diseased skin [332]. But, segmentation masks are expensive to obtain, and instead of directly segmenting healthy skin, we apply the YCbCr algorithm to mask skin pixels [344]. We compare Fitzpatrick labels on the entire dataset with ITA calculated on the full images and the YCbCr masks.

The YCbCr algorithm takes as input an image in RGBA color space and applies the following masking thresholds.

$$R > 95 \tag{4.1}$$

$$R > G \tag{4.2}$$

$$R > B \tag{4.3}$$

$$G > 40 \tag{4.4}$$

$$B > 20 \tag{4.5}$$

$$|R - G| > 15 \tag{4.6}$$

$$A > 15 \tag{4.7}$$

Then, the image is converted from RGBA to YCbCr color space, and applies a further masking along the following thresholds:

$$Cr > 135 \tag{4.8}$$

$$Cr \geq (0.3448 \cdot Cb) + 76.2069 \tag{4.9}$$

$$Cr \geq (-4.5652 \cdot Cb) + 234.5652 \tag{4.10}$$

$$Cr \leq (-1.15 \cdot Cb) + 301.75 \tag{4.11}$$

$$Cr \leq (-2.2857 \cdot Cb) + 432.85 \tag{4.12}$$

where $R - G - B - A$ are the respective Red-Green-Blue-Alpha components of the input image, and $Y - Cb - Cr$ are the respective luminance and chrominance components of the color-converted image. As a result, the YCbCr algorithm attempts to segment healthy skin from the rest of an image.

We calculate the ITA of each full and YCbCr masked image by converting the input image to $CIE - LAB$ color space, which contains L: luminance and B: yellow, and applying the following formula [423]:

$$ITA = arctan(\frac{L^* - 50}{B^*}) \cdot \frac{180}{\pi} \tag{4.13}$$

where L^* and B^* are the mean of non-masked pixels with values within one standard deviation of the actual mean.

4.4.2 Results

In Table 4.5, we compare ITA calculations on both the full images and YCbCr masks with Fitzpatrick skin type labels. Furthermore, we compare two different methods for calculating Fitzpatrick type given ITA, as described in Equations 4.14 and 4.15. For each entry, we calculate the proportion of ITA scores in the range of plus or minus one of the annotated Fitzpatrick score.

$$Fitzpatrick(ITA) = \begin{cases} 1 & ITA > 55 \\ 2 & 55 \geq ITA > 41 \\ 3 & 41 \geq ITA > 28 \\ 4 & 28 \geq ITA > 19 \\ 5 & 19 \geq ITA > 10 \\ 6 & 10 \geq ITA \end{cases} \tag{4.14}$$

$$Fitzpatrick(ITA) = \begin{cases} 1 & ITA > 40 \\ 2 & 40 \geq ITA > 23 \\ 3 & 23 \geq ITA > 12 \\ 4 & 12 \geq ITA > 0 \\ 5 & 0 \geq ITA > -25 \\ 6 & -25 \geq ITA \end{cases} \tag{4.15}$$

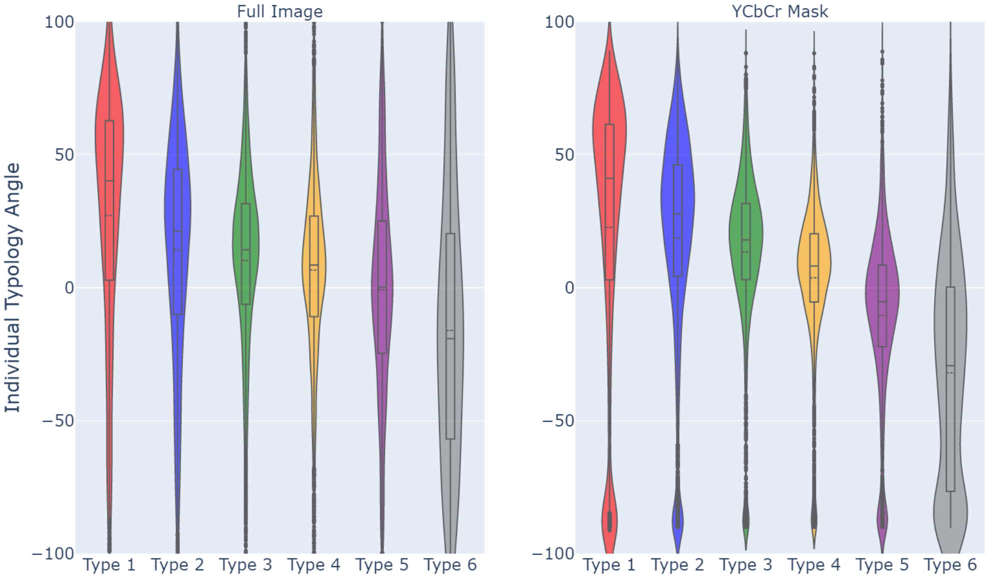


Figure 4-1: Observed distribution of individual typology angles by Fitzpatrick.

In Table 4.5, the columns labeled "Kinyananjui" compare Fitzpatrick skin type labels with ITA following Equation 4.14, a modified version of the ITA thresholds described by Kinyanjui et al. [332]. The columns labeled "Empirical" follow Equation 4.15, which we developed based on the empirical distribution of ITA scores minimizing overall error. Figure 4-1 plots the empirical distribution of ITA scores for each Fitzpatrick skin type label. The discrepancy between Fitzpatrick skin type labels and the ITA approach appears to be driven mostly by high variance in the ITA algorithm as Figure 4-3 reveals.

4.5 Conclusion

We present the *Fitzpatrick 17k*, a new dataset consisting of 16,577 clinical images of 114 different skin conditions annotated with Fitzpatrick skin type labels. These images are sourced from Atlas Dermatologico and Derma Amin and contain 3.6 times more images of the two lightest Fitzpatrick skin types than the two darkest Fitzpatrick skin types. By annotating this dataset with Fitzpatrick skin type labels, we reveal both an underrepresentation of dark

	Full Image		YCbCr Mask	
	Kinyanjui	Empirical	Kinyanjui	Empirical
Overall	45.87%	60.34%	53.30%	70.38%
Type 1	50.97%	65.35%	52.22%	66.00%
Type 2	42.60%	59.57%	49.15%	69.47%
Type 3	35.43%	55.20%	45.13%	66.41%
Type 4	34.09%	58.54%	40.24%	72.10%
Type 5	78.21%	65.49%	93.41%	82.26%
Type 6	74.80%	65.04%	90.71%	79.69%

Table 4.5: Plus or minus one concordance of individual typology angle (ITA) with Fitzpatrick skin type labels. Each column shows the percent of ITA scores that are within plus or minus 1 point of the annotated Fitzpatrick labels after converting ITA to Fitzpatrick types via Equations 4.14 and 4.15.

skin images in online dermatology atlases and accuracy disparities that arise from training a neural network on only a subset of skin types.

By training a deep neural network based on an adapted VGG-16 architecture pre-trained on ImageNet, we achieve accuracy results that approach the levels reported on a much larger dataset [187]. We find that the skin type in the images on which a model is trained affects the accuracy scores across Fitzpatrick skin types. Specifically, we find that models trained on data from only two Fitzpatrick skin types are most accurate on holdout images of the closest Fitzpatrick skin types to the training data. These relationships between the type of training data and holdout accuracy across skin types are consistent with what has been long known by dermatologists: skin conditions appear differently across skin types [9].

An open question for future research is in which skin conditions do accuracy disparities appear largest across skin types. Recent research shows that diagnoses by medical students and physicians appears to vary across skin types [161, 198]. Future research at the intersection of dermatology and computer vision should focus on specific groups of skin conditions where accuracy disparities are expected to arise because visual features of skin conditions (e.g. redness in inflammatory conditions) do not appear universally across skin types.

The large set of Fitzpatrick skin type labels enable an empirical evaluation of ITA as an automated tool for assessing skin tone. Our comparison reveals that ITA is prone to error

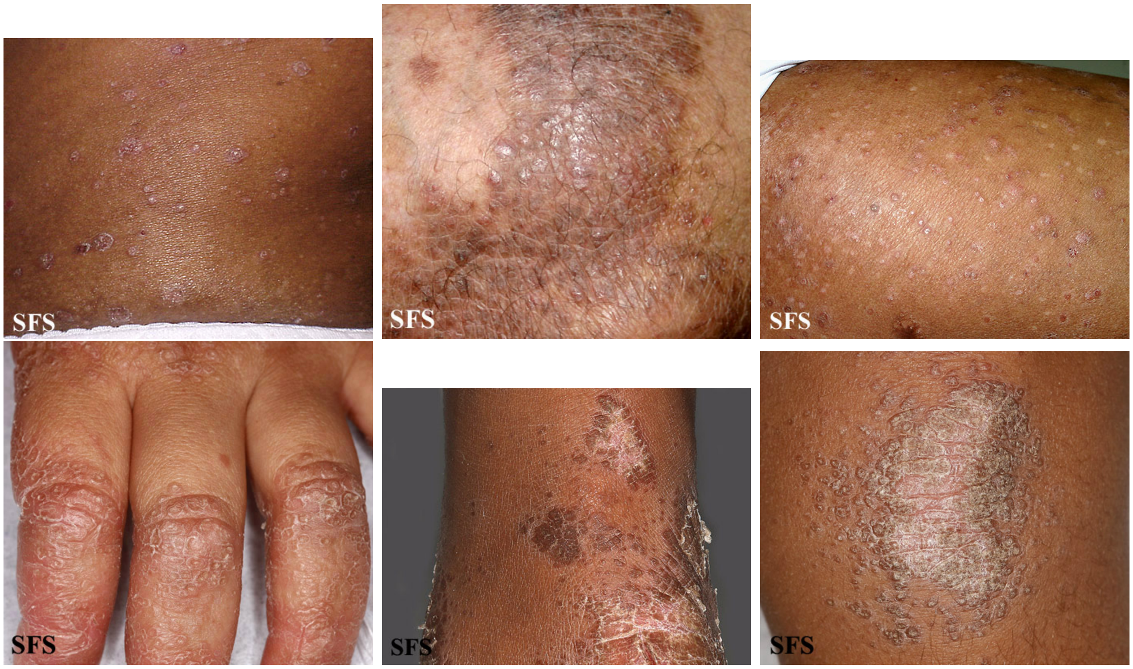


Figure 4-2: Example images of pityriasis rubra pilaris from Atlas Dermatologico that were accurately classified by the neural network trained on DermaAmin images. On the 174 images from Atlas Dermatologico labeled pityriasis rubra pilaris, 24% are accurately identified, 35% are accurately identified in the top 2 most likely predictions, and 45% are accurately identified in the top 3 most likely predictions.

on images that human labelers can easily agree upon. The most accurate ITA scores are off by more than one point on the Fitzpatrick scale in about one third of the dataset. One limitation of this comparison is that we calculated ITA based on either the entire image or an automatic segmentation mask. Future work should refine this comparison based on more precise segmentation masks.

We present this dataset and paper in the hopes that it inspires future research at the intersection of dermatology and computer vision to evaluate accuracy across sub-populations where classification accuracy is suspected to be heterogeneous.

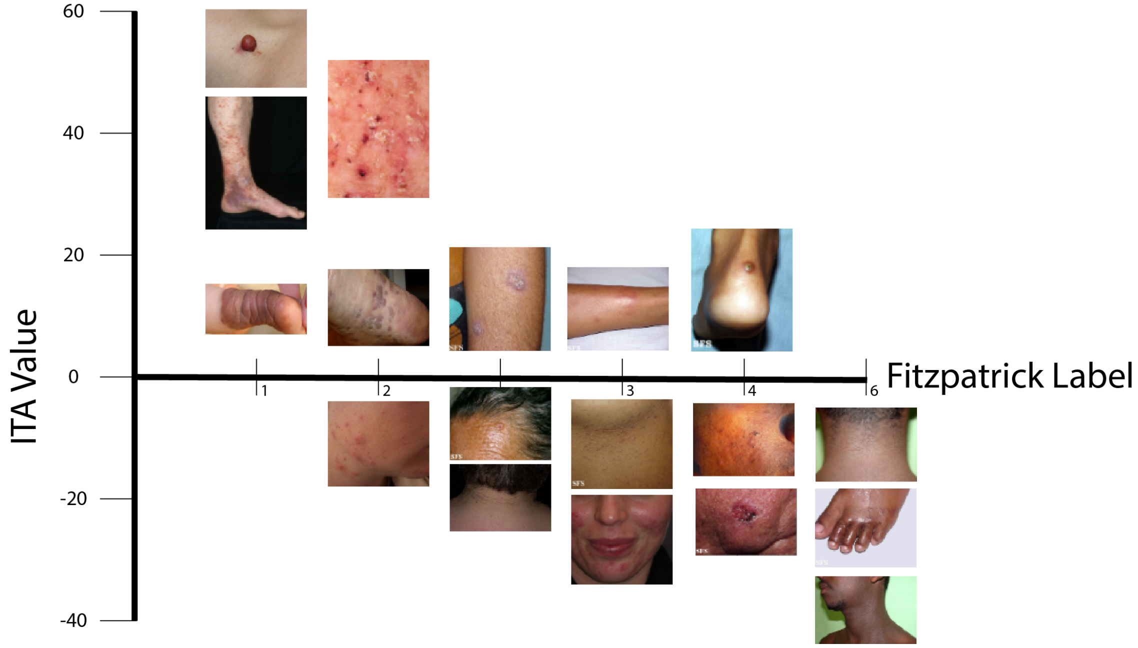


Figure 4-3: 18 images plot arranged based on ITA values and Fitzpatrick labels.

4.6 Data and Code Availability

The datasets and code generated and analyzed during the current study are available in our public Github repository, `https://github.com/mattgroh/fitzpatrick17k`.

Chapter 5

Towards Transparency in Dermatology Image Datasets with Experts, Crowds, and an Algorithm

Abstract

While artificial intelligence (AI) holds promise for supporting healthcare providers and improving the accuracy of medical diagnoses, a lack of transparency in the composition of datasets exposes AI models to the possibility of unintentional and avoidable mistakes. In particular, public and private image datasets of dermatological conditions rarely include information on skin color. As a start towards increasing transparency, AI researchers have appropriated the use of the Fitzpatrick skin type (FST) from a measure of patient photosensitivity to a measure for estimating skin tone in algorithmic audits of computer vision applications including facial recognition and dermatology diagnosis. In order to understand the variability of estimated FST annotations on images, we compare several FST annotation methods on a diverse set of 460 images of skin conditions from both textbooks and online dermatology atlases. These methods include expert annotation by board-certified dermatologists, algorithmic annotation via the Individual Typology Angle algorithm, which is then converted to estimated FST (ITA-FST), and two crowd-sourced, dynamic consensus protocols for annotating estimated FSTs. We find the inter-rater reliability between three board-certified dermatologists is comparable to the inter-rater reliability between the board-certified dermatologists and either of the crowdsourcing methods. In contrast, we find that the ITA-FST method produces annotations that are significantly less correlated with the experts' annotations than the experts' annotations are correlated with each other.

These results demonstrate that algorithms based on ITA-FST are not reliable for annotating large-scale image datasets, but human-centered, crowd-based protocols can reliably add skin type transparency to dermatology datasets. Furthermore, we introduce the concept of dynamic consensus protocols with tunable parameters including expert review that increase the visibility of crowdwork and provide guidance for future crowdsourced annotations of large image datasets.[1]

5.1 Motivation

Artificial intelligence (AI) algorithms hold promise for improving image-based clinical diagnosis tasks ranging from identifying breast cancer in mammograms [418] to classifying skin lesions based on a single image [187] to predicting the diagnosis of hundreds of diverse skin conditions based on a few images and a brief patient history [388]. The combination of algorithmic predictions with physician diagnostic skill has the potential to create large efficiency and welfare gains in healthcare [530]. In particular, AI systems can enhance specialists' diagnostic performance on specific tasks (e.g. identifying pneumonia on chest radiographs [487] and predicting hypoxaemia risk from operating room data [395]) but incorrect predictions from an AI system can mislead specialists and generalists alike [243, 299, 608]. In fact, inaccurate advice regardless of whether it comes from an AI or human tends to decrease physicians' accuracy on diagnostic tasks [221]. Moreover, the algorithm appreciation effect [389] suggests that inaccurate advice from an algorithm is likely to have more negative effects than the same advice given by a human.

Given the consequences of inaccurate advice in healthcare, ethical and responsible algorithm-in-the-loop decision systems should require the systems to be both accurate and also unbiased with regard to sensitive attributes like race and gender. Moreover, these systems should be transparent such that medical experts can reliably assess algorithmic performance [238, 256]. These principles for ethical systems are particularly important because algorithms are prone to make unexpected errors on out-of-distribution data. Due to biases in dataset representation, protected classes are more likely to be out-of-distribution [17, 49, 386]. Moreover,

[1] This chapter, which is co-authored by Caleb Harris, Roxana Daneshjou, Arash Koochek, and Omar Badri, appeared in the proceedings for the Computer Supported Collaborative Work (CSCW) 2022 [246].

when accurate yet non-equitable algorithmic risk assessments are used as decision support tools they have been shown to alter decision maker's risk aversion and lead to unexpected and sometimes unwanted shifts in human decision-making [239].

Yet the vast majority of AI algorithms for diagnostic tasks in dermatology are trained on datasets that lack transparency with regards to demographic and skin tone attributes [142, 635]. Due to this lack of transparency, it is difficult to assess what data may be out-of-distribution and this leads to the potential for unexpected errors that could have otherwise been addressed. For example, given the under representation of dark skin in educational resources [9, 23, 179, 378, 391] and online dermatology atlases [248], it is unknown the full extent to which dark skin is under-represented in many of the large dermatology image datasets. For the few datasets that do include skin type information, dark skin types are underrepresented [142]. This is particularly problematic because AI algorithms for classifying the skin condition in an image are more accurate on images that match the skin color upon which the algorithm is trained than images that do not match the skin color [248]. An analysis of three AI algorithms (ModelDerm [263], DeepDerm [187], and Ham 1000 [607]) reveal that images of dark skin show a drop in all three models' accuracy rates relative to rates in images of light skin [145].

One approach for increasing transparency in dermatology image datasets and their resulting AI algorithms is to annotate skin tone with Fitzpatrick Skin Type (FST) like the algorithmic audit of accuracy disparities in facial recognition by Buolamwini and Gebru 2018 [87]. FST is a clinical measurement developed and used by dermatologists to assess patients' sun sensitivity for dosing phototherapy or chemophototherapy. Clinical FST has been criticized for subjectivity [254], is not designed for classifying race or skin color [632], and often involves not just assessing skin tone but assessing a patient's hair color and eye color. Despite the imperfections and biases of clinical FST as a proxy for skin tone and an assessment of differential healthcare risks [472], AI researchers have appropriated FST to estimate skin tone labels for algorithmic audits of tasks like classifying skin disease [174, 248, 388, 498] and facial recognition algorithms [87, 272]. In this paper, we distinguish FST as recorded in a clinical patient-provider interaction as "clinical FST" and FST as recorded based on a single image as “estimated FST.”

While estimated FST has been frequently used in computer vision tasks, basic questions have not been explored about its use for labeling image datasets: Who is qualified to annotate images with estimated FSTs? More specifically, should estimated FST annotations on large-scale datasets be limited to board-certified dermatologists? How concordant are board-certified dermatologists, particularly on the kinds of datasets used in computer vision? Would the annotations of trained annotators, crowdsourced labor, or algorithms differ significantly from board-certified dermatologists? These are empirical questions, which are connected more broadly to questions about what makes desirable data and how race and gender should be annotated in image datasets [553, 554]. Notably, we limit the focus on estimated FST because it is a method used in algorithmic audits based on clinical medicine and it allows granular analysis which would not be captured by race alone [87]. While most large image datasets with estimated FST annotations are labeled by dermatologists [87, 174, 388, 498], the “Casual Conversations” dataset is annotated by trained annotators [272] and the “Fitzpatrick 17k” annotations are generated by applying a dynamic consensus protocol to crowdsourced annotations [248].

As an alternative to human-annotated estimated FST, researchers have proposed and used the Individual Typology Angle to FST (ITA-FST) algorithm, a computer vision algorithm that converts the RGB values of an image into a single metric for constitutive pigmentation, to estimate apparent skin tone from images [332, 352]. Prior work shows that ITA-FST is strongly correlated with Melanin Index [640], which is sometimes used in assigning clinical FSTs [182]. However, recent research in photodermatology suggests that ITA used for constitutive pigmentation is a poor proxy for clinical FST [477].

Prior work suggests that crowdsourced estimated FST annotations are generally within 1 point of an expert board-certified dermatologist’s annotation, but Groh et al (2021) compared crowd annotations with only a single expert, do not include statistical analyses of inter-rater reliability, do not compare ITA with experts’ annotations, and do not examine nuances around the compositions of the crowd or edge cases where the crowd is prone to err [248]. We present evidence that the inter-rater reliability between three board-certified dermatologists is comparable to the inter-rater reliability between board-certified dermatologists and crowdsourcing methods but not the ITA-FST algorithm. However, for a subset of images with high disagreement between crowd annotators, we find higher inter-rater reliability between board-certified dermatologists than board-certified dermatologists and the crowd.

In summary, our contributions are the following:

(1) We evaluate the inter-rater reliability between three medical experts, a computer vision algorithm, and two crowdsourcing approaches for annotating images of skin conditions with estimated FST, which is useful for increasing transparency into how algorithms perform on images of different skin tones. On a set of 320 images drawn from dermatology textbooks [73, 241, 255, 298, 309, 311, 430, 648], we do not find a statistically significant difference when comparing the Pearson Correlation Coefficients (ρ) between three medical experts with the ρ between each medical expert and either of the crowdsourcing methods. In contrast, we do find a statistically significant difference in the annotations produced by the ITA-FST algorithm. These results suggest that crowdsourcing (but not the ITA-FST algorithm) can be a reliable source for generating estimated FST annotations on large-scale datasets of images intended for training and evaluating AI models to classify skin disease. However, we include important caveats. First, our qualitative results show the crowd will sometimes make errors that the medical experts would be unlikely to make. Second, a quantitative follow-up with 140 images drawn from two online dermatology atlases [16, 215] shows the results are robust to 70 images randomly sampled from the 91% of images with relatively low crowd disagreement but on a random sample of 70 images from the 9% of images with relatively high crowd disagreement, we find the crowd annotations can be significantly different from the experts' annotations. Third, the image-based estimated FST annotations are subject to lighting, image quality, and pose variability that are not an issue for in-person assessments

(2) In order to increase visibility into the process of human annotation of large image datasets and guide future work, we introduce and describe a dynamic consensus protocol for aggregating crowdsourced estimated FST annotations using the following transparent, adjustable criteria: (1) **consensus thresholds**, (2) **qualified annotations**, (3) **failure reports** [90], (4) **agreement metrics** and (5) **expert review**. We apply this procedure to the publicly available "Fitzpatrick 17k" dataset of 16,577 images to evaluate inter-rater reliability across crowdsourcing annotation methods, estimate the proportion of images that experts should review, and conduct expert review on 140 images.

5.2 Background and Related Work

5.2.1 Data Documentation for Increasing Transparency and Accountability in Algorithms

Critical frameworks documenting both machine learning datasets and their resulting models promote transparency and accountability by enabling nuanced analyses that can expose unwanted biases. Examples of guiding frameworks for detailing data, its definitions, and its associated models' potential harms include *Data Statements for Natural Language Processing*, *The Dataset Nutrition Label* [279], *Model Cards for Model Reporting* [433], and *Datasheets for Datasets* [222]. The seminal algorithmic audit of accuracy disparities in facial recognition by Buolamwini and Gebru 2018 relied on documenting estimated FST annotations and evaluating algorithmic performance across FSTs and found significant intersectional accuracy disparities [87]. Estimated FST annotations have also been helpful in documenting accuracy in machine learning models for dermatology [145, 248]. With appropriate, inclusive data, algorithms can increase accountability by both serving as a diagnostic tool to detect discrimination and formalizing our definitions around a social problem like inequities in healthcare across gender, race, and skin color [4, 337].

Beyond cataloguing the elements of a dataset, data documentation can also question the existence of categories within the data and inform the question posed by Miceli et al 2022: "Is this information sufficient in itself to explicate unjust outcomes" [427]? For a large number of datasets with images of humans, the definitions of both race and gender in databases lack critical engagements, are overly reductive, and require more than an outside observer looking at a photograph to annotate appropriately [429, 554]. This is particularly problematic because the definition of a category, class, or outcome will impact how disparate treatment and disparate impact arise in the data [49]. For example, Obermeyer et al 2019 report that an algorithm for predicting health risk of millions of people in the United States using cost of care as a proxy for health needs led to the following bias: "At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses" [470]. This racial bias in a healthcare setting is not only a problem of selecting the right outcome measure, but a deeper problem that involves a history in the United States of "segregated hospital facilities, racist medical curricula, and unequal insurance structures, among other factors" [59]. Dataset documentation is an initial step that enables critical researchers to both identify empirical biases, question the definitions of specific data features, and inspect the data generating process. Data documentation is particularly helpful to bridge collaboration between data scientists and subject matter experts to make knowledge and processes explicit such that both groups of people can ask the right question [402]. As bias is uncovered in data, researchers can offer new insights into bias as a starting point for "studying up" [452] with a critical focus on accountability and power dynamics in the underlying data generation process [45].

Extracting categories and clusters from complex data involves value judgments. For example, Scheuerman et al 2021 highlight the tensions in the development of computer vision datasets between efficiency and care, universality and contextuality, impartiality and positionality, and model work and and data work [553]. In crowd sourcing tasks, categorizing can become problematic when crowdworkers have limited attention and expertise [25] and when crowdworkers are overly constrained by power dynamics such that the crowd annotates data based on their expectations of how a client sees the world rather than their own sense of how the world looks [428]. One recent applied example from CSCW shows that accessible interfaces with high degrees of freedom enable crowdworkers to categorize data that can appropriately filter harmful content generated by AI [401]. Another example from recent research in CSCW highlights the potential for failure reports [90] – open-ended descriptions of model errors – to help navigate unexpected systematic failures. We expand on the concept of failure reports in Section 3.3 where crowdworkers can transcend the menu of multiple choice annotations (in our case estimated FST I-VI and "not applicable") to free-text responses where images can be flagged for being incorrectly labeled, inappropriate, or irrelevant.

In dataset documentation, epistemic authority is an important value judgment. How should data be annotated and who or what should do it? Data annotation is often less straightforward and more complex than it seems. Data work is often time-consuming, opaque (unless there's good documentation), and not well rewarded; in field interviews with data workers, one interviewee exclaimed, "Everyone wants to do the model work, not the data work," which is a sentiment shared by many interviewees [549]. Moreover, reasonable people often disagree on color classification [234] and medical experts often disagree on medical diagnoses [515, 520]. In fact, in one study comparing referral and final diagnoses across 280 patients, significant disagreements appear in 21% of cases [617]. Instead of assigning epistemic authority to any particular individual or algorithm for a subjective task, we follow prior work that treats epistemic authority based on inter-annotator agreement [124, 207, 208, 283]. Disagreement between annotators is not necessarily indicative of poor quality annotations or bias, but instead, disagreement can help reveal the subjectivity involved in a particular task and a particular example [34].

In order to answer who or what is epistemically qualified to annotate data with information that can provide transparency and accountability into potential biases, we need to examine the level of agreement produced by different methods. The first step involves measuring the subjectivity of the task by measuring the degree of disagreement among experts. Next, we compare alternative methods (e.g. an algorithm and crowd methods) to the level of disagreement among experts. If an alternative method does not disagree significantly more with experts than experts do with each other, then we can call the alternative method generally comparable to experts. While an alternative method may be generally comparable to experts, edge cases may arise where experts have significantly lower disagreement among themselves than with the crowd. For example, in the case of estimated FST images where a rare skin disease has transformed the color of the skin, non-experts may have higher levels of disagreement than experts. In the framework of Muller et al 2019 *How Data Science*

Workers Work with Data [446], this annotation process would be described as "Ground Truth as Created" where human expertise applied to images informs an analysis of the similarity of various methods for annotating estimated FST. While Muller et al 2021 [447] write that "it is widely agreed that SME-labeled data [data labeled by subject matter experts] is the "gold standard" data source for high quality labeled data for specialized tasks," we take a step back from this assumption and empirically evaluate how well crowds and an algorithm compare to estimated FST annotations of images by board-certified dermatologists. The dynamic consensus threshold process described in 3.3 can represent both Muller et al 2021's "Principled design" and "Iterative design" to ground truth annotation because the annotation process is planned and well-defined, but aspects of the dynamic consensus threshold process (e.g. failure reports and expert review) allow for clarifications, adjustments, and potential redefinitions based on collaboration back and forth between the human annotators examining data at the record level and data scientists examining the data at the dataset level.

5.2.2 Designing Transparency into Clinical Decision Support Systems

Clinical decision support systems (CDSSs) are systems designed to support healthcare providers in medical decision-making. Past work at CSCW has documented the following relevant onboarding criteria for healthcare providers to interact with CDSSs: capabilities and limitations, functionality, design objective, relative strengths and weaknesses of an algorithm, performance of a model on domain specific cases including how the model's idiosyncrasies compare with human idiosyncrasies [91, 92]. Transparency on subgroups within the data is an integral component to onboarding healthcare providers such that they develop an understanding for when they should override the system, which is important when an algorithm makes an erroneous prediction [35]. In practice, healthcare professionals seek to compare algorithmic errors in CDSSs with their own errors [93]. Well-documented data enables healthcare providers (and researchers) to examine subcategories on which algorithms are likely to error, which is important for establishing trust that the algorithm will lead to positive results for vulnerable patients [181, 621] and useful for identifying what kind of data should be collected to reduce accuracy disparities [112].

Recent deployments of deep learning systems for health reveal that algorithms trained on retrospective datasets may not be ecologically valid [56]. In particular, an algorithm applied to data that deviates from the training data is prone to unexpected errors on the out-of-distribution examples. One approach to handling out-of-distribution data is training a classifier to predict whether an image is out-of-distribution and if it is to abstain from generating an algorithmic classification [537]. Decision support systems tend to be less effective on out-of-distribution examples; in evaluations of algorithms that generally outperform humans, the performance gap in accuracy between humans informed by the algorithm and the algorithm is lower in out-of-distribution examples than in-distribution examples [386]. If a particular skin tone is out-of-distribution for a particular disease, this is important for clinicians and model developers to know, so they are aware what kind of data might constitute a context shift [242].

Recently, Jain et al 2021 completed a retrospective study that showed how a deep learning based CDSS may help non-specialists such as primary care physicians and nurse practitioners diagnose skin disease with higher accuracy (defined as agreement with reference conditions) and possibly reduce biopsy and referral rates to dermatologists than the providers would without the system [300]. Decision support systems have the potential to improve the quality of dermatological care, and as such, it is important to evaluate the underlying skin disease classification algorithm on diverse skin tones to address potential accuracy disparities given the context of skin tone and race in the United States healthcare system and computer vision applications. However the algorithm used in Jain et al 2021 was trained on only 46 images of FST VI skin and 510 images of FST V, which was 0.3% and 3.2% of the entire training set, respectively [388]. The lack of images of dark skin types in this dataset means this model may be prone to a higher level of unexpected algorithmic errors on future images of dark skin types [248].

5.3 Methods for Fitzpatrick Skin Type Annotations

The Fitzpatrick labeling system is a six-point clinical scale used by dermatologists for classifying skin types based on photo-reactivity of skin and was originally intended to be used for photochemotherapy [209]. See Table 5.3 in Appendix for a copy of the original description of the Fitzpatrick Scale. We note that the original scale does not include nuanced skin tones or color beyond white, brown, and black. While the Fitzpatrick scale is highly correlated with an individual's melanin index (measured by narrow-band spectrophotometric devices), the Fitzpatrick scale is a subjective measure [326]. In clinical practice, clinical FSTs are visually assessed by dermatologists based on the colors of a patient's skin, hair, and eyes and their history of sunburns [640]. In a study comparing self-reports to a single dermatologist's clinical FST determination, the dermatologist's assessment was found to be significantly more reliable than individuals' self reports [182]. Recently, researchers have used the estimated FST to annotate images and evaluate algorithmic fairness of AI models across apparent skin tones [87, 142, 248, 300, 388]. While the original Fitzpatrick scale was not designed to categorize skin color, it is often used as such in clinical practice [632] and it serves as a starting point (albeit imperfect given the coarseness of categories for skin color along sepia tones) for algorithmic audits [87].

5.3.1 Expert Labels from Board-Certified Dermatologists (N=3)

We asked three board-certified dermatologists – experts with deep experience examining skin conditions and assessing patients' clinical FST – to annotate images with the estimated FST. Each expert provided independent estimated FST annotations for 320 images collected from dermatology textbooks and 160 images collected from online dermatology Atlases. We collected 1,380 estimated FST annotations from experts. Given the inherent subjectivity of this task, we present ranges of experts' annotation across these images: 3-5% Type I, 28-31% Type II, 29-30% Type III, 14-15% Type IV, 14-15% Type V, and 4-9% Type VI. Likewise, the distribution across the the 160 images is 4-20% unknown, 0-3% Type I, 17-28% Type II, 26-34% Type III, 20-24% Type IV, 9-13% Type 5, and 1-8% Type 6.

The experts noted that estimated FSTs will not necessarily match in-person assessments because clinical FST relies on not just skin color but eye color, hair type and color, and history of sunburns. Moreover, clinical FST based on an in-person assessment considers an individual's entire body across varying lighting conditions while estimated FSTs based on a single image are restricted to a limited view of the body under a single lighting condition. As such, image-based estimated FST assessments will be have less information and be fundamentally more noisy with less inter-annotator agreement than clinical, in-person assessments. For example, clinical images of dermatological conditions differ in what part of the body is photographed, how the photograph is framed (from the camera's angle and zoom level to the patient's pose), how the lighting illuminates the image, and how the skin disease has transformed the patient's skin. We discuss further limitations of estimated FST annotations in the Limitations section.

5.3.2 Algorithmic Labels from Individual Typology Angle

Following computer vision papers using ITA-FST for algorithmic audits [332, 352], we compute ITA-FST annotations for each image. ITA was designed to classify skin color in a Caucasian population based on healthy skin in an image [109]. While ITA was not designed for all people, research shows ITA-FST correlates with both Melanin Index and clinical FST [151, 640]. However, ITA-FST and clinical FST are designed to measure constitutive pigmentation and sun-reactivity, respectively, and recent research suggests they are poor proxies of one another [477]. In order to calculate ITA-FST more precisely, researchers developed YCbCr masks to mask pixels outside a range of pre-specified colors [344] to reduce the noise of ITA-FST estimates. YCbCr masks are imperfect and often mask healthy skin or fail to mask non-skin parts of an image in the range of skin colors, but without YCbCr masks ITA-FST estimates are even more varied because very light or very dark backgrounds can influence the estimate. For example, YCbCr often fails to mask white underwear of dark skin people leading to the ITA-FST algorithm making errors in estimating skin tone that a reasonable human would not make.

We calculate ITA using the default D65 illuminant over the healthy skin pixels identified by YCbCr masks, and we convert the scores to estimated FSTs that minimize discrepancy between algorithmic labels and the experts' labels on the 320 textbook images and 140 images from dermatology atlases following procedures described by Groh et al 2021 and Krishnapriya et al 2022 [248, 352]. See Algorithm 2 in Appendix for details on transforming ITA scores to FSTs.

5.3.3 Dynamic Consensus Protocol for Crowd Labels (N>10,000 participants)

In order to crowdsource estimated FST annotations for images, we collaborated with Scale AI and Centaur Labs, two companies that specialize in labeling large image datasets via dynamic consensus protocols applied to crowdworkers' annotations. In this section, we identify five key components of a dynamic consensus protocol based on the process at Centaur Labs.

Dynamic consensus refers to the process of transforming multiple annotations from independent sources at different times on a single image into a consensus annotation. A dynamic consensus differs from a standard consensus metric like a mean, median, or mode because a dynamic consensus pre-specifies a **consensus threshold**, which must be met before annotations are transformed into accepted responses. For example, the annotations produced by Centaur Labs included a consensus threshold defined as either (a) a single category (across the 6 FSTs and a category for not applicable) has 3 more annotations than any other category or (b) the majority label if a consensus has not been reached after 20 annotations.

Qualified annotations are defined as annotations by individuals who have passed a task specific quality control procedure. In contrast, disqualified annotations are annotations by individuals who have failed the task specific quality control. The third category, non-qualified annotations, are annotations by individuals who have not yet been assessed by a task specific quality control procedure. In general, quality control is determined by the proportion of an individual's annotations that correspond to a set of expert annotations. Given the subjectivity of estimated FST, we use both an expert's annotations to compare against 320 annotations collected from both Scale AI and Centaur Labs and the dynamic crowd consensus annotations on the rest of the images as measures on which to evaluate annotation quality. For both Scale AI and Centaur Labs, we seed the dynamic crowd consensus protocol with expert annotations to avoid crowd prejudice equilibria that can arise in cold-start annotation tasks [152]. In the dynamic crowd consensus protocol devised with Centaur Labs, we included a qualified minimum agreement of 40% and qualified minimum and maximum annotations at 25 and 50, which means an individual is qualified only after attaining 40% agreement on 25 images and then an individual only remains qualified as long as her agreement remains above 40% for the 50 most recently annotated images. We selected the 40% minimum agreement threshold with Centaur Labs for two reasons: first, it is significantly above random guessing, which would be 16.7%, and second, we had previously found that 48% of consensus annotations by Scale AI matched expert 1's annotation exactly, so we rounded down to the nearest multiple of ten. The qualified minimum and maximum annotation levels were suggested by Centaur Labs based on past performance of their crowdworkers on other similar datasets. We did not include a dynamic quality control procedure for (dis)qualifying annotations with Scale AI.

For the 320 textbook images, Scale AI provided 156,566 annotations (ranging from 378 to 1094 annotations per image) and Centaur Labs provided 7,999 qualified annotations (ranging from 2 to 93 qualified annotations per image). For an additional 16,577 images from the Fitzpatrick 17k dataset, Scale AI provided 62,710 annotations (with an interquartile range of 4 to 4 annotations per image) and Centaur Labs provided 265,279 qualified annotations (with an interquartile range of 9 to 20 qualified annotations per image) [248]. In total, we collected 492,554 estimated FST annotations from crowd workers.

In addition to estimated FST annotations, we collected **agreement metrics** for measuring the agreement and difficulty of annotating images with estimated FSTs. These agreement metrics are weighted by each individual annotator's agreement with the expert annotations and defined for each image as follows: agreement is the weighted, qualified annotations with the consensus label divided by the weighted, qualified annotations; difficulty is the weighted, qualified annotations without the consensus label divided by the weighted, qualified annotations. In algebraic notation, agreement and difficulty can be written as $A = \frac{Q_c}{Q}$

and difficulty as $D = \frac{Q \notin Q_c}{Q}$ where Q_c is the weighted number of qualified annotations with the consensus label, Q is the weighted number of qualified annotations, and $Q \notin Q_c$ is the weighted number of qualified annotations that are not the consensus label.

Another criteria for assessing and improving the reliability of an images' annotations is the incorporation of **failure reports** [90]. Failure reports are comments on flagged images by annotators indicating that an image is either incorrectly labeled or inappropriate or irrelevant. Failure reports allow crowdsourced workers to transcend the 7 multiple choice labels (the FST scale and the not applicable option) to provide text-based feedback on the image. In the annotations by Centaur Labs, we stop labeling any image which was flagged as inappropriate or irrelevant once or flagged as incorrect twice. Across, the 320 textbook images, we received 20 failure reports on 17 images. We discuss the details of these failure reports in Section 4.2.

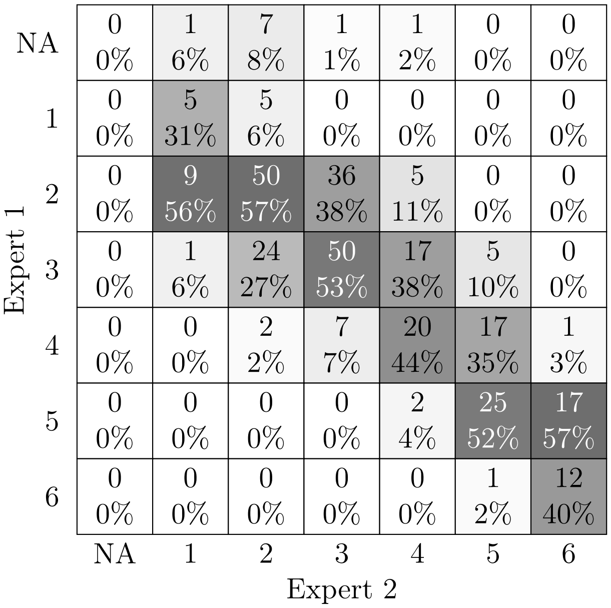
The final criteria for crowdsourcing is **expert review**, which is particularly useful for focusing the efforts by experts on the edge cases with high disagreement among crowd annotators. Expert review consists of experts reviewing flagged images without seeing the distribution of labels to adjudicate the annotation. We discuss the results of expert review in Section 4.3.

5.4 Results Comparing Annotations on 320 Textbook Images

We asked three board-certified dermatologists to annotate 320 images with FSTs, and we find that the annotations of any two experts match exactly on 50-55% of images and match within one unit on 92-94% of images. Figure 5-1 presents a confusion matrix comparing the annotations of the first two experts.

In comparison to the two experts' labels, the algorithmically generated annotations for the 320 images are much less similar. The ITA-FST algorithm produces Fitzpatrick labels identical to expert 1, 2, and 3 in 27%, 31%, and 40% of images, respectively, and is off by no more than a single unit (i.e., FST I vs FST II) in 70%, 69%, and 76% of images, respectively. See Figure 5-6 and Figure 5-7 in the Appendix for confusion matrices examining annotation discrepancies between experts 1 and 2 and the Scale AI, Centaur Labs, and ITA-FST algorithm.

The inter-rater reliability between the two experts' and the crowds' annotations is much more similar across the 320 images. The labels produced by Scale AI and Centaur Labs match expert 1 exactly in 48% and 40% of images, match expert 2 exactly in 50% and 38% of images, and match expert 3 exactly in 58% and 43% of images, all respectively. Likewise, the annotations produced by Scale AI and Centaur Labs are off by no more than a single unit from expert 1's annotations in 94% and 87% of images, expert 2's annotations in 91% and 79% of images, and expert 3's annotations in 93% and 88% of images.



Expert 1 \ Expert 2	NA	1	2	3	4	5	6
NA	0 0%	1 6%	7 8%	1 1%	1 2%	0 0%	0 0%
1	0 0%	5 31%	5 6%	0 0%	0 0%	0 0%	0 0%
2	0 0%	9 56%	50 57%	36 38%	5 11%	0 0%	0 0%
3	0 0%	1 6%	24 27%	50 53%	17 38%	5 10%	0 0%
4	0 0%	0 0%	2 2%	7 7%	20 44%	17 35%	1 3%
5	0 0%	0 0%	0 0%	0 0%	2 4%	25 52%	17 57%
6	0 0%	0 0%	0 0%	0 0%	0 0%	1 2%	12 40%

Figure 5-1: Confusion matrix comparing two board-certified dermatologists' Fitzpatrick skin type annotations on 320 images from dermatology textbooks.

5.4.1 Quantitative Assessment of Inter-Rater Reliability

In light of the subjectivity of estimated FST annotations, we evaluate annotation performance by comparing inter-rater reliability between pairs of experts with the inter-rater reliability between experts and each non-expert annotation method. Specifically, we measure inter-rater reliability using the Pearson Correlation Coefficient (ρ) between two annotation methods, and we evaluate the statistical significance following the Fisher Z transformation for comparing independent correlations [207]. We describe the pseudocode for comparing ρ_{E_i,E_j} with $\rho_{E_{i,Method}}$ in Algorithm 1 in the Appendix where E_i and E_j refer to one of the three experts and E_{Method} refers to one of the non-expert annotation methods. When calculating the $\rho_{X,Y}$ between two annotation methods X and Y, we drop annotations that either method marks as not applicable.

We find the inter-rater reliability of the ITA-FST algorithm is significantly lower than the inter-rater reliability of experts. The correlation between the first two experts' annotations is $\rho_{E_1,E_2} = .84$ (E_1 and E_2 refer to expert 1 and 2, respectively) whereas the correlation between the ITA-FST algorithm and any of the experts is $\rho_{ITA,E_1} = .57$, $\rho_{ITA,E_2} = .52$, and $\rho_{ITA,E_3} = .55$. The differences between any pair of experts ρ_{E_i,E_j} and $\rho_{E_1,ITA}$, $\rho_{E_2,ITA}$, and $\rho_{E_3,ITA}$ are statistically significant ($p < 0.00000001$). We present the correlations and the p-value of the comparisons of correlations in Table 5.1. We also present a heatmap of inter-rater reliability as measured by ρ in Figure 5-2.

In contrast to the low inter-rater reliability between experts and the algorithm, we find the inter-rater reliability of expert and crowdsourced annotations to be comparable. Notably, the crowdsourced annotations are slightly more correlated with experts' annotations in five of six comparisons – $\rho_{E_1,S} = .88$, $\rho_{E_1,C} = .88$, $\rho_{E_2,S} = .86$, $\rho_{E_2,C} = .83$, $\rho_{E_3,S} = .87$, $\rho_{E_3,C} = .87$ (S and C refer to Scale AI and Centaur Labs, respectively) – than experts' annotations are correlated with each other ($\rho_{E_1,E_2} = .84$, $\rho_{E_2,E_3} = .85$, and $\rho_{E_1,E_3} = .86$). We do not find statistically significant differences between the experts' correlation with each other and either expert's correlation with any of the crowdsourced methods.

Method	E_1 (ρ)	E_1 p-value	E_2 (ρ)	E_2 p-value	E_3 (ρ)	E_3 p-value
E_1			0.84	0.73	0.86	0.66
E_2	0.84	0.44			0.85	0.66
E_3	0.86	0.44	0.85	0.73		
ITA-FST	0.57	<0.001	0.52	<0.001	0.55	<0.001
Scale AI	0.88	0.08	0.86	0.43	0.88	0.08
Centaur Labs	0.88	0.08	0.83	0.50	0.87	0.32

Table 5.1: Inter-rater reliability based on Fisher Z transformations of Pearson Correlation Coefficients (ρ). The E_x (ρ) columns display the correlation between the method in the row and the method in the column. The p-value columns show the minimum p-value based on Algorithm 1 in the Appendix applied to all pairwise correlations of experts; as an example, the cell in the E_1 p-value column and ITA-FST row presents the minimum p-value comparing (a) $\rho_{E_1,ITA}$ and ρ_{E_1,E_2}, (b) $\rho_{E_1,ITA}$ and ρ_{E_1,E_3}, and (c) $\rho_{E_1,ITA}$ and ρ_{E_2,E_3}
.

In addition to examining the inter-rater reliability across methods, we examine how inter-rater reliability changes depending on the number of non-qualified annotations. Instead of assessing FSTs based on a dynamic consensus procedure, we compare expert 1's annotations with the crowd mean of 25 random draws from the Scale AI annotations (which were non-qualified meaning that crowdworkers were not filtered by a task specific quality control procedure) in samples of the following sizes: 3, 6, 12, 24, 48, and 96 annotations. We find a logarithmic relationship between ρ_{S,E_1} and sample size that plateaus with ρ_{S,E_1} approaching 0.88; see Figure 5-2 for a visualization of this relationship. For example, an increase from 3 to 12 annotations per image is associated with a 10 percentage point increase in ρ_{S,E_1}; the mean ρ_{S,E_1} is 0.74 with a standard deviation of 0.026 when evaluating across 3 annotations per image and 0.84 with a standard deviation of 0.01 when evaluating across 12 annotations per image. A further increase from 12 to 24 annotations per image is associated with another 2 percentage points increase in ρ_{S,E_1}. We also find a similar relationship when comparing the varying size of the crowd with expert 2 and 3.

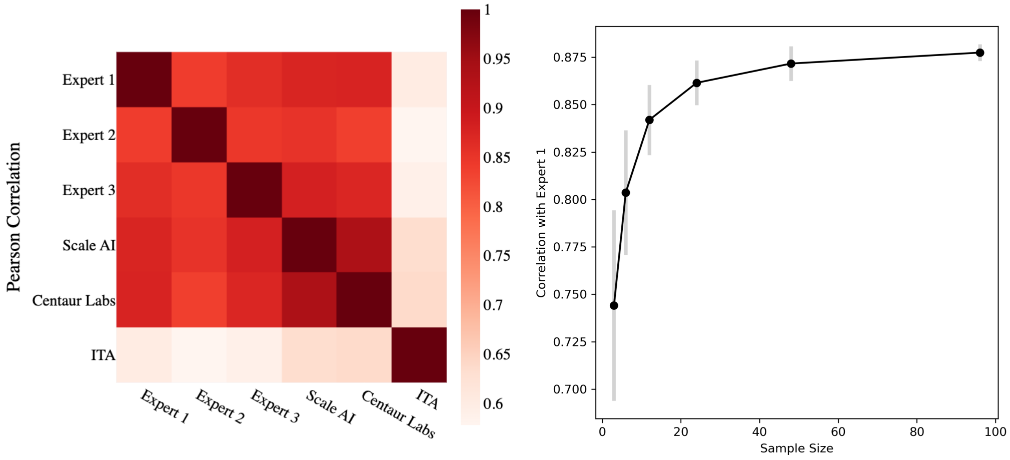


Figure 5-2: **Left**: Heatmaps showing inter-rater reliability as measured by Pearson's Correlation Coefficient. These heatmaps include 296 images and exclude the 24 images rated by any expert or crowdsourcing method as "Not Applicable." **Right**: Inter-rater reliability by crowd size based on 25 random bootstrapped samples from the Scale AI annotations. The y-axis presents the correlation between expert 1's annotations and the crowd's mean FST annotation. The x-axis presents the number of annotations per image. The gray bars represent the 95% confidence interval. As the number of annotations increases the confidence interval decreases and the Pearson Correlation Coefficient (ρ) approaches 0.88.

5.4.2 Qualitative Assessment of Inter-Rater Reliability

We examine inter-rater reliability qualitatively by illustrating similarities and differences in annotations across methods and examining images flagged by failure reports. In Figure 5-3, we present **qualitative confusion matrices** that showcase how different annotation methods lead to different annotations. These qualitative confusion matrices are intended to contextualize and illustrate similarities and discrepancies in subjective annotations and build upon the finding that alternative representation of confusion matrices can improve non-expert understanding of performance [564].

Across the 320 textbook images, annotators flagged 17 images as inappropriate or incorrect. Three of these flagged images were originally marked by expert 1 as "Not Applicable." Unlike most images, all three of these images contain multiple photographs under multiple lighting conditions, expert 2 provided a different annotation than expert 1, and the Centaur

Labs and Scale AI crowd labels are discordant. Another two of these flagged images are marked as confusing and neither the expert annotations or the crowd annotations agree with one another. The final 12 of the flagged images contain messages that the annotator is confident that the expert's label is wrong; in 5 of these 12 images, expert 2 and both crowd consensuses agree that expert 1's annotation is one unit off, in 6 of these 12 images, expert 2 agrees with expert 1 while both crowd consensuses disagree with the experts, and in 1 of these 12 images, there is disagreement across experts and crowd consensuses. These results suggest failure reports are generally useful in identifying images that are likely to be problematic and extremely subjective for one reason or another.

5.5 Scaling Annotations on the Fitzpatrick 17k

For resource constrained developers of large-scale image datasets, it is orders of magnitude less resource intensive to annotate images with an algorithm or crowdsourcing than with board-certified dermatologists [476, 545, 611]. Given the lower inter-rater reliability of the ITA-FST algorithm, we limit our analysis of scaling annotations on the full Fitzpatrick 17k dataset [248] to crowdsourcing methods. 85% of consensus FST annotations by Scale AI and Centaur Labs are within one unit of each other. In Figure 5-4, we present a confusion matrix, which reveals that large discrepancies in annotations between sources are rare. An expert review of all applicable annotation discrepancies that are off by more than one unit would involve examining 9% (1,365 of the 13,865 images) of the Fitzpatrick 17k dataset. Error reports by annotators from Centaur Labs indicate that the consensus annotation for 166 images are incorrectly labeled and 21 images are inappropriate or irrelevant for the task.

5.5.1 Expert Review of Scaled Annotations

As a final step in evaluating dynamic consensus protocols, we collect labels from 3 board-certified dermatologists on 140 images randomly selected from the 16,577 images in the

Fitzpatrick 17k dataset. We stratified this random selection on two features: (1) Scale AI's estimated FST annotations and (2) a binary variable for discrepancy between Scale AI's and Centaur Labs' annotations of more than 1 estimated FST annotation. As a result, there are 20 images with each Scale AI estimated FST type and 20 images annotated by Scale AI as not applicable. In addition, 70 of these images have been annotated by Scale AI and Centaur Labs within 1 estimated FST of each other and the other 70 images have been annotated with estimated FST that differ by more than 1.

For the 70 images with similar annotations, the correlation between experts ranges from 83% to 87% and the crowds correlation with experts ranges from 86% to 89%. We do not see any statistically significant difference between experts and crowds.

However, for the 70 images with greater than 1 unit discrepancies across the two crowd methods, we do find significant differences between inter-annotator reliability across experts and the crowd. The correlation between experts ranges from 59% to 66% and the crowds' correlation with experts ranges from 32% to 63%. We examine the inter-rater reliability between Scale AI and Centaur Labs and experts by conducting 12 tests of statistical significance to cover all possible comparison permutations. We find that 3 of 6 comparisons of inter-rater reliability between Scale AI and experts show Scale AI's annotations are less correlated and the p-value is less than the standard 5% threshold for statistical significance. Likewise, we find that 1 of 6 comparisons of inter-rater reliability between Centaur Labs and experts show Centaur Labs' annotations are less correlated and the p-value is less than the standard 5% threshold for statistical significance. In Table 5.2, we present the inter-rater reliability Pearson correlation coefficients and lowest p-values for tests of statistical significance. This table also includes an examination of estimated ITA-FST on these 70 images, and we find estimated the correlation between ITA-FST and experts' annotations approaches 0 for this selection of images for expert review.

The comparison of inter-annotator agreement on the images selected for expert review reveals important nuances that researchers should keep in mind when annotating future datasets.

Method	E_1 (ρ)	E_1 p-value	E_2 (ρ)	E_2 p-value	E_3 (ρ)	E_3 p-value
E_1			0.59	0.29	0.66	0.29
E_2	0.59	0.29			0.66	0.58
E_3	0.66	0.29	0.66	0.58		
ITA-FST	0.05	<0.001	-0.06	<0.001	0.08	<0.001
Scale AI	0.50	0.04	0.32	<0.001	0.57	0.19
Centaur Labs	0.63	0.54	0.47	0.02	0.53	0.09

Table 5.2: Analysis of subset of 70 images with high disagreement showing the inter-rater reliability based on Fisher Z transformations of Pearson Correlation Coefficients (ρ). The E_x (ρ) columns display the correlation between the method in the row and the method in the column. The p-value columns show the minimum p-value based on Algorithm 1 in the Appendix applied to all pairwise correlations of experts; as an example, the cell in the E_1 p-value column and ITA-FST row presents the minimum p-value comparing (a) $\rho_{E_1,ITA}$ and ρ_{E_1,E_2}, (b) $\rho_{E_1,ITA}$ and ρ_{E_1,E_3}, and (c) $\rho_{E_1,ITA}$ and ρ_{E_2,E_3}
.

While the inter-rater reliability on estimated FST is just as high between experts as it is between experts and the crowd consensus for most images, the annotations by crowds on images with low agreement may be less reliable than experts' annotations. By incorporating **expert review** into a subset of crowd annotations with low agreement, a dynamic consensus protocol can adjudicate edge cases such that adjudication leads to a higher likelihood of agreement with other experts.

This particular expert review of 140 images highlights edge cases where experts tend to agree with each other more often than they agree with dynamic consensus labels from crowdworkers. However, it is important to note that inter-rater reliability across experts on the 70 images randomly drawn from the 9% of images with two discordant crowd ratings ranges from 59% to 66% whereas inter-rater reliablity on the other 70 images (randomly drawn from the 91% of images with two concordant crowd ratings) ranges from 83% to 87%. On these 70 images with discordant crowd annotations, experts agree with each other significantly less than they do on the overwhelming majority of images. This lower rate of expert agreement and significantly lower rate of crowd worker and expert agreement demonstrates the subjectivity of estimating FST of an individual in an image can vary considerably across images.

5.6 Discussion

How well does the ITA-FST algorithm and various crowdsourcing methods compare to board-certified dermatologists in annotating images with estimated FSTs? Our results reveal that the inter-rater reliability between three board-certified dermatologists (as measured by ρ_{E_x,E_y}) is comparable to the inter-rater reliability between each board-certified dermatologist and each of the two crowdsourcing methods (as measured by $\rho_{E,Crowd}$). However, inter-rater reliability of the ITA-FST algorithm (as measured by $\rho_{E,ITA}$) is significantly lower than the inter-rater reliability between any two experts.

Estimated FST annotations on images are highly subjective. We find that three experts agree with each other exactly on estimated FST in only 50-55% of images (although they agree with each other within a one unit difference in 92-94% of images). Rather than treat this subjectivity as a bias, we treat subjectivity on a per annotation basis as a measure of signal and noise. We find the differences between the three experts' annotations are not significantly larger than the differences between the experts and either crowdsourced method. In other words, expert annotations generally have the same amount of signal and noise as crowd annotations. This general finding comes with a caveat: there are identifiable edge cases where experts' annotations demonstrated significantly higher inter-rater reliability than crowdsourced annotations. Nonetheless, our results suggest that crowdsourcing methods (but not the ITA-FST algorithm) can be reliable for annotating large scale dermatology image datasets with skin type annotations especially when expert review is included.

This is particularly important for increasing transparency in machine learning for dermatology because skin type annotations are one of the items on the CLEAR Derm checklist for the evaluation of image-based AI algorithms [144] and an important consideration for evaluating medical AI devices for FDA approvals [651]. Transparency on skin tone information can be useful for evaluating both the distribution (and potential under-representation) of various skin tones in image datasets and how AI algorithms in dermatology perform across different skin tones, which is then useful as evidence for holding the fields of computer vision and dermatology accountable for addressing the unwanted biases.

While crowdsourced annotations are comparable with experts' annotations in aggregate, there are many examples where experts agree with each other yet the crowd differs. One approach for reducing crowdsourcing disagreement with experts is to include more annotations per image, which we find is effective for reducing errors from crowd sizes of 3 to 12 but less effective for reducing errors from larger crowd sizes. A second approach is to integrate expert review into crowdsourcing. In particular, expert review examines edge cases that are flagged based on failure reports, agreement metrics (e.g. low agreement scores, high difficulty scores), and random samples for review. For the overwhelming majority of images, experts and the crowd have similar inter-rater reliability, but for the edge cases, expert review can offer additional reliability because inter-rater reliability of experts on edge cases can be higher than inter-rater reliability of crowds on edge cases.

The comparison between methods for annotating subjective labels provides a replicable methodology for answering when an algorithm or crowdsourced methodology can reliably be used in lieu of experts for annotating data. The goal of this kind of data annotation is to increase transparency in dataset biases to motivate greater accountability in sociotechnical decision-making systems. However, this kind of transparency comes at a cost. Human labor by experts or crowd annotators requires time and energy and should be compensated appropriately whereas the resources needed to compute ITA-FST scores are neglible. The low agreement between ITA-FST and experts is best to avoid because it may leave analyses prone to data cascade errors [549]. On the other hand, the relatively high agreement between experts and the crowd (and the opportunity to augment crowd labels with expert review) makes crowd annotations of estimated FST on images more attractive than expensive experts. We note that the crowd labels here come from Scale AI and Centaur labs, which represent very different ecosystems than the decentralized requester marketplace of Amazon Mechanical Turk (AMT) [405]. In particular, Scale AI and Centaur Labs work directly with individuals rather than through AMT, and as such, both these services avoid the "root problem... of unfair requesters" [267] in the AMT marketplace and the problem of turkers' uncertainty about the fairness of a particular requester [293]. Moreover, the ability to submit error reports with Centaur Labs creates a tractable opportunity for communication between crowdworkers analyzing the images and data scientists analyzing the data.

5.7 Limitations

We focused our comparisons on how three experts, one algorithm, and two crowdsourcing methods retrospectively annotate estimated FST across 320 images collected from dermatology textbooks and 140 images collected from online dermatology atlases. The 320 images showcase skin of all six skin types, but the distribution of skin types is not uniform across these images because dark skin types are underrepresented in dermatology textbooks [9, 23, 179]. Based on experts' annotations, only 18-26% of the 320 images show the two darkest skin types.

In our evaluation, we consider the ITA-FST algorithm applied to images with YCbCr masks, and we find it exhibits higher variability than expert and crowd-based annotations. While the ITA-FST algorithm may not be a reliable method for annotating estimated FST, future algorithms applied to images (especially segmented portions of images) may be able to match the inter-rater reliability of experts.

The lighting conditions are heterogeneous across these images, which makes assessing estimated FST more difficult than it would be in images with a single, consistent cross-polarized light source. Guidelines for photographing images of skin conditions on dark skin suggest images use indirect, natural light and a separate light for the hair and should avoid backgrounds with bright colors or patterns [379]. A recent perspective piece in the British Journal of Dermatology presents a series of images where the only difference is lighting source (cross-polarized light vs. white light) that reveals cross-polarized light reduces specular reflections and increases the contrast between healthy and unhealthy skin [471].

The variability of estimated FST annotations in images is much higher than in-person assessments because in-person assessments are not limited by lighting sources and enable a dermatologist to include an assessment of the patients' skin color, eye color, hair color, and history of sunburns. We leave the comparison of in-person FST annotation to image-based estimated FST annotations to future research.

The Fitzpatrick scale is a starting point but not a perfect method for annotating skin color [87, 632]. The Fitzpatrick classification system was originally designed for classifying skin based on skin's reaction to the sun (burning vs tanning) and not skin color [209]. Moreover, the original Fitzpatrick classification labeled FST I-IV as white, FST V as brown, and FST VI as black, which contrasts with how researchers describe today's usage of FST as pale-white for I, white for II, beige for III, brown for IV, dark brown for V, and black for VI [537]. We re-created the original scale in Table 5.3 in the Appendix for quick reference. Annotating images with estimated FSTs helps to document the diversity of dermatology datasets and inspect algorithms for discrimination based on the color of one's skin, but estimated FSTs serve as a blunt proxy (biased towards lighter skin colors) that fail to capture the global diversity of skin colors [632]. In order to avoid singularly optimizing future AI algorithms on a biased proxy [444], future research and data collection should consider additional methods and metrics for annotating the diversity and complexity of skin color including factors such as self-reported versus observer reported skin tone [437], in-person or image based assessment, and the number of response categories [512].

5.8 Conclusion

By annotating large datasets of dermatology images with FSTs, researchers can increase transparency and enable relatively straight-forward evaluations of algorithmic performance across skin types for AI models trained to classify skin conditions. While image-based FST annotations are subjective, we find the annotations of experts and crowds are highly comparable while the annotations produced by the ITA-FST algorithm are more variable. In light of the higher variability of annotations generated by the ITA-FST algorithm, we recommend that researchers do not augment their datasets of clinical dermatology images algorithmically and instead use a crowdsourcing or expert-based approach.

We find some instances where the experts concur yet the crowd consensus disagrees. We recommend the most efficient and thorough approach to annotating images of skin conditions with FSTs is to combine experts and the crowd. Expert review can adjudicate both images flagged for error reports and images with low agreement or high difficulty scores. While we propose this approach for annotation of FSTs, our recommendation for hybrid dynamic consensus protocols with experts and crowds may extend to other domains in which annotations are similarly subjective for experts and non-experts alike.

5.9 Data and Code Availability

The datasets and code generated and analyzed during the current study are available in our public Github repository, `https://github.com/mattgroh/fitzpatrick17k`.

Acknowledgments

We thank Erik Duhaime and Kira Prentice at Centaur Labs and Aerin Kim at Scale AI for providing data annotation services for free, the many crowdworkers for annotating the images, Bruke Wossenseged for valuable research assistance, and Rosalind Picard, Ziv Epstein, and Luis Soenksen for thoughtful feedback.

5.10 Appendix

Algorithm 1 Fisher Z transformation for comparing independent correlations

```
for Each Expert do
    for Each Method do
        $z_{E_A,E_B} \leftarrow \frac{1}{2} \ln \frac{1+\rho_{E_A,E_B}}{1-\rho_{E_A,E_B}}$
        $z_{Expert,Method} \leftarrow \frac{1}{2} \ln \frac{1+\rho_{Expert,Method}}{1-\rho_{Expert,Method}}$
        $Z \leftarrow |\frac{z_{E_A,E_B}-z_{Expert,Method}}{\sqrt{2/(n-3)}}|$ where n= # of images
        Convert Z score to p-value
    end for
end for
```

Algorithm 2 Individual typology angle threshold adjustment

```
$T_{12} \leftarrow Mean(ITA_1.Quantile(1), ITA_2.Quantile(3))$
$T_{23} \leftarrow Mean(ITA_2.Quantile(1), ITA_3.Quantile(3))$
$T_{34} \leftarrow Mean(ITA_3.Quantile(1), ITA_4.Quantile(3))$
$T_{45} \leftarrow Mean(ITA_4.Quantile(1), ITA_5.Quantile(3))$
$T_{56} \leftarrow Mean(ITA_5.Quantile(1), ITA_6.Quantile(3))$
$T \leftarrow \{T_{12}, T_{23}, T_{34}, T_{45}, T_{56}\}$
for all $t \in T$ do
    $Max\_Concordant \leftarrow Sum(Annotation\_E_1 = Annotation\_E_2 = ITA(t))$
    $I \leftarrow \{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$
    for all $i \in I$ do
        $t_i \leftarrow t + i$
        $Concordant \leftarrow Sum(Annotation\_E_1 = Annotation\_E_2 = ITA(t_i))$
        if $Concordant > Max\_Concordant$ then
            $Max\_Concordant \leftarrow Concordant$
            $t \leftarrow t_i$
        end if
    end for
end for
```

Skin Type	Skin Color	Sunburn	Tan
I	White	Yes	No
II	White	Yes	Minimal
III	White	Yes	Yes
IV	White	No	Yes
V	Brown	No	Yes
VI	Black	No	Yes

Table 5.3: The Fitzpatrick skin type scale from Fitzpatrick et al 1988 [209]. The scale is intended for classifying sun-reactive skin types. Notably, the original scale does not include nuanced skin tones or color beyond white, brown, and black. In dermatology practice, the Fitzpatrick scale is commonly used to describe constitutive skin color [632]. Recent research published in the *Medical Image Analysis* describes the Fitzpatrick skin types as pale-white, white, beige, brown, dark brown, and black [537]. We informed crowd annotators by presenting example images of each FST.

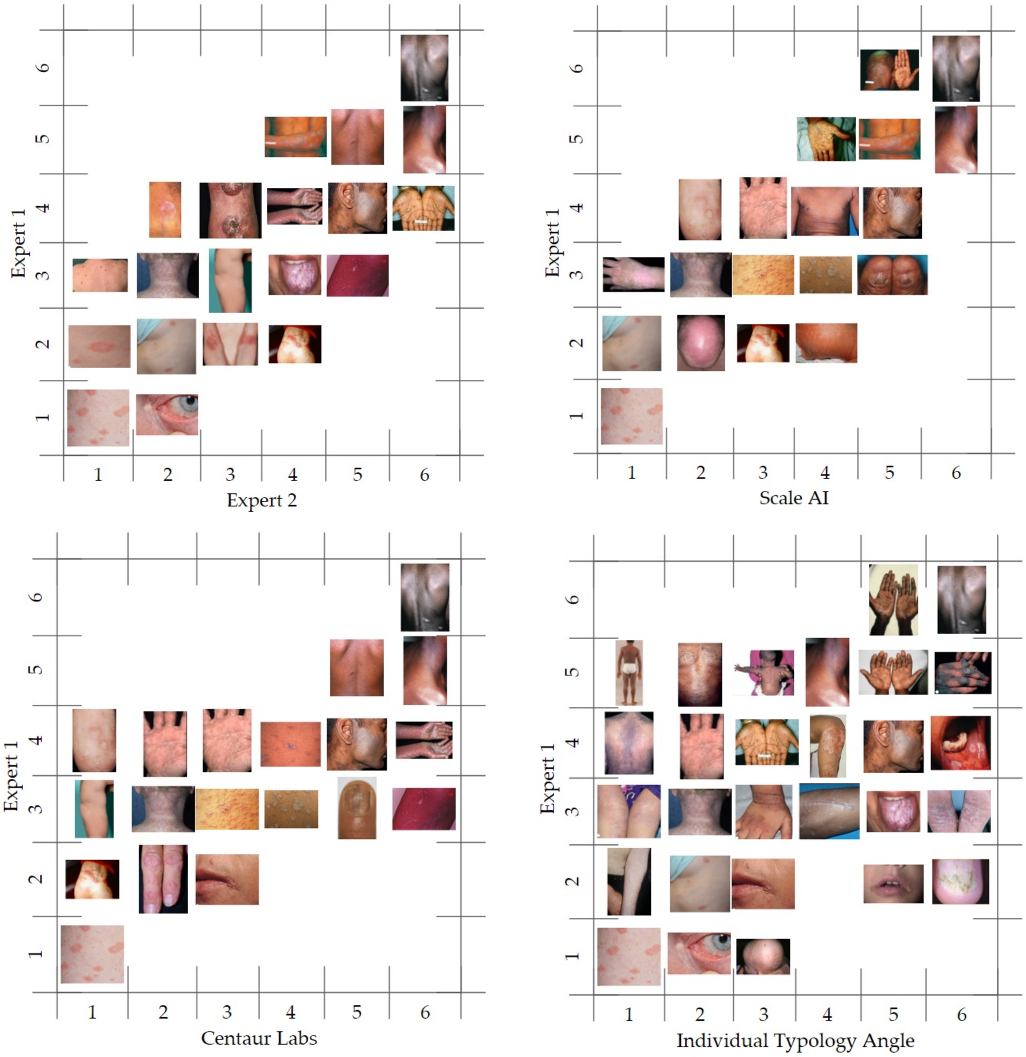
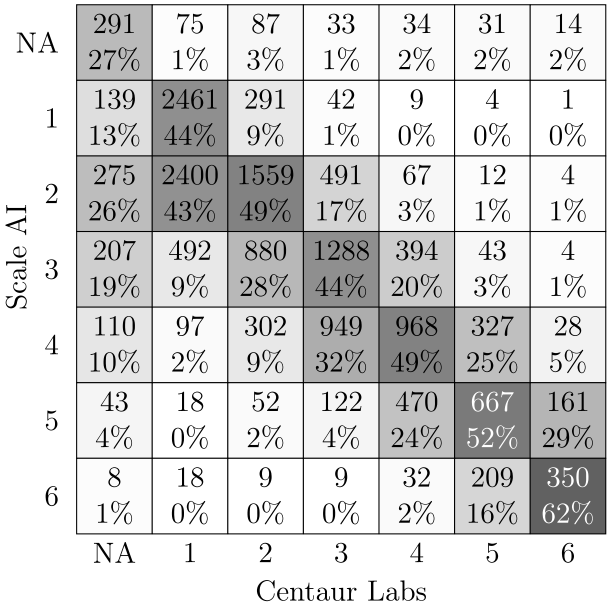


Figure 5-3: Textbook images [73, 241, 255, 298, 309, 311, 430, 648] of skin conditions plotted according to Expert 1's annotations (on the Y-axis) and 4 other methods (Expert 2, ITA-FST algorithm Scale AI, and Centaur Labs on the X-axis).



Scale AI \ Centaur Labs	NA	1	2	3	4	5	6
NA	291 27%	75 1%	87 3%	33 1%	34 2%	31 2%	14 2%
1	139 13%	2461 44%	291 9%	42 1%	9 0%	4 0%	1 0%
2	275 26%	2400 43%	1559 49%	491 17%	67 3%	12 1%	4 1%
3	207 19%	492 9%	880 28%	1288 44%	394 20%	43 3%	4 1%
4	110 10%	97 2%	302 9%	949 32%	968 49%	327 25%	28 5%
5	43 4%	18 0%	52 2%	122 4%	470 24%	667 52%	161 29%
6	8 1%	18 0%	9 0%	9 0%	32 2%	209 16%	350 62%

Figure 5-4: Confusion matrix comparing two crowdsourcing methods for annotating the 16,577 images in the Fitzpatrick 17k dataset

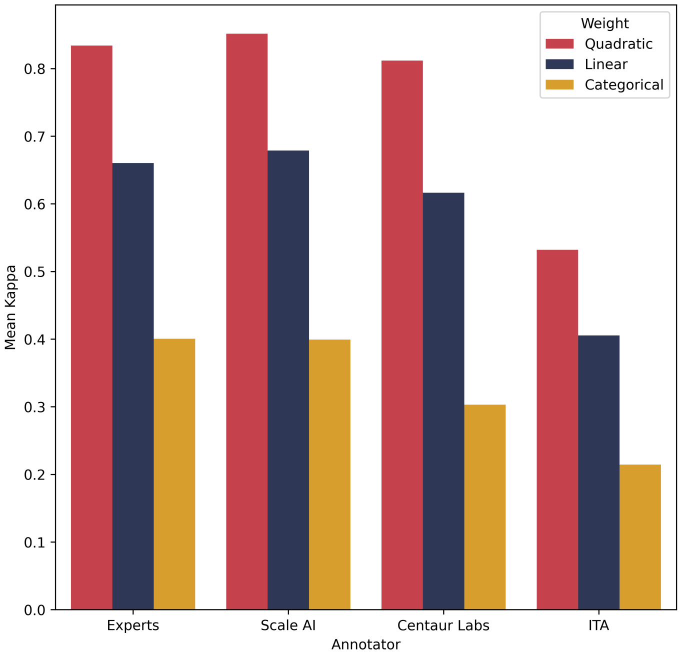
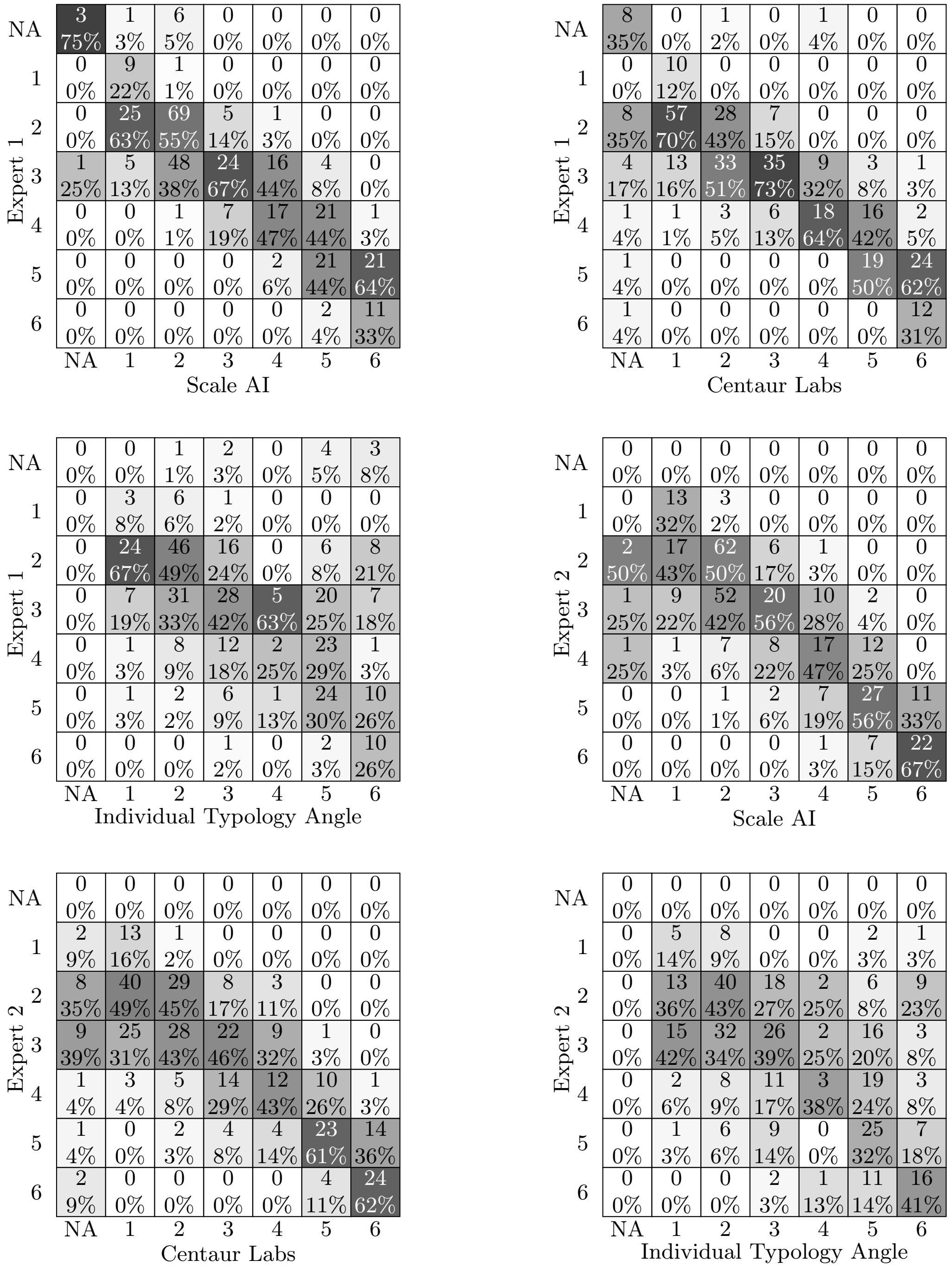


Figure 5-5: Inter-rater reliability measured via mean Cohen's kappa between experts and other annotators. Cohen's kappa based on categorical weighting is lower than Cohen's kappa based on linear and quadratic weighting. Many annotations are only off by a single unit and categorical weighting penalizes annotations off by a single unit the same as annotation off by multiple units.



Expert 1	NA	1	2	3	4	5	6
NA	3 75%	1 3%	6 5%	0 0%	0 0%	0 0%	0 0%
1	0 0%	9 22%	1 1%	0 0%	0 0%	0 0%	0 0%
2	0 0%	25 63%	69 55%	5 14%	1 3%	0 0%	0 0%
3	1 25%	5 13%	48 38%	24 67%	16 44%	4 8%	0 0%
4	0 0%	0 0%	1 1%	7 19%	17 47%	21 44%	1 3%
5	0 0%	0 0%	0 0%	0 0%	2 6%	21 44%	21 64%
6	0 0%	0 0%	0 0%	0 0%	0 0%	2 4%	11 33%

Scale AI

Expert 1	NA	1	2	3	4	5	6
NA	8 35%	0 0%	1 2%	0 0%	1 4%	0 0%	0 0%
1	0 0%	10 12%	0 0%	0 0%	0 0%	0 0%	0 0%
2	8 35%	57 70%	28 43%	7 15%	0 0%	0 0%	0 0%
3	4 17%	13 16%	33 51%	35 73%	9 32%	3 8%	1 3%
4	1 4%	1 1%	3 5%	6 13%	18 64%	16 42%	2 5%
5	1 4%	0 0%	0 0%	0 0%	0 0%	19 50%	24 62%
6	1 4%	0 0%	0 0%	0 0%	0 0%	0 0%	12 31%

Centaur Labs

Expert 1	NA	1	2	3	4	5	6
NA	0 0%	0 0%	1 1%	2 3%	0 0%	4 5%	3 8%
1	0 0%	3 8%	6 6%	1 2%	0 0%	0 0%	0 0%
2	0 0%	24 67%	46 49%	16 24%	0 0%	6 8%	8 21%
3	0 0%	7 19%	31 33%	28 42%	5 63%	20 25%	7 18%
4	0 0%	1 3%	8 9%	12 18%	2 25%	23 29%	1 3%
5	0 0%	1 3%	2 2%	6 9%	1 13%	24 30%	10 26%
6	0 0%	0 0%	0 0%	1 2%	0 0%	2 3%	10 26%

Individual Typology Angle

Expert 2	NA	1	2	3	4	5	6
NA	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%
1	0 0%	13 32%	3 2%	0 0%	0 0%	0 0%	0 0%
2	2 50%	17 43%	62 50%	6 17%	1 3%	0 0%	0 0%
3	1 25%	9 22%	52 42%	20 56%	10 28%	2 4%	0 0%
4	1 25%	1 3%	7 6%	8 22%	17 47%	12 25%	0 0%
5	0 0%	0 0%	1 1%	2 6%	7 19%	27 56%	11 33%
6	0 0%	0 0%	0 0%	0 0%	1 3%	7 15%	22 67%

Scale AI

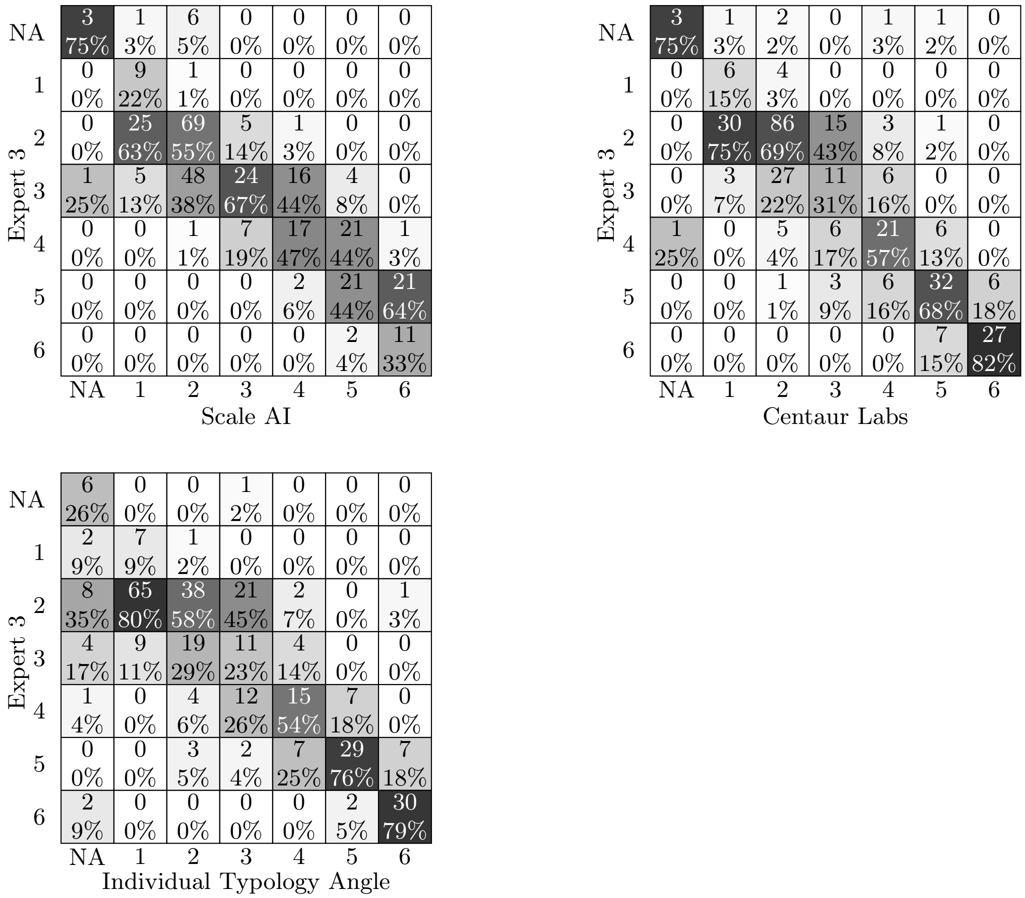
Expert 2	NA	1	2	3	4	5	6
NA	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%
1	2 9%	13 16%	1 2%	0 0%	0 0%	0 0%	0 0%
2	8 35%	40 49%	29 45%	8 17%	3 11%	0 0%	0 0%
3	9 39%	25 31%	28 43%	22 46%	9 32%	1 3%	0 0%
4	1 4%	3 4%	5 8%	14 29%	12 43%	10 26%	1 3%
5	1 4%	0 0%	2 3%	4 8%	4 14%	23 61%	14 36%
6	2 9%	0 0%	0 0%	0 0%	0 0%	4 11%	24 62%

Centaur Labs

Expert 2	NA	1	2	3	4	5	6
NA	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%
1	0 0%	5 14%	8 9%	0 0%	0 0%	2 3%	1 3%
2	0 0%	13 36%	40 43%	18 27%	2 25%	6 8%	9 23%
3	0 0%	15 42%	32 34%	26 39%	2 25%	16 20%	3 8%
4	0 0%	2 6%	8 9%	11 17%	3 38%	19 24%	3 8%
5	0 0%	1 3%	6 6%	9 14%	0 0%	25 32%	7 18%
6	0 0%	0 0%	0 0%	2 3%	1 13%	11 14%	16 41%

Individual Typology Angle

Figure 5-6: Confusion matrices comparing experts 1 and 2 to the ITA algorithm and crowdsourcing methods.



Expert 3 \ Scale AI	NA	1	2	3	4	5	6
NA	3 75%	1 3%	6 5%	0 0%	0 0%	0 0%	0 0%
1	0 0%	9 22%	1 1%	0 0%	0 0%	0 0%	0 0%
2	0 0%	25 63%	69 55%	5 14%	1 3%	0 0%	0 0%
3	1 25%	5 13%	48 38%	24 67%	16 44%	4 8%	0 0%
4	0 0%	0 0%	1 1%	7 19%	17 47%	21 44%	1 3%
5	0 0%	0 0%	0 0%	0 0%	2 6%	21 44%	21 64%
6	0 0%	0 0%	0 0%	0 0%	0 0%	2 4%	11 33%

Expert 3 \ Centaur Labs	NA	1	2	3	4	5	6
NA	3 75%	1 3%	2 2%	0 0%	1 3%	1 2%	0 0%
1	0 0%	6 15%	4 3%	0 0%	0 0%	0 0%	0 0%
2	0 0%	30 75%	86 69%	15 43%	3 8%	1 2%	0 0%
3	0 0%	3 7%	27 22%	11 31%	6 16%	0 0%	0 0%
4	1 25%	0 0%	5 4%	6 17%	21 57%	6 13%	0 0%
5	0 0%	0 0%	1 1%	3 9%	6 16%	32 68%	6 18%
6	0 0%	0 0%	0 0%	0 0%	0 0%	7 15%	27 82%

Expert 3 \ Individual Typology Angle	NA	1	2	3	4	5	6
NA	6 26%	0 0%	0 0%	1 2%	0 0%	0 0%	0 0%
1	2 9%	7 9%	1 2%	0 0%	0 0%	0 0%	0 0%
2	8 35%	65 80%	38 58%	21 45%	2 7%	0 0%	1 3%
3	4 17%	9 11%	19 29%	11 23%	4 14%	0 0%	0 0%
4	1 4%	0 0%	4 6%	12 26%	15 54%	7 18%	0 0%
5	0 0%	0 0%	3 5%	2 4%	7 25%	29 76%	7 18%
6	2 9%	0 0%	0 0%	0 0%	0 0%	2 5%	30 79%

Figure 5-7: Confusion matrices comparing expert 3 to the ITA algorithm and crowdsourcing methods.

Chapter 6

Diagnostic Accuracy across Light and Dark Skin by Dermatologists, Primary Care Physicians, and Physician-Machine Partnerships

Abstract

Recent advances in deep learning systems (DLS) for image-based medical diagnosis demonstrate the potential to augment clinical decision-making, but the effectiveness of physician-machine partnerships remains an open question because physicians and algorithms are susceptible to systematic errors especially on underrepresented populations. We present results from a large-scale digital experiment (N=390 board-certified dermatologists and N=460 primary care physicians from 39 countries) to evaluate the accuracy of physicians submitting up to 4 differential diagnoses on 364 images of 46 skin conditions in a store-and-forward teledermatology simulation. We find specialists achieve 38% accuracy, specialists and generalists alike are 4 percentage points less accurate on images of dark skin than light skin, and fair DLS decision support improves physicians' diagnostic accuracy by 33%, possibly reduces accuracy disparities of specialists, but exacerbates accuracy disparities of generalists. These results reveal image-based diagnosis is challenging, but well-designed physician-machine partnerships can enhance physician performance. [1]

[1] This chapter, which is co-authored by Omar Badri, Roxana Daneshjou, Arash Koochek, Caleb Harris, Luis Soenksen, P. Murali Doraiswamy, and Rosalind Picard is currently under peer review.

6.1 Motivation

The future of machine learning in medicine is unlikely to involve substituting machines for physicians but instead involves physician-machine partnerships where domain-specific interfaces built on top of machine learning models may support clinical expertise in providing more accurate diagnoses for patients [111, 300, 327, 351, 487, 608, 619, 638]. However, an emerging literature on human and artificial intelligence collaboration reveals that physician-machine partnerships are not guaranteed to be better than either physicians or machines alone [101, 221, 243, 369, 561, 575]. In particular, experts may have trouble recognizing when to override or defer to algorithmic advice, which may be systematically biased in ways unknown to the expert. Initial research in store-and-forward teledermatology suggests clinical decision support based on a deep learning system (DLS) can improve diagnostic accuracy by generalists [300], but open questions remain about how physician-machine partnerships perform across levels of physician expertise and across underrepresented populations [519, 612].

Racial bias in medicine is well documented [21, 150, 470, 573, 641]. In dermatology, there is a lack of representation of diverse skin tones that permeates textbooks [9, 23], residency programs [380], dermatology research [378], non-specialists' diagnostic accuracy [162, 198], and training data for machine learning algorithms [143]. While deep learning models show promise for enhancing clinical decision-making in dermatology [187, 388], algorithmic audits of deep learning models for dermatology reveal these applied models often exhibit systematic errors on subsets of the data, especially on dark skin [146, 248]. Recent research in machine learning applied to dermatology has focused on increasing transparency in large-scale dermatology image datasets by annotating images with estimated Fitzpatrick skin type (FST) [246], developing new datasets with a focus on diversity [146] and creating synthetic images with diffusion models [547]. These solutions can address some of the current issues of transparency and performance disparities [113], but an open question remains of how accurately specialist and generalist physicians diagnose skin disease across skin tones with store-and-forward teledermatology and how a physician-machine partnership may help to reduce (or possibly exacerbate) any potential differences in diagnostic accuracy across skin color.

Methods from digital experiments in social sciences can be used for evaluating accuracy and bias in medical decision making and physician-AI interaction. Similarly to how crowdworkers on MTurk enabled the transformation of experimentation in social and behavioral science a decade ago [62, 483, 522], physician platforms offer an opportunity to recruit large numbers of physicians for surveys and diagnostic accuracy experiments [169–171]. We recruit a large number of physician participants by paying a nominal fee and designing the experiment to be a fun learning experience drawing on research on gamified behavioral experiments [390]. In addition to following designs from gamified experimentation, we follow guidance from integrative experimentation [20] and identify a reproducible experimental design space that covers the following dimensions: skin conditions, skin color, physician expertise, physician-machine partnerships, clinical decision support accuracy, and user interaction designs. Our experiment focuses on diagnostic accuracy and follows methods from algorithmic audits [387], which serve as a useful tool for systematically evaluating errors, exposing bias, and promoting transparency in machine learning algorithms [87]. We build upon recent work in diagnosing physician error [445] to demonstrate that diagnostic accuracy experiments can offer insights into performance of physicians and physician-machine partnerships.

Specifically, we designed a custom, digital experiment to evaluate physicians' diagnostic accuracy on images of inflammatory appearing skin conditions. We curated 364 images of 46 skin conditions. The vast majority of images (78%) depict the following 8 main diseases of which we have at least 29 images for each disease: atopic dermatitis, cutaneous t-cell lymphoma, dermatomyositis, lichen planus, lyme disease, pityriasis rosea, pityriasis rubra pilaris, and secondary syphilis. The selected images represent a near uniform distribution across skin color as measured by estimated FST. We hosted these images in a image-only, simulated store-and-forward experiment, which is outlined in Figure 6-1 and is a setting that limits the amount of information available to the physician relative to the information available in an in-person clinical visit. The experiment begins with randomized assignment of participants to two sets of conditions: two versions of the DLS and two interfaces for clinical decision support. The control DLS is a neural network architecture trained to classify 9 classes (the 8 main conditions and an other class to represent all other conditions) and has a top-1 accuracy of 47% and is a fair classifier in the sense that accuracy is highly similar across FST. The treatment DLS is Wizard of Oz classifier, which is a synthetically enhanced version of the control DLS, where we randomly re-assigned 65% of wrong classifications to be correct classifications resulting in 84% top-1 accuracy. The synthetically enhanced DLS is designed to anticipate future DLS systems that may be significantly more accurate than today's leading systems. The control clinical decision support interface consists of three buttons in the follow order "Update my top prediction with [condition]" "Update my differential to include [condition]" and "Keep my differential." The treatment interface consists of the same three buttons in reverse order as seen in Figure 6-11. For full details about either the DLS or interface for clinical decision support, see the Deep Learning System Development subsection in the Methods.

The experiment began by presenting participants with 7 pre-survey questions, instructions, and the diagnostic accuracy task where we ask participants to provide a differential diagnosis of 3 diseases (see Figure 6-7, Figure 6-8, Figure 6-9 in the Appendix for screenshots of the experimental interface). Next, we presented physicians with clinical decision support and asked physicians to decide whether or not to include the decision support suggested diagnosis in their differential (see Figure 6-11 in the Appendix). In this experiment, we motivated participant engagement by showing the reference condition after each trial and overall performance after 10 trials, which allowed physicians to learn about the content (e.g. Which images correspond to which condition? How often is the decision support correct?) and themselves (e.g. Did the participant diagnose the image correctly? How accurate is the participant compared to other specialists, generalists, and the DLS?).

In the results section, we evaluate how accurately specialist and generalist physicians diagnose images of inflammatory appearing skin disease. We consider three measures of accuracy: top-1 accuracy (Does the participant's leading diagnosis match the skin condition in the image?), top-3 accuracy (Do any of the participant's initial three differential diagnoses match the skin condition in the image?), and top-4 accuracy (Do any of the participant's initial three differential diagnoses or the decision support suggestion – if included by the participant – match the skin condition in the image?). We further evaluate how diagnostic accuracy differs across different skin tones in the images and physicians' experience with different skin tones. Finally, we consider how DLS-based decision support influences diagnostic accuracy.

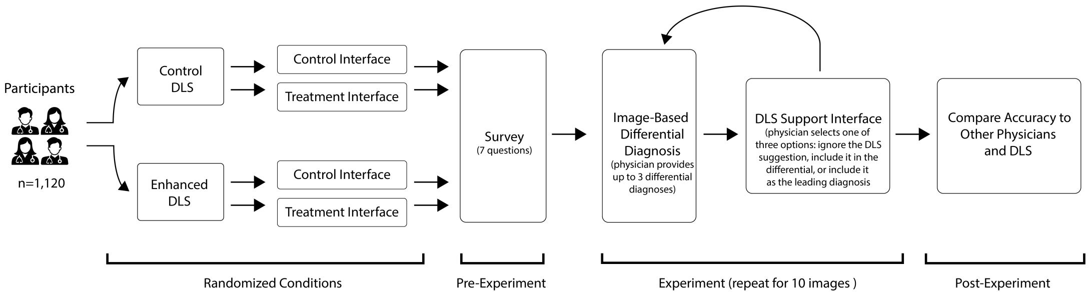

Participants
n=1,120
Control DLS
Enhanced DLS
Control Interface
Treatment Interface
Control Interface
Treatment Interface
Survey
(7 questions)
Image-Based Differential Diagnosis
(physician provides up to 3 differential diagnoses)
DLS Support Interface
(physician selects one of three options: ignore the DLS suggestion, include it in the differential, or include it as the leading diagnosis
Compare Accuracy to Other Physicians and DLS
Randomized Conditions
Pre-Experiment
Experiment (repeat for 10 images)
Post-Experiment

Figure 6-1: Illustration of the experimental design. See Figures 6-7 - 6-12 for screenshots of the experiment's user interface.

6.2 Results

6.2.1 Physician Characteristics (N=1,120)

In our digital, diagnostic accuracy experiment, we collected 14,261 differential diagnoses from 1,118 individuals on 364 images. This includes 5,365 differentials from 389 board-certified dermatologists (BCDs), 1,691 differentials from 116 dermatology residents, 5,458 differentials from 459 individual primary care physicians (PCPs), and 1,747 differentials from 154 other physicians. The first image shown in the experiment is an image of a woman with acne, which serves as an attention check that physicians at all levels of expertise should be able to diagnose accurately. We find 98% of BCDs, PCPs, and other physicians pass the attention check and 96% of dermatology residents pass the attention check. Moreover, 76% of BCDs and PCPs, 73% of other physicians, and 72% of dermatology residents pass the attention check and provide differential diagnoses on at least 10 images. After participants provide 10 differential diagnoses, we thank each participant for completing the experiment, reveal the aggregate performance of other participants to the participant, and offer the participant to continue diagnosing skin conditions in the experiment if they would like (see Figure 6-1 for an illustration of the experimental design).

In the results sections on diagnostic accuracy, we focus our analysis on the first ten differentials provided by participants who passed the attention check and provided at least 10 differentials. This includes 2,660 differentials from 296 BCDs, 747 differentials from 83 dermatology residents, 3,150 differentials from 350 PCPs, and 1,015 differentials from 113 other physicians. Our results are robust to other selection criteria, such as only participants from the United States, participants who provided fewer than 10 differentials, and all participants who pass the attention check. This experiment includes physicians living in 39 countries; half of these physicians live in the United States.

6.2.2 Image quality

In order to ensure the skin condition reference labels accurately represent the skin conditions in the images, we followed a five-step quality control process with three BCDs, conducted a post-hoc quality review, and evaluated accuracy rates across image sources which we describe in the Methods section.

6.2.3 Overall Diagnostic Accuracy

In this experiment, participants do not know which skin conditions will appear, and as such, the accuracy of random guessing is near 0% (see Experimental Interface subsection in Methods for more details). The top-3 accuracy of BCDs, dermatology residents, PCPs, and other physicians, as measured by any of their three differential diagnoses matching the reference label, is 38%, 36%, 19%, and 18%, respectively across all images in this experiment (excluding the attention check image) and 37%, 35%, 17%, and 16%, respectively across images of the eight main conditions in this experiment.

The top-1 accuracy, the accuracy of the leading diagnosis only, for BCDs, dermatology residents, PCPs, and other physicians is 27%, 24%, 14%, and 13%, respectively across all images in this experiment (excluding the attention check image) and 27%, 24%, 13%, and 12%, respectively across images of the eight main conditions in this experiment.

The top row of Figure 6-2 presents the mean diagnostic accuracy of participants' split by their primary, secondary, and tertiary diagnoses for images of the eight main conditions in this experiment.

The bottom row of Figure 6-2 presents the top-3 accuracy of BCDs' and PCPs' full differential diagnosis across the eight main conditions and a category labeled "Other", which aggregates the auxiliary 38 skin conditions into a single category. BCDs significantly outperform PCPs at visually diagnosing skin conditions from images across seven of the eight skin conditions and the other category.

We find the most common leading diagnosis for each image by BCDs and PCPs is correct in 48% and 33% of observations, respectively. At least one BCD identified the reference label in their differential diagnosis in 77% of images while at least one PCP identified the reference label in their differential diagnosis in 58% of images. After seeing a correct DLS prediction, at least one BCD included the reference label in their differential diagnosis in 98% of images.

6.2.4 Diagnostic Accuracy across Light and Dark Skin

Across all images, we find skin conditions in dark skin (estimated FST 5 & 6) are diagnosed less accurately than skin conditions in light skin (estimated FST 1-4). Across all participants, we find the top-1 and top-3 accuracy for skin conditions in dark skin is 4 percentage points ($p < 0.001$ and $p = 0.001$, respectively) lower than skin conditions in light skin. All statistical comparisons in this paper are based on ordinary least squares regression with robust standard errors clustered at the participant level unless otherwise noted. When we examine the physician types separately, we find BCDs, dermatology residents, PCPs, and other physicians are lower by 5 percentage points ($p = 0.011$), 5 percentage points ($p = 0.114$), 3 percentage points ($p = 0.006$), and 5 percentage points ($p = 0.012$) for images of dark skin than light skin, respectively. The top-3 diagnostic accuracy of BCDs, dermatology residents, PCPs, and other physicians is lower by 3 percentage points($p = 0.117$), 5 percentage points ($p = 0.113$), 4 percentage points ($p = 0.008$), and 4 percentage points ($p = 0.092$) for images of dark skin than light skin, respectively. We find qualitatively similar results in a series of robust checks including only participants who live in the United States, include participants who provide fewer than 10 responses, and include all responses from all participants who pass the attention check reveal similar results.

The top and middle row of Figure 6-3 presents top-3 diagnostic accuracy across skin conditions for BCDs and PCPs, respectively. BCDs diagnosed seven out of eight skin conditions and the other category with higher accuracy for light skin than dark skin images. The only skin condition in which BCDs are more accurate on dark skin than light skin is lichen planus. We do not find statistically significant differences in top-3 accuracy across skin color across individual skin conditions for BCDs, but we find statistically significant differences in BCDs' top-1 accuracy across light and dark skin images in four conditions: atopic dermatitis, Lyme disease, pityriasis rosea, and CTCL, respectively, which are 18 percentage points ($p = 0.007$), 20 percentage points ($p < 0.001$), 19 percentage points ($p = 0.001$), and 10 percentage points ($p = 0.009$) lower on dark skin, respectively. We find statistically significant and large differences in the top-3 and top-1 diagnostic accuracy of PCPs across light and dark skin images in three conditions: atopic dermatitis, Lyme disease, and pityriasis rosea, respectively.

We find that accuracy disparities across skin color are moderated by the diversity of patients seen by PCPs and PCPs' training. In particular, we find that PCPs who report seeing mostly or all white patients are 7 percentage points ($p = 0.009$) less accurate (top-3) on dark skin images than light skin images. We do not find statistically significant differences for BCDs based on self-reported patient diversity. Likewise, we find PCPs' who report sufficient training are 5 percentage points ($p = 0.079$) more accurate (top-3) than PCPs' who report insufficient training on images of dark skin than light skin. We do not find statistically significant differences in BCDs' top-1 or top-3 accuracy with respect to their self-reported sufficient training on dark skin. Likewise, we do not find statistically significant differences in BCDs' or PCPs' top-1 or top-3 accuracy with respect to their years of experience or self-reported difficulty with white patients relative to non-white patients.

6.2.5 Deep Learning System for Clinical Decision Support

We find the decision support tool significantly increases diagnostic accuracy while leading to the inclusion of relatively few incorrect diagnoses. With access to suggestions from the control DLS, BCDs' and PCPs' top-1 accuracy on the main eight conditions increases from 27% to 36% ($p < 0.001$, t-test) and 13% to 22% ($p < 0.001$, t-test), respectively. We find even larger accuracy gains when moving from top-3 accuracy without the DLS support to top-4 accuracy with the DLS support on the main eight conditions: BCDs' accuracy increases from 37% to 60% and PCPs' accuracy increases from 17% to 47%. Figure 6-4 shows physicians' top-1 accuracy (on the top) and top-3 and top-4 accuracy (on the bottom) before and after participants see the DLS-based suggestions.

On images where the DLS makes an incorrect suggestion, we find minimal effects on BCDs' and PCPs' top-1 accuracy, which both decrease by 1.2 percentage points ($p = 0.517$ and 0.312, respectively, t-test). In instances where the DLS provides an incorrect suggestion, we find that BCDs and PCPs override their correct leading diagnosis with an incorrect suggestion in fewer than 2% of observations. In contrast when the decision support provides an incorrect suggestion and BCDs' and PCPs' three differential diagnoses are all incorrect, we find that BCDs and PCPs include incorrect suggestions as leading diagnoses in 10% and 14% of observations, respectively. The BCDs' top-4 accuracy with decision support includes 1.58 incorrect diagnoses per observation and a top-3, top-2, and top-1 accuracy without the decision support of 1.40, 1.05, and 0.59 incorrect diagnoses per image, respectively. In contrast, PCPs' top-4 accuracy with the decision support includes 1.72 incorrect diagnoses per observation whereas top-3, top-2, and top-1 accuracy without the decision support includes 1.55, 1.26, and 0.82 incorrect diagnoses per image, respectively.

With respect to top-1 accuracy, we find BCDs without decision support are 5 percentage points ($p < 0.001$, t-test) more accurate than PCPs with the control DLS decision support but 4 percentage points ($p = 0.022$, t-test) less accurate than PCPs with enhanced DLS decision support.

In Table 6.2 in the Appendix, we present main effects of physician expertise, skin tone in an image, DLS suggestions, and interactions between these variables. In this regression table where we focus on BCDs and PCPs, we present top-1 accuracy in the first column and top-4 accuracy in the second column. For top-1 accuracy, we find BCDs are 13 percentage points more accurate than PCPs ($p < 0.001$), participants are 3 percentage points less accurate on images of dark skin ($p = 0.006$), the DLS suggestions leads to 8 percentage points higher performance overall ($p < 0.001$), and the treatment DLS leads to an additional 8 percentage point increase in accuracy ($p = 0.002$). Likewise, we find the control DLS suggestion exacerbates the accuracy disparities in PCPs' diagnoses by 5 percentage points ($p = 0.008$ and 0.048, respectively for top-1 and top-4 accuracy), but we do not find statistically significant evidence that accuracy disparities increase for BCDs. The three way interaction between BCDs, dark skin, and the DLS suggestion shows that the DLS suggestions on dark skin lead to a marginal 4 and 8 percentage point increase top-1 and top-4 accuracy ($p = 0.227$ and $p = 0.034$), respectively. As a result in Figure 6-5, we continue to find statistically significant evidence for accuracy disparities for PCPs but not for BCDs.

6.2.6 User Interaction Design

We do not find any statistically significant differences in whether participants chose to ignore or include suggestions in their differential diagnoses between the control and treatment conditions. However, we find a significant effect of the order of options on participants' choice to update their leading diagnosis with suggestion versus updating their differential diagnosis to include the suggestion. Specifically, we find the treatment condition (with "Update my top prediction" on the bottom) leads participants to select "Update my differential" 9 percentage points ($p < 0.001$) more often and "Update my top prediction" 9 percentage points ($p < 0.001$) less often. Table 6.3 in the Appendix presents regressions showing average treatment effects of the interface randomization on participants' choices to update their differential diagnoses. As a consequence, we find BCD-machine partnerships and PCP-machine partnerships assigned to the treatment condition are 12 percentage points ($p < 0.001$) and 7 percentage points ($p = 0.011$) lower, respectively, in top-1 accuracy than the partnerships assigned to the control condition.

6.3 Discussion

As we move towards a future where algorithms and physicians work collaboratively, it is important to understand the baseline bias of physicians and how algorithms will influence those biases. Using skin disease as a case study, we assessed the baseline accuracy of specialist and generalist physicians in diagnosing skin disease across skin tones in a simulated store-and-forward teledermatology setting. By first establishing a benchmark for diagnostic accuracy of physicians in this well-defined task, we then assessed how specialist and generalist physicians perform with suggestions from a decision support interface based on a deep learning system (DLS).

As a baseline, we find the top-3 diagnostic accuracy of BCDs is 38% and PCPs is 19% (and 42% and 19% for United States based BCDs and PCPs, respectively) on images of inflammatory appearing skin conditions in this experiment. These results match past research demonstrating specialists significantly outperform generalists at skin disease diagnosis but show lower diagnostic accuracy than past studies with different experimental setups [116, 195, 196, 438, 606]. Given our quality control protocol, the post-hoc qualitative review, and the similar error rates across sources, which are described in the Methods, these results cannot be explained by mislabeled images. Instead, our results, which may seem surprising due to the low accuracy rate of specialists on inflammatory appearing skin conditions, are best explained by the difficulty of diagnosing these conditions with free response (as opposed to multiple choice) answers and the differences between this store-and-forward teledermatology setting (where a physician has access to only a single image) and an in-person patient interaction (where a physician has access to much more information such as better lighting, field of view, and ability to inquire about the patient's symptoms, lifestyle, clinical history, family history, and more). While in person clinical visits are the gold standard, image-based store and forward teledermatology has gained traction in triage [579] and can serve as a use case for looking at baseline physician accuracy and physician-AI interaction. Future research in applied machine learning for classifying skin conditions should expand to consider non-visual features in addition to visual features.

We find diagnostic accuracy disparities across patients' skin color for specialists and generalists alike. When comparing participants' three differential diagnoses to the quality controlled skin disease reference labels, we find BCDs and PCPs are four percentage points more accurate on images of light skin (FST 1-4) than dark skin (FST 5-6), and these differences are statistically significant. Given BCDs' and PCPs' accuracy rates of 38% and 19%, respectively, images of dark skin are diagnosed 10% less accurately than images of light skin by BCDs and 22% less accurately by PCPs. These results contribute to an emerging literature on diagnostic accuracy disparities across patient skin color [162, 198] and present evidence that the diagnostic accuracy of medical professionals on images of dark skin is lower than images of light skin.

We find that PCPs who report seeing mostly or all white patients or report insufficient training on skin of color are less accurate on dark skin images than light skin images. This accuracy disparity across skin color for PCPs who see few non-white patients or report insufficient training reveals a gap in healthcare expertise that disproportionately affects people of color. In contrast for BCDs, we do not find evidence that self-reported sufficient training on dark skin or self-reported difficulty with non-white patients is associated with accuracy disparities across images of light and dark skin in this experiment. However, absence of evidence does not mean evidence of absence.

In this experiment, we operationalized the concept of physician-machine partnerships by providing decision support to physicians after they provide an initial differential diagnosis for the skin condition in an image. We find the decision support increases top-1 diagnostic accuracy by 33% for BCDs and 69% for PCPs. However, DLS-based decision support exacerbates diagnostic accuracy disparities in light and dark skin of non-specialists although it does not significantly influence diagnostic accuracy disparities of specialists. These results, though in a limited diagnostic setting, suggest that physician-machine partnerships may improve diagnostic accuracy beyond the performance of unaided physicians but may increase diagnostic accuracy disparities of physicians.

The physician-machine partnerships in the form of physicians interacting with decision support based on a DLS in this experiment led to minimal errors. We find physicians rarely override their leading diagnosis when it is correct, but specialists and generalists can be influenced by the DLS to include incorrect diagnoses in their differential diagnosis. We find that a minor design choice – the order of whether to include a DLS suggestion as a leading diagnosis, one of the diagnoses, or ignore the suggestion – significantly influences participants' choices. This indicates that in addition to the accuracy of the classifier, the presentation interface is an important consideration for human-AI interactions.

6.4 Limitations and Recommendations

This digital experiment for evaluating diagnostic accuracy resembles a store-and-forward teledermatology setting but does not fully match a clinical evaluation in either teledermatology or an in-person examination. A single image contains significantly less information than an in-person interaction (or even a video call), which could include additional visual information (e.g., adjustments in light and angle of view), a patient's symptoms, clinical history, behavioral information, and more. This paper serves as an assessment of physicians' "know what" on a very specific, constrained task where a physician has access to a single image, but not physicians' "know how" [368] of interacting with a patient and diagnosing the patient's condition. Nonetheless, BCDs' top-1 accuracy without decision support remains higher than PCPs' top-1 accuracy with decision support.

Future work should consider diagnostic accuracy in clinical settings and further examine how DLS based decision support compares to collective human intelligence based decision support [108, 450]. In the meantime, physicians should seek additional support in diagnosing dark skin conditions to avoid the potential for systematic misdiagnoses in clinical settings that may mirror the systematic differences found in diagnosing light and dark skin in this experiment.

6.5 Methods

6.5.1 Ethics Approval

This research complies with all relevant ethical regulations, and the Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects determined this study to fall under Exempt Category 3 – Benign Behavioral Intervention. This study's exemption identification numbers are E-2875 and E-3675. All participants who participated in the experiment on `https://diagnosing-diagnosis.media.mit.edu` are informed that "This is an MIT research project. We will first ask 7 brief survey questions. Then, we will show you images of skin conditions and ask you to try to diagnose the skin conditions. After you diagnose conditions in 10 images, we will show you how you perform relative to other healthcare providers. All submissions are collected anonymously for research purposes. For questions, please contact dermatology-diagnosis@mit.edu. Participation is voluntary."

6.5.2 Experimental Interface

We designed and deployed a custom website at `https://diagnosing-diagnosis.media.mit.edu` to host the diagnostic accuracy experiment. Upon clicking the link to our website, participants are directed to the landing page where we provide informed consent and ask several questions as shown in Figure 6-7. After participants fill out the survey, the website directs participants to instructions via a modal window as shown in Figure 6-8. Once participants close the modal, they can begin the experiment as shown in Figure 6-9. All participants see the same first image of a woman with acne, which serves as a relatively easy image to diagnose and a robustness check to confirm participants are participating seriously. Participants are asked, “Can you accurately diagnose this skin condition?” and they are informed how many images they have seen and that they will see how they compare to others after seeing ten images. Participants can provide up to three differential diagnoses, and the three text response forms say “Type leading diagnosis,” “Type secondary differential diagnosis,” and “Type tertiary differential diagnosis.” Participants can move a slider to provide how confident they are from 0% confident to 100% confident. In addition, participants (with the exception of BCDs) are asked to check the boxes for whether they would refer the patient for a biopsy or a dermatologist for a second opinion.

When a participant begins to type their diagnosis in the free response text boxes, predictive text appears as shown in Figure 6-10. We designed this experiment with free responses instead of multiple choice responses to maintain as much ecological validity to clinical practice as possible. Free response is more difficult than multiple choice for two main reasons: first, multiple choice enables correct answers via uninformed guessing whereas free responses do not, and second, multiple choice primes the participant on what a particular condition might be whereas free responses do not. We supported free responses with predictive text based on 445 possible diagnoses to promote standardized responses. These 445 diagnoses include the 46 skin conditions in this experiment, the 419 skin conditions in Liu et al 2020 [388], which have significant overlap with the skin conditions in this experiment, and similar clinical terms for skin conditions. Three examples of similar clinical terms include atopic dermatitis and eczema, cutaneous t-cell lymphoma and mycosis fungoides, and lyme disease and erythema migrans. The predictive text appears as a function of the first characters typed, and in order to encourage participants to choose from the list, we attempted to include as many ways of writing conditions as possible (e.g. “erythema migrans (Lyme)” and “lyme (erythema migrans)” or “ctcl (cutaneous t-cell lymphoma)” and “cutaneous t-cell lymphoma (ctcl).”

Once a participant clicks submit (and assuming the participants’ differential diagnosis differs from the AI’s prediction), the website directs participants to a page showing the AI’s prediction. Participants have three options: “Keep my differential,” “Update my differential to include [suggestion condition],” or “Update my top prediction with [suggested condition]” as shown in Figure 6-11. Next (or if the participant’s differential matched the suggestion), the website directs participants to a page offering feedback on what the correct diagnosis is and what the most common incorrect diagnosis for this image was as shown in Figure 6-12. When participants click “Next Image” on the feedback page, participants are redirected to a page that looks like Figure 6-9 with a different image and the experiment repeats for as long as a participant is willing to participate. After a participant sees 10 images, we show participants a bar graph showing how diagnostic accuracy compares across the DLS, specialists, and generalists.

6.5.3 Clinical Image Curation

The experiment contains 364 images of 46 different skin conditions. The vast majority of images show eight relatively common skin conditions; there are 31 images of atopic dermatitis, 48 of cutaneous t-cell lymphoma, 34 of dermatomyositis, 30 of erythema migrans (lyme disease), 32 of lichen planus, 33 of pityriasis rosea, 47 of pityriasis rubra pilaris, and 29 of secondary syphilis. We decided to focus our analysis on these 8 conditions based on three criteria: first, three practicing BCDs identified these conditions as the most likely conditions on which we may find accuracy disparities across patients' skin color; second, these conditions are relatively common, and third, these conditions appear frequently enough in dermatology textbooks and dermatology image atlases such that we could select at least 5 images of the two darkest skin types after applying a quality control review by BCDs. We sourced the 284 images of the eight conditions based on 241 publicly available images online from dermatology atlases and search engines, 30 images from 14 textbooks, and 13 images from dermatologists' slides and education material [16, 22, 28, 33, 39, 73, 89, 96, 155–157, 173, 184, 215, 241, 280, 290, 310, 339, 466, 469, 525, 548, 585, 613]. The number of images from each source is provided in Table 6.1 in the Appendix.

The remaining 80 images represent 38 skin conditions and are all drawn from the Fitzpatrick 17k dataset [248] except for the attention check, which is sourced from a magazine article on inflammatory conditions in dark skin [329]. We included these additional conditions primarily to promote the ecological validity of the experiment. In particular, we designed this experiment such that participants do not know which skin conditions will appear in the experiment, and as such, participants cannot simply treat this as a multiple-choice test. Beyond the eight conditions of direct interest, there are 8 images of scleroderma, 6 of lupus erythematosus, 6 of acne, 4 of vitiligo, 3 of rosacea, 3 of tungiasis, 3 of urticaria pigmentosa, 3 of sarcoidosis, 2 of cheilitis, 2 of calcinosis cutis, 2 of allergic contact dermatitis, 2 of factitial dermatitis, 2 of fixed eruptions, 2 of granuloma annulare, 2 of keloid, 2 of keratosis pilaris, 2 of acanthosis nigricans, 2 of rhinophyma, 2 of necrobiosis lipoidica, 2 of tick bite, 2 of papilomatosis confluentes and reticulate, 2 of psoriasis, 2 of scabies, 1 of livedo reticularis, 1 urticaria, 1 of Steven Johnson syndrome, 1 of statis edema, 1 of seborrheic dermatitis, 1 of erythema nodosum, 1 of erythema elevatum diutinum, 1 of lichen simplex, 1 of neurotic excoriations, 1 of hidradenitis, 1 of nematode infection, 1 of lichen amyloidosis, and 1 of xanthomas.

We curated the images of skin conditions via the following five steps. First, we collected all images of the eight skin conditions from online sources and textbooks and the attention check image from an online magazine. Second, we annotated images with estimated Fitzpatrick skin type (FST) labels. One BCD curated 351 of the highest quality images of the eight conditions of interest for each of the six Fitzpatrick skin types by dragging and dropping images into folders on their computer specifying the skin condition and FST label. Due to a lack of images of secondary syphilis in light-skin instances and lyme disease in dark skin, this first BCD supplemented the dataset with 11 images from their educational materials. Third, a second BCD reviewed the initially selected images and identified 66 images as low-quality due to image resolution or questions about the diagnostic label. We removed these 66 images from the data set to leave 285 images of the eight conditions remaining. Fourth, we added 79 images of 38 skin conditions from the Fitzpatrick 17k dataset that have been reviewed and assessed by two BCDs as high-quality and diagnostic of the underlying condition. Fifth, a third BCD reviewed the images and found no clear objections.

While the gold standard label for skin conditions such as cutaneous malignant neoplasm is histopathological diagnosis [141], the majority of non-neoplastic skin conditions (including skin conditions) are considered readily diagnosable with an in-patient exam and a patient's clinical history [269]. The images in this experiment come from external sources (textbooks, dermatology atlases, online search engines, and dermatologist education materials) and were curated and confirmed to be correctly labeled by three BCDs to the best of their knowledge based on the visual features in the images.

As a post-hoc quality review, three board certified dermatologists reviewed the three most and least accurately diagnosed images for light and dark skin in each of the eight skin conditions. The analysis of these images by three BCDs indicates that the most accurately diagnosed images appear to be relatively classic presentations of each skin condition (e.g. a heliotrope sign and gottron's papules for dermatomyositis, rashes of the hands and feet for secondary syphilis, bullseye rash for Lyme) while the least accurately diagnosed images appear to be atypical presentations.

As an additional quality control measure, we also present Table 6.1 in the Appendix to summarize the sources upon which we draw these images and how accurately BCDs identify the reference label across sources. For images of the main 8 conditions that no BCD diagnosed correctly, 15% of those images come from dermatology textbooks. This is slightly higher than the proportion of textbook images in the 284 images of the 8 conditions, which is 11%.

6.5.4 Skin Tone Annotations

We annotated images by initially hiring crowdworkers to provide estimated FSTs for each image and then asking BCDs to update the FST label appropriately. The images are relatively balanced across FST; 32% of images show people with the two darkest FST (FST 5-6) and 68% of images show people with the four lightest FST (FST 1-4). We compare the two darkest FST to the four lightest FST because the original FST scale indicates FST 1-4 as "white" and FST 5 and 6 as "black" and "brown." Our findings are robust to comparisons between the three lightest and three darkest skin conditions and comparisons between the two lightest and two darkest skin conditions. We note the Fitzpatrick skin type scale is imperfect (and its imperfections have been widely discussed [87, 246, 437, 472, 632]) but remains a useful starting point for examining diagnostic accuracy disparities across skin color.

6.5.5 Deep Learning System Development

In order to offer computer vision-based predictions of diagnoses, we trained a convolutional neural network to classify nine labels: the eight skin conditions of interest and an other category. This neural network is a VGG-16 architecture pre-trained on ImageNet, which is the same general architecture as Esteva et al 2017 and the identical architecture of Groh et al 2021 [187, 248]. Following insights that fine-tuning on diverse data can close performance gaps between light and dark skin tones [146], we fine-tuned the model on 31,219 diverse clinical dermatology images which come from the Fitzpatrick 17k dataset and an additional collection of images collected from textbooks, dermatology atlases, and online search engines. The fine-tuning includes a number of transformations to images including randomly resizing images to 256x256 pixels, randomly rotating images 0 to 15 degrees, randomly altering the brightness, contrast, saturation, and hue of each image, randomly flipping the image horizontally or not, center cropping the image ot be 224x224 pixels, and normalizing the image arrays by the ImageNet means and standard deviations.

We evaluated the model on the 364 images in this experiment, which neither appear in the pre-training ImageNet data nor in the fine-tuning clinical dermatology images dataset, and we find the model is 47% accurate at predicting the nine labels on the 364 images.

We do not compare the DLS system directly to physician performance because the DLS system is trained to classify only nine labels whereas physicians are tasked with diagnosing images without knowing the short list of what the possible skin conditions might be.

In this experiment, we refer to the VGG-16 architecture pre-trained on ImageNet and fine-tuned on 31,219 clinical dermatology images as the control DLS.

In addition to the control DLS, we consider a treatment DLS, which is Wizard of Oz classifier that is a synthetically enhanced version of the control DLS. In order to create the treatment DLS, we randomly re-assigned 65% of wrong classifications by the control DLS to be correct classifications, which resulted in 84% top-1 accuracy.

We note that the control and treatment DLS are fair classifiers from a disparate impact perspective. Both classifiers have relatively similar top-1 accuracy across skin tones on the eight conditions; the control DLS is 58% accurate on dark skin and 56% accurate on light skin and the enhanced DLS is 82% accurate on dark skin and 84% accurate on light skin.

Following the MI-CLAIM [463] checklist, we examine the control DLS performance with two examination techniques. First, specialists examined the model's performance across images and find that correct predictions often (but not always) correspond to classic presentations of a disease. Second, we examined the model's performance across FST and we do not find meaningful differences in the model's performance across skin types. In the context of the visual diagnosis of skin disease task, we did not find saliency maps particularly helpful for interpretability because they highlighted skin lesions but did not provide any additional information on what differentiates one skin lesion from another.

6.5.6 Randomization Protocol

We randomly assigned the order in which images appear to participants for all images except the first. All participants see the same first image, and all subsequent images are drawn randomly from the remaining images.

We conducted two randomized experiments where participants were assigned to control and treatment conditions. We randomly assigned participants to see suggestions from a control model (the 47% accurate model) or a synthetically enhanced treatment model (the 84% accurate model). We also randomly assigned the order in which the options appear for including or ignoring the suggestion in a participant's differential diagnosis. The treatment group saw "Keep my differential" on top and "Update my top prediction with [condition]" on the bottom as seen in Figure 6-11 whereas the control group saw the opposite where "Update my top prediction with [condition]" appeared on the top.

6.5.7 Participants

We recruited participants by word-of-mouth and direct emails by Sermo, a secure digital (online) platform designed for physician networking and anonymous survey research, to their verified physician network. Sermo sent emails to 7,900 BCDs and 10,000 PCPs and offered $10 for BCDs and $5 for PCPs to complete the survey (see 6-13 for a copy of the invitation email). 68% of BCDs and 94% of PCPs in this experiment came from Sermo and the rest came from authors reaching out to other physicians via email and social media. We recruited dermatology residents by identifying the email addresses of dermatology resident coordinators at 142 programs across the United States and requesting coordinators to forward an invitation to residents to participate in this study.

The countries with more than 10 participants include: United States (551 total with 167 BCDs, 47 dermatology residents, 295 PCPs, and 42 other physicians), India (134 total with 67 BCDs, 15 dermatology residents, 20 PCPs, and 32 other physicians), Canada (91 total with 18 BCDs, 1 dermatology resident, 59 PCPs, and 13 other physicians), United Kingdom (53 total with 18 BCDs, 3 dermatology residents, 25 PCPs, and 7 other physicians), Italy (45 total with 13 BCDs, 18 dermatology residents, 6 PCPs, and 8 other physicians), Germany (35 total with 16 BCDs, 8 dermatology residents, 5 PCPs, and 6 other physicians), Nigeria (30 total with 3 dermatology residents, 6 PCPs, and 21 other physicians), Brazil (22 total with 11 BCDs, 4 dermatology residents, 5 PCPs, and 2 other physicians), Spain (21 total with 19 BCDs and 2 dermatology residents), Australia (18 total with 3 BCDs, 1 dermatology resident, 8 PCPs, and 6 other physicians), France (14 total with 5 BCDs, 2 dermatology residents, 3 PCPs, and 4 other physicians), and South Africa (14 total with 3 BCDs, 7 PCPs, and 4 other physicians).

In the pre-experiment survey, we asked physicians how many years they have practiced medicine, what is the distribution of their patients' skin color, what is the frequency of difficulty for diagnosing skin conditions in white and non-white patients, and how do they view the training they received for diagnosing skin conditions in patients with skin of color. In this experiment, 40% of physicians have been practicing medicine for 20 years or more, 26% have been practicing for 10 to 20 years, 22% have been practicing for 2 to 10 years, 3% have been practicing for 0 to 2 years, and the rest are doing residencies, fellowships, or internships. In response to the question "How would you describe the distribution of your patients' skin colors?", 32% of participants responded about an equal portion of white and non-white patients, 43% responded mostly white patients, 2% responded all white patients, 15% responded mostly non-white, 7% responded all non-white patients, and 1% responded that the question is not applicable. This overall distribution is similar but slightly more diverse than the distribution for participants from the United States, which is skewed slightly more towards mostly white patients with 49% mostly white patients, 36% equal portion of white and non-white patients, and 13% mostly or all non-white patients.

We find PCPs report significantly higher rates of difficulty in diagnosing skin conditions for both light and dark skin than BCDs. Specifically, we find 8% of PCPs report difficulties diagnosing skin conditions in one in two white patients and 15% of PCPs report difficulties diagnosing skin conditions in one in two non-white patients while less than 3% of BCDs report difficulties diagnosing skin conditions in one in two patients of any skin color. For participants in the United States, 70% of BCDs and 72% of PCPs report the same diagnostic difficulty between white and non-white patients while 10% of BCDs and 20% of PCPs reporting more difficulties in diagnosing non-white patients compared to white patients. When asked, "Do you feel you received sufficient training for diagnosing skin conditions in patients with skin of color (non-white patients)?" 67% of all PCPs respond no and 33% of all BCDs respond no (similarly, 68% of United States PCPs respond no and 28% of United States BCDs respond no).

6.5.8 Annotating Participants' Differential Diagnoses

We collected 14,261 differential diagnoses, which include 2,348 unique text strings. As a function of our experimental interface which asked participants to provide differential diagnoses in free response text boxes supported by predictive text, 43% of the leading diagnosis text strings do not exactly match any of the text strings in the initial list of 445 conditions. However, the majority of these off-list responses are easily matched to the list. For example, 14% of the 14,261 leading diagnoses are "atopic dermatitis" which we match to "atopic dermatitis (eczema)" in the list, 4% of participants submitted "Lyme" which we match to "lyme (erythema migrans)" in the list, 3% of participants submitted "pityriasis rubra pilaris" which we match to "pityriasis rubra pilaris (prp)" in the list, and 3% of participants submitted "cutaneous t-cell lymphoma" which we match to "cutaneous t-cell lymphoma (ctcl)" in the list). The remaining 19% of leading diagnoses match 1,447 unique text strings. In order to evaluate diagnostic accuracy as accurately as possible, we reviewed all diagnoses and marked responses as correct if they appear to be misspellings or shorthand for the correct answer. For example, we included the following answer as correct for lichen planus: lichen planus, lichen ruber planus, lichens planus, lichen plan, lichen planes, lichen planhs, lichen planis, lichen plannus, lichen plans, lichen planus linearis, lichen planus., luchen planus, lichen_planus, lichen plane, linear lichen planus, linen planu, and liquen plano. As a second example, we included the following answers as correct for cutaneous t-cell lymphoma: cutaneous t-cell lymphoma, t cell lymphoma, cutaneous t cell lymphoma, cutaneous t cell, ctcl, mycosis fungoides, lymphoma, mucosità fungoide, micosi fungoide, myocses fungoides, mycosis fungiodies, mycoses fungoides, plaque type, mf, cuttaneoua t-cell lymph, linfoma, linfoma células t,linfoma t, lmphoma, lymphome, malignant skin cancer, t cell lyphoma, t-cell lyphoma, mucosis fungoides, mycoses fungoides, mycoses glfungoide, mycosis, mycosis fongicide, mycosis fungoides/ctcl, mycosis fungoidis, mycosis fungoidus, micose fungoide, micosis fungoide, micosis fungoides, cutaneous t-cell lymphoma (ctcl), ctcl (cutaneous t-cell lymphoma), cutaneous_t-cell_lymphoma, t-cell lymphoma, cutaneous lymphoma, and cutaneous lympoma.

6.5.9 Standards for Reporting Diagnostic Accuracy Studies (STARD)

The updated STARD 2015 guidelines are designed to help readers of diagnostic accuracy studies recognize for which patient groups and settings a diagnostic accuracy study is relevant [79, 125]. While this study focuses on physician diagnostic accuracy, which differs significantly from standard diagnostic accuracy studies which focus on medical test accuracy, we followed the STARD 2015 checklist to clarify the study objectives, experimental design, analysis, limitations, and implications for clinical dermatology practice and designing physician-machine partnerships.

6.6 Data and Code Availability

The images and data collected during the study and the code to reproduce the results of this study are available in our Github repository, https: //github.com/mattgroh/diagnosis-diagnosis, which will be set to public upon peer-reviewed publication.

Author's contributions

M.G., O.B., R.D., A.K., L.S., and M.D. conceived the experiments, M.G., O.B., R.D., A.K., C.H., and L.S. curated the stimulus set, M.G. analyzed the data, M.G. wrote the initial draft, M.G., C.H., R.D., L.S., A.K., O.B., M.D., and R.P. reviewed and edited the manuscript.

Acknowledgments

The authors would like to graciously acknowledge Sermo for platform support to recruit physicians to participate in this experiment, Apollo Hospitals for forwarding invites to their physicians, MIT Media Lab member companies and the Harold Horowitz (1951) Student Research Fund for financial support, the participants for their time and care participating in this experiment, and Bruke Wossenseged for excellent research assistance in the early phases of this paper.

Image Source	P	N
Textbook (Stratigos 2009)	0.00	1
Textbook (Oakley 2017)	0.00	1
Textbook (Du Vivier 2002)	0.33	3
Textbook (Bolognia 2018)	0.50	2
Dermatologist Education Material	0.55	11
DermisNet	0.57	7
Textbook (James 2020)	0.67	3
Regional Derm	0.67	3
Textbook (Archer 2008)	0.67	3
Textbook (Callen 1993)	0.67	3
Textbook (Buxton 2009)	0.67	3
Derma Amin	0.71	85
Atlas Dermatologico	0.72	95
Dermato Web Net	0.73	15
Textbook (Usatine 2009)	0.75	4
Enzyklopaedie Dermatologie	0.75	4
Hellenic Derm Atlas	0.75	4
Google Derm	0.77	22
Bing Derm	0.88	43
Dermnet	0.92	13
AAD Slides	1.00	2
Textbook (Wolf 2017)	1.00	1
Textbook (Griffiths 2016)	1.00	2
Textbook (Nouri 2008)	1.00	1
Textbook (Salzman 2020)	1.00	2
Textbook (Knoop 2010)	1.00	1
Danderm	1.00	2
Derm 101	1.00	4
Dermnet NZ	1.00	13
Iconique	1.00	1
SD198	1.00	1

Table 6.1: Table of image sources with P indicating the proportion of images from a particular source in which at least one board-certified dermatologist provided a top-3 diagnosis matching the source image's label. N indicates the number of images from each source.

	Top-1	Top-4
	(1)	(2)
Constant	0.15***	0.20***
	(0.01)	(0.01)
Specialist	0.13***	0.19***
	(0.02)	(0.02)
Dark Skin	-0.03**	-0.04**
	(0.01)	(0.01)
DLS Assistant	0.08***	0.25***
	(0.02)	(0.02)
Enhanced DLS Assistant	0.08**	0.16***
	(0.03)	(0.03)
DLS Assistant * Specialist	-0.01	-0.09***
	(0.02)	(0.03)
Enhanced DLS Assistant * Specialist	-0.03*	-0.06*
	(0.04)	(0.04)
DLS Assistant * Dark Skin	-0.05**	-0.05*
	(0.02)	(0.02)
Enhanced DLS Assistant * Dark Skin	0.05*	0.01
	(0.03)	(0.03)
Specialist * Dark Skin	-0.01	0.01
	(0.02)	(0.02)
DLS Assistant * Specialist * Dark Skin	0.04*	0.08*
	(0.03)	(0.04)
Enhanced DLS Assistant * Specialist * Dark Skin	-0.03	-0.07*
	(0.05)	(0.05)
Observations	11,619	11,619
Number of Board-Certified Dermatologists	296	296
Number of Primary Care Physicians	350	350
R^2	0.04	0.11
Note:	*p<0.5; **p<0.01; ***p<0.001	

Table 6.2: Ordinary Least Squares regressions with robust standard errors clustered on physician participants. This regression includes only board-certified dermatologist (BCD) and primary care physician (PCP) participants. We code the *Specialist*, *Dark Skin*, *AI Assistance* binary variables as follows: *Specialist* equals 1 for BCDs and 0 for PCPs, *Dark Skin* equals 1 for FST 5 and 6 and 0 for FST 1 to 4, and *AI Assistance* equals 1 for participant responses with access to the DLS prediction and 0 for participant responses before access to the DLS predictions.

	No Update	Update Differential	Update Leading
	(1)	(2)	(3)
Constant	0.47***	0.28***	0.24***
	(0.02)	(0.01)	(0.02)
Keep My Differential on Top	-0.00	0.09***	-0.09***
	(0.02)	(0.02)	(0.02)
Observations	5,982	5,982	5,982
R^2	0.00	0.01	0.01

Note: *p<0.05; **p<0.01; ***p<0.001

Table 6.3: Average treatment effects of user interaction design with "Keep My Differential" on top based on ordinary least squares regressions with robust standard errors clustered on participants.

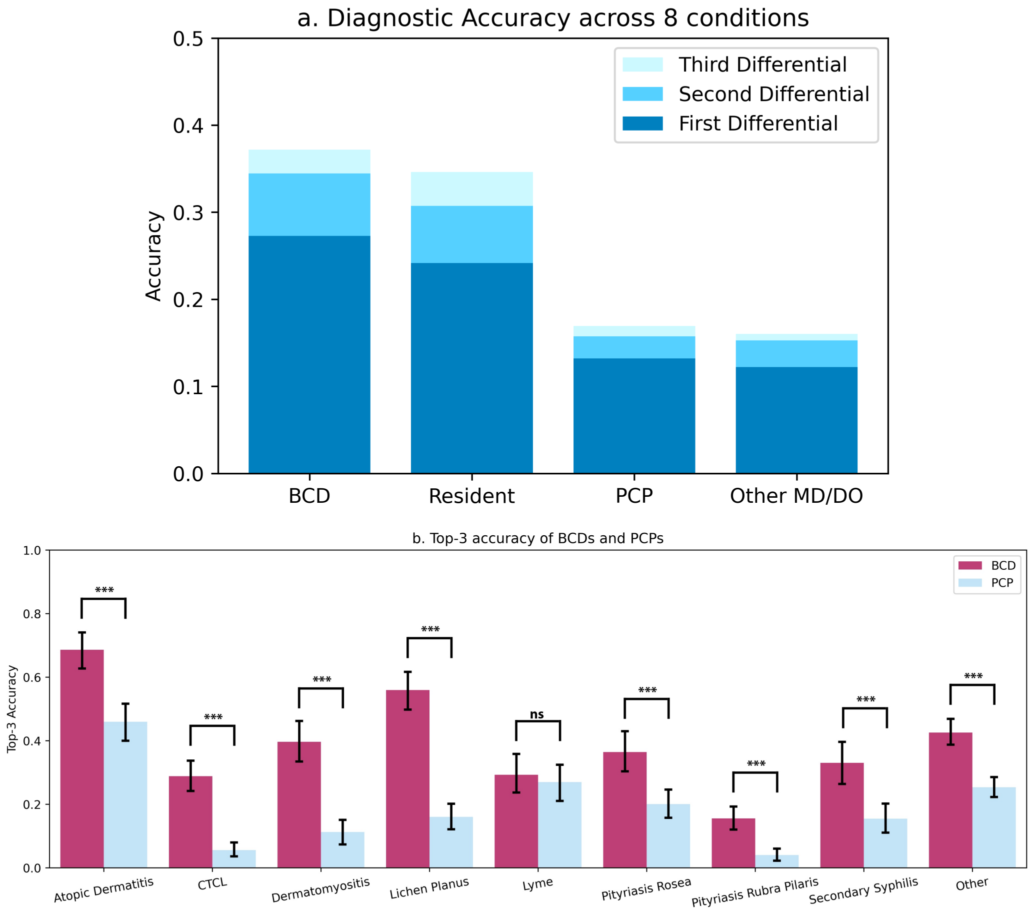


Figure 6-2: Top: Diagnostic accuracy of physician participants on the eight main inflammatory skin conditions in this experiment with shades of blue indicating the diagnostic accuracy of the first, second, and third differential, respectively. BCD and PCP are acronyms for board-certified dermatologist and primary care physician; residents refer strictly to dermatology residents. Bottom: Top-3 diagnostic accuracy of BCDs and PCPs on each of the eight main skin conditions and the auxiliary 38 conditions aggregated in the other category. The error bars represent the 95% confidence interval of the true mean. *** indicates the p-value is less than 0.001 and ns indicates the p-value is greater than 0.05.

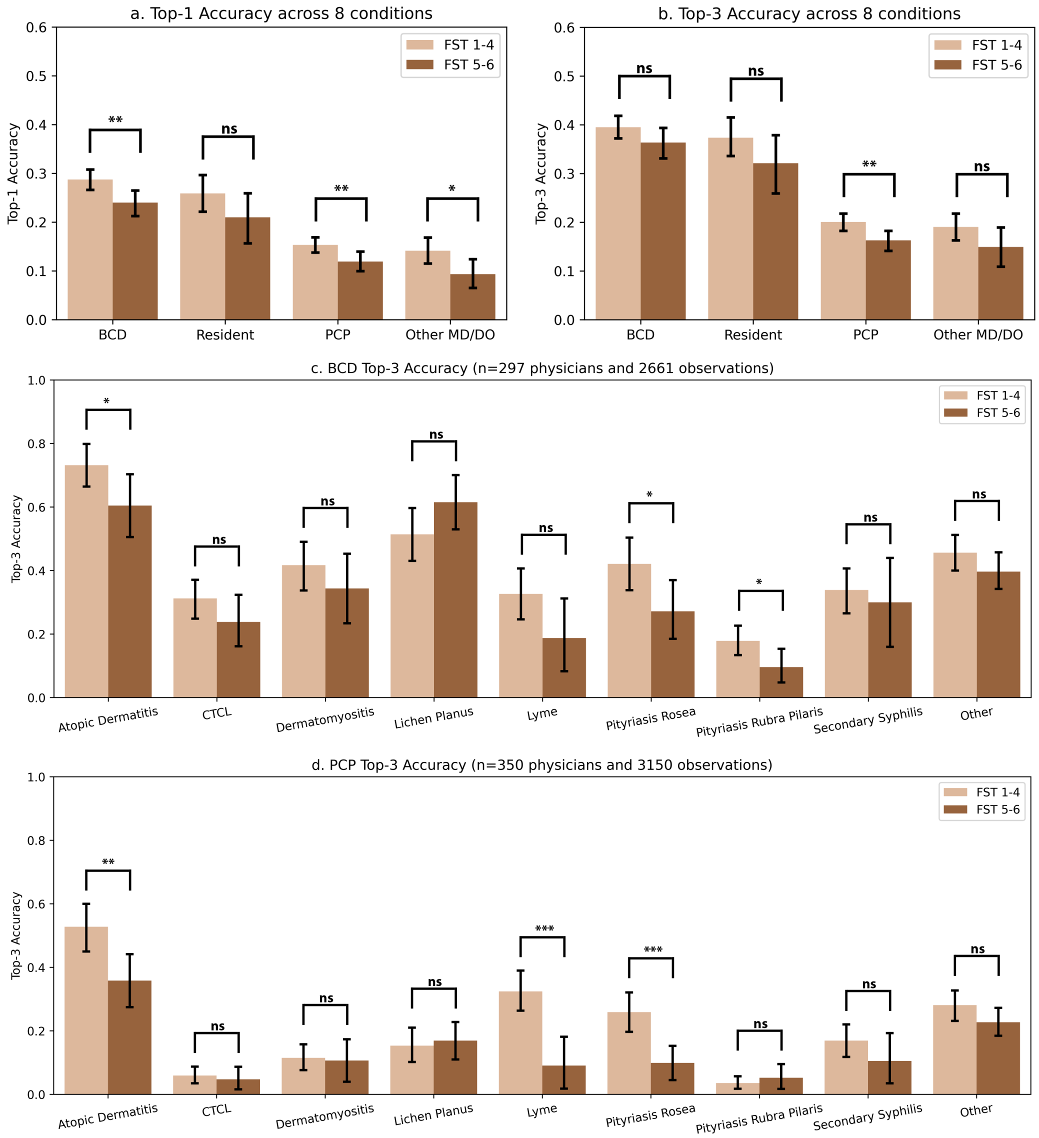


Figure 6-3: Top left: Top-1 diagnostic accuracy of physician participants – board certified dermatologists (BCDs), dermatology residents, primary care physicians (PCPs), and other physicians – across estimated Fitzpatrick skin types (FSTs) on the eight main inflammatory conditions. The error bars represent the 95% confidence interval of the true mean. *, **, and *** indicates the p-value is less than 0.05, 0.01, and 0.001, respectively and ns indicates the p-value is greater than 0.05. Top right: Top-3 diagnostic accuracy of physician participants across estimated FSTs on the eight main inflammatory conditions. Middle: Top-3 diagnostic accuracy of BCDs across skin diseases and FSTs. Bottom: Top-3 diagnostic accuracy of PCPs across skin diseases and FSTs.

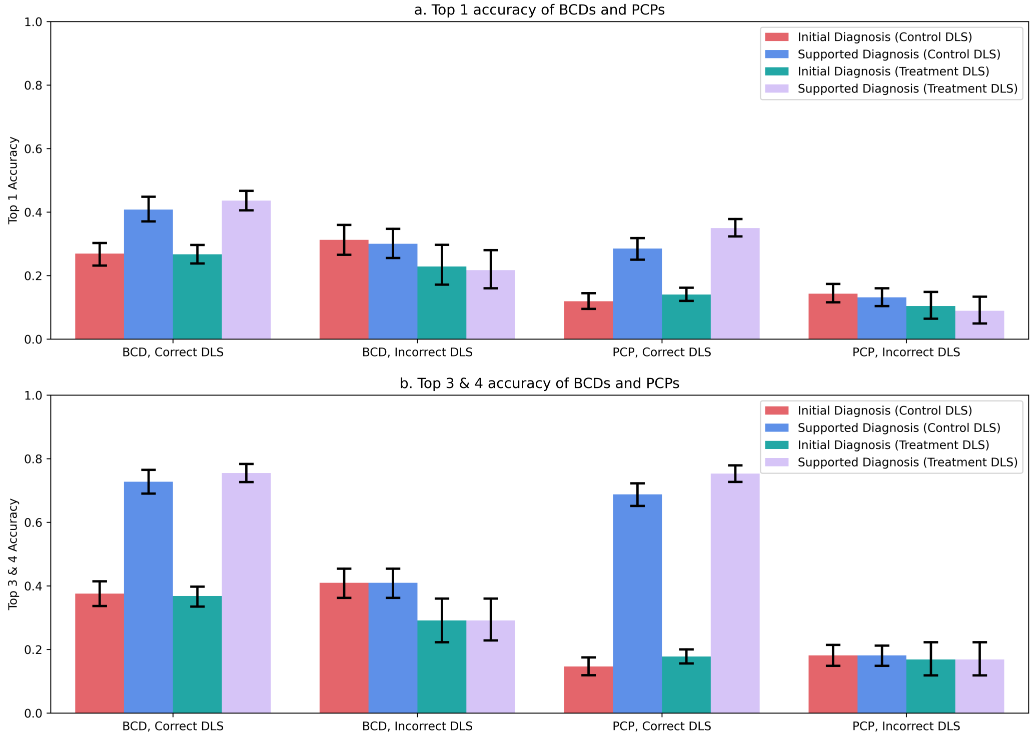


Figure 6-4: Top: Top-1 accuracy for physicians before and after seeing either the control or treatment deep learning system (DLS) suggestion. Bottom: Top-3 and top-4 accuracy for physician before and after seeing the control or treatment DLS suggestion. BCD and PCP refer to board-certified dermatologist and primary care physician, respectively, and resident refers to dermatology resident. The error bars represent the 95% confidence interval of the true mean.

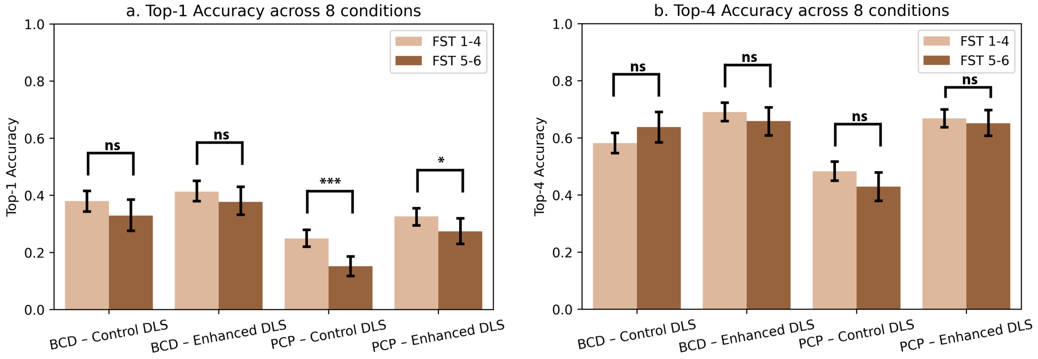


Figure 6-5: Top-1 and Top-4 accuracy of physician-machine partnerships across light and dark skin. *, **, and *** indicates the p-value is less than 0.05, 0.01, and 0.001, respectively and ns indicates the p-value is greater than 0.05

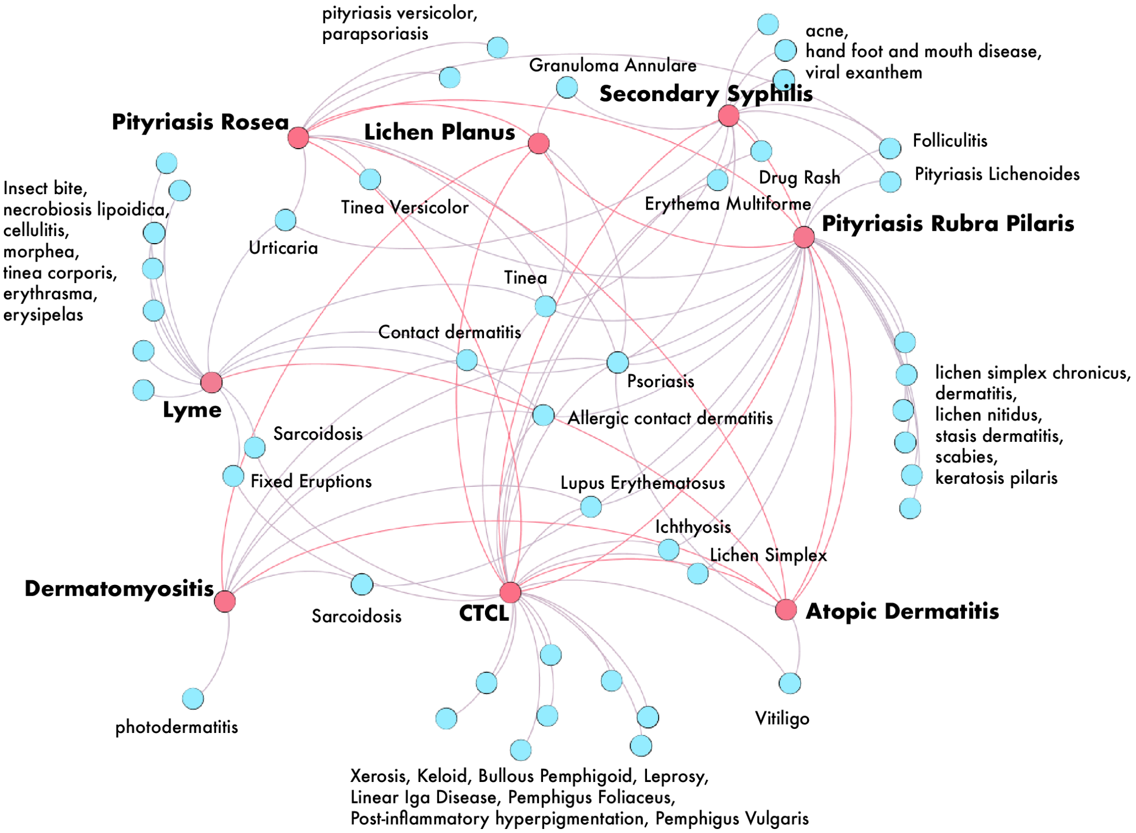
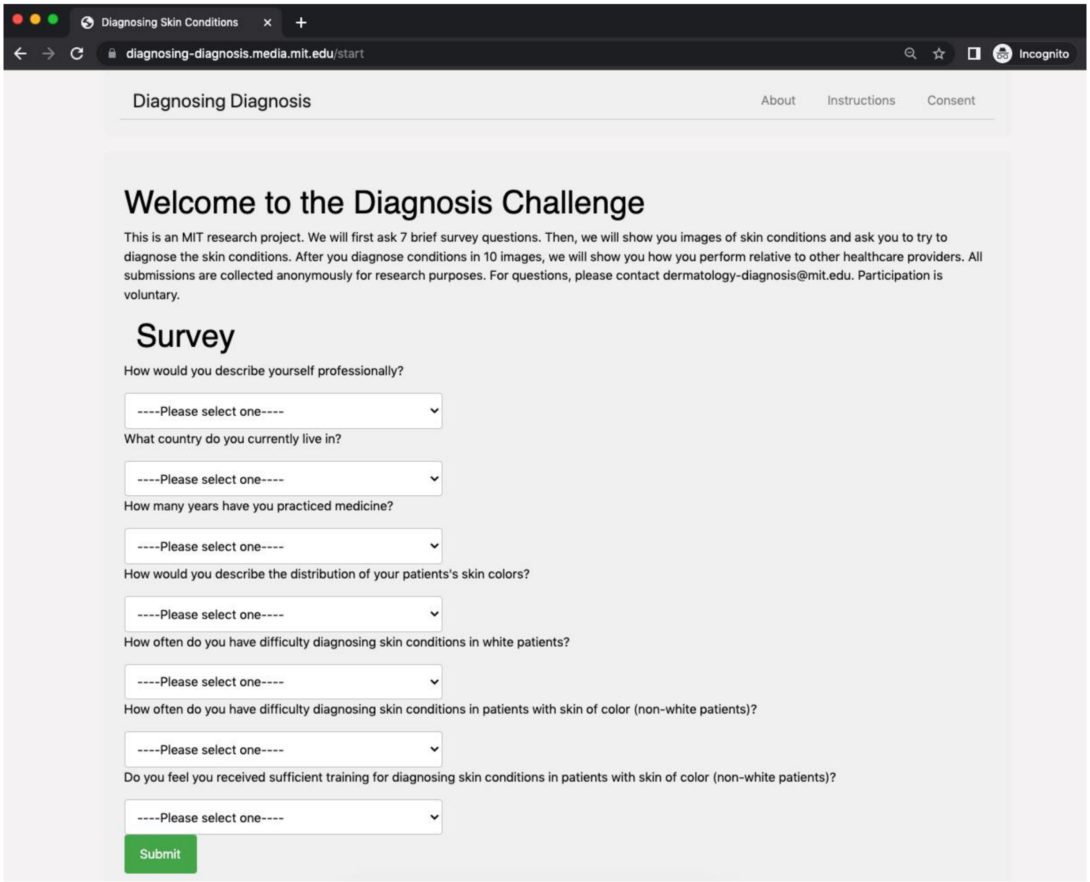


Figure 6-6: Network graph of eight inflammatory skin conditions (red) and the most common conditions listed in differential diagnoses (blue) for images labeled with these eight inflammatory skin conditions: Lyme, dermatomyositis, pityriasis rubra pilaris, lichen planus, atopic dermatitis, pityriasis rosea, secondary syphilis, and cutaneous t-cell lymphoma (CTCL). We show all conditions listed in differential diagnoses by board-certified dermatologists that appeared at least five times for each skin condition category.


Diagnosing Skin Conditions
diagnosing-diagnosis.media.mit.edu/start
Incognito
Diagnosing Diagnosis
About
Instructions
Consent

Welcome to the Diagnosis Challenge

This is an MIT research project. We will first ask 7 brief survey questions. Then, we will show you images of skin conditions and ask you to try to diagnose the skin conditions. After you diagnose conditions in 10 images, we will show you how you perform relative to other healthcare providers. All submissions are collected anonymously for research purposes. For questions, please contact dermatology-diagnosis@mit.edu. Participation is voluntary.

Survey

How would you describe yourself professionally?

----Please select one----

What country do you currently live in?

----Please select one----

How many years have you practiced medicine?

----Please select one----

How would you describe the distribution of your patients's skin colors?

----Please select one----

How often do you have difficulty diagnosing skin conditions in white patients?

----Please select one----

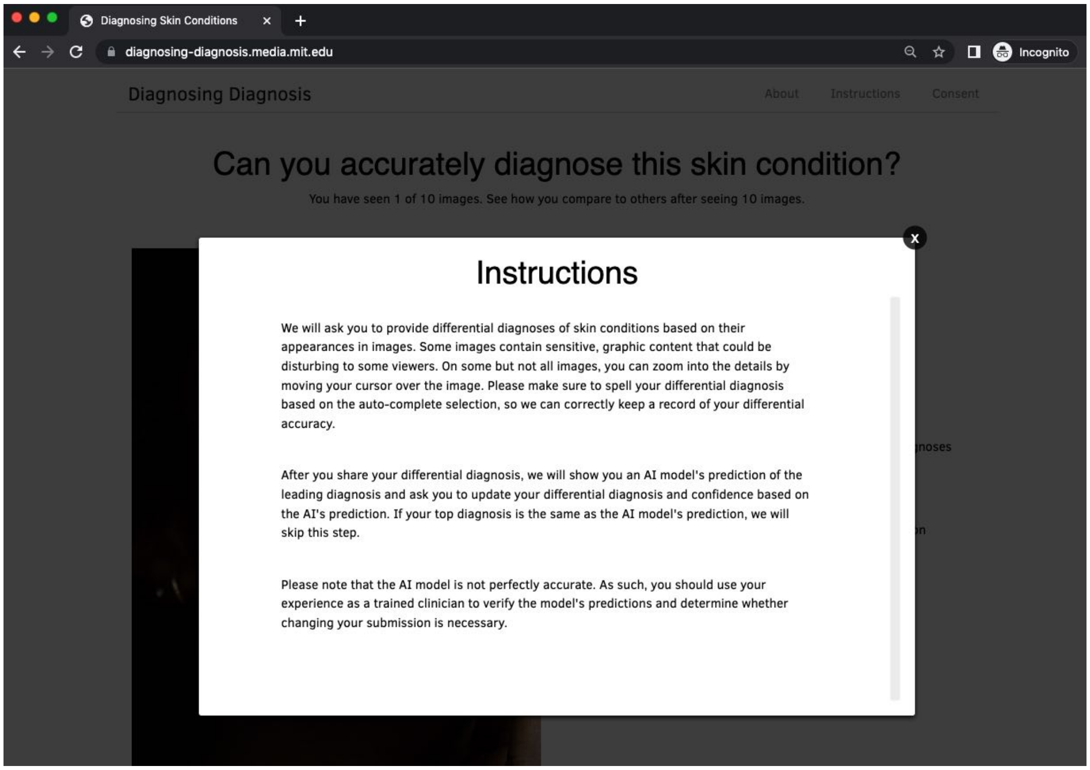
How often do you have difficulty diagnosing skin conditions in patients with skin of color (non-white patients)?

----Please select one----

Do you feel you received sufficient training for diagnosing skin conditions in patients with skin of color (non-white patients)?

----Please select one----

Figure 6-7: Screenshot from the Diagnosing Diagnosis experiment website showing the welcome landing page


Diagnosing Skin Conditions
diagnosing-diagnosis.media.mit.edu
Incognito
Diagnosing Diagnosis
About
Instructions
Consent
x

Can you accurately diagnose this skin condition?

You have seen 1 of 10 images. See how you compare to others after seeing 10 images.

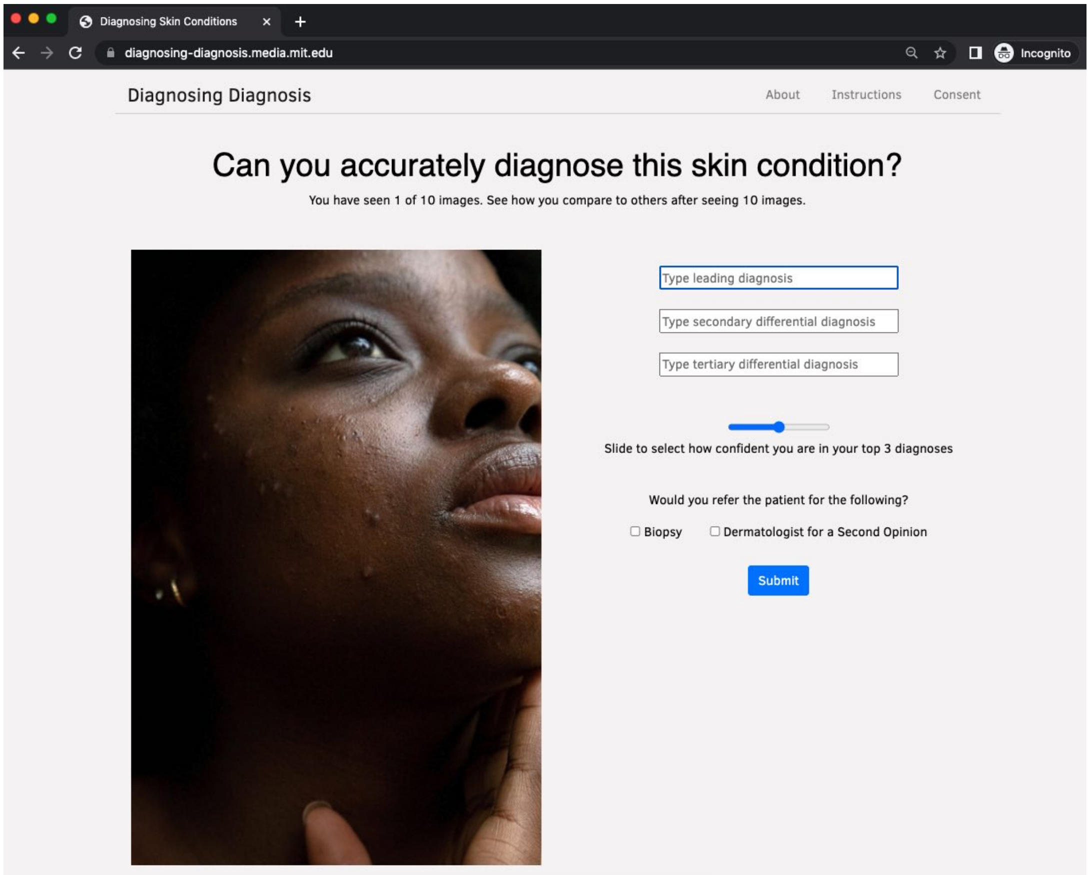
Instructions

We will ask you to provide differential diagnoses of skin conditions based on their appearances in images. Some images contain sensitive, graphic content that could be disturbing to some viewers. On some but not all images, you can zoom into the details by moving your cursor over the image. Please make sure to spell your differential diagnosis based on the auto-complete selection, so we can correctly keep a record of your differential accuracy.

After you share your differential diagnosis, we will show you an AI model's prediction of the leading diagnosis and ask you to update your differential diagnosis and confidence based on the AI's prediction. If your top diagnosis is the same as the AI model's prediction, we will skip this step.

Please note that the AI model is not perfectly accurate. As such, you should use your experience as a trained clinician to verify the model's predictions and determine whether changing your submission is necessary.

Figure 6-8: Screenshot from the Diagnosing Diagnosis experiment website showing the instructions


Diagnosing Skin Conditions
diagnosing-diagnosis.media.mit.edu
Incognito
Diagnosing Diagnosis
About
Instructions
Consent

Can you accurately diagnose this skin condition?

You have seen 1 of 10 images. See how you compare to others after seeing 10 images.

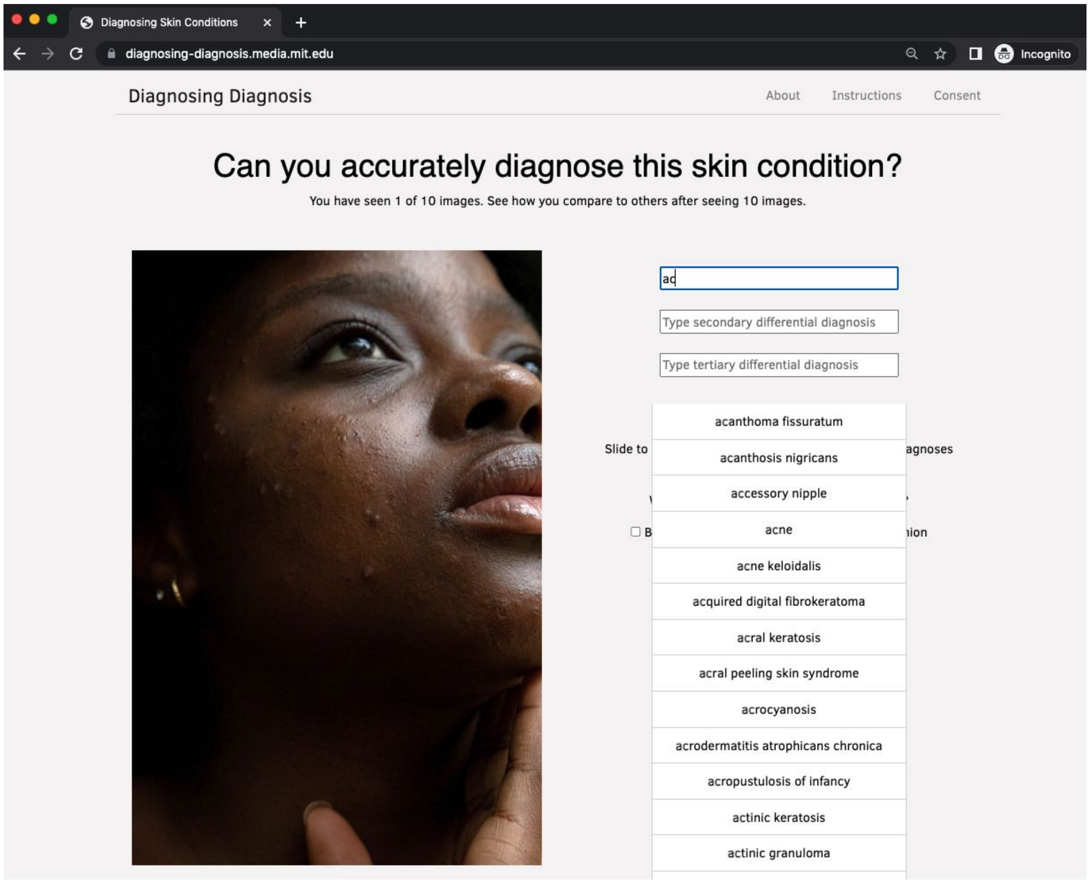
Type leading diagnosis
Type secondary differential diagnosis
Type tertiary differential diagnosis

Slide to select how confident you are in your top 3 diagnoses

Would you refer the patient for the following?

□ Biopsy □ Dermatologist for a Second Opinion

Figure 6-9: Screenshot from the Diagnosing Diagnosis experiment website showing diagnostic task

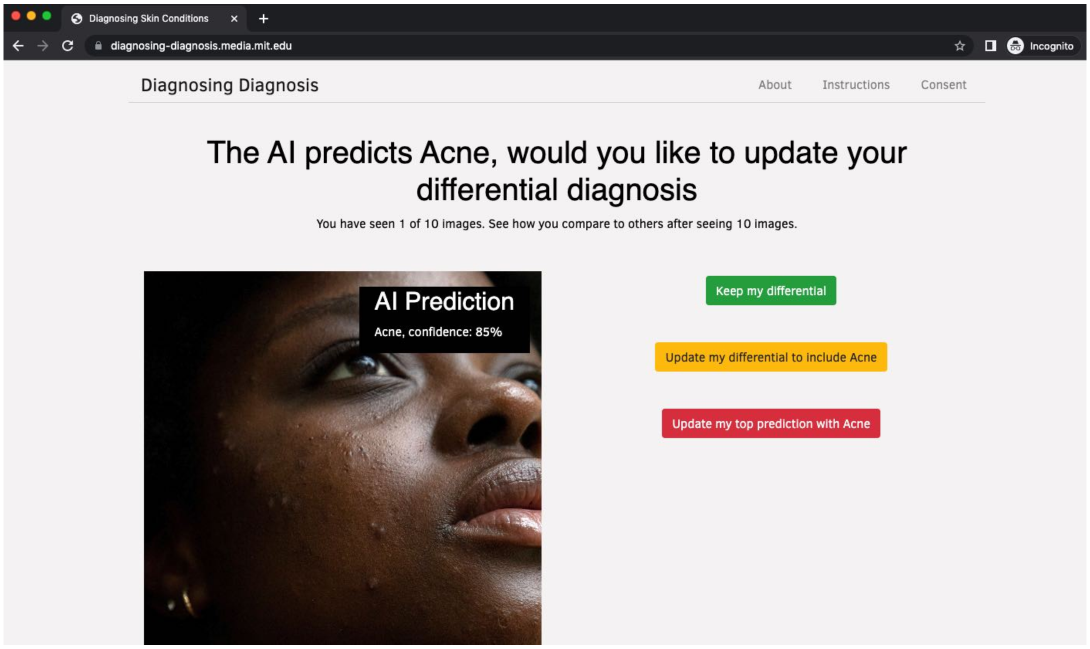


Diagnosing Skin Conditions
diagnosing-diagnosis.media.mit.edu
Incognito
Diagnosing Diagnosis
About
Instructions
Consent
ac
Type secondary differential diagnosis
Type tertiary differential diagnosis
acanthoma fissuratum
acanthosis nigricans
accessory nipple
acne
acne keloidalis
acquired digital fibrokeratoma
acral keratosis
acral peeling skin syndrome
acrocyanosis
acrodermatitis atrophicans chronica
acropustulosis of infancy
actinic keratosis
actinic granuloma
Slide to
agnoses
□ B
ion

Can you accurately diagnose this skin condition?

You have seen 1 of 10 images. See how you compare to others after seeing 10 images.

Figure 6-10: Screenshot from the Diagnosing Diagnosis experiment website showing predictive text for selecting diagnoses


Diagnosing Skin Conditions
diagnosing-diagnosis.media.mit.edu
Incognito
Diagnosing Diagnosis
About
Instructions
Consent
Keep my differential
Update my differential to include Acne
Update my top prediction with Acne

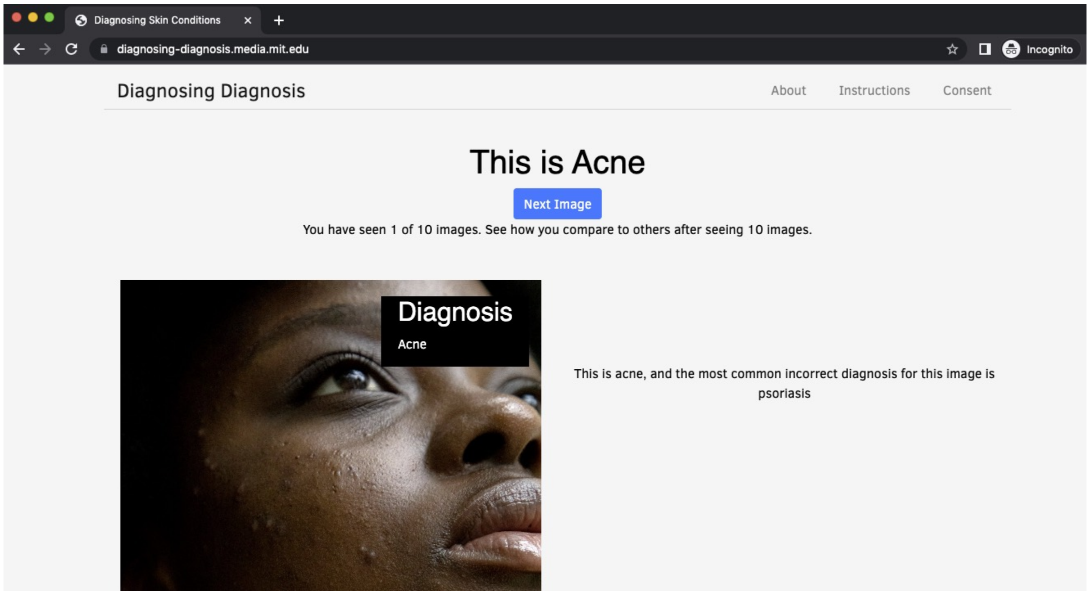
The AI predicts Acne, would you like to update your differential diagnosis

You have seen 1 of 10 images. See how you compare to others after seeing 10 images.

AI Prediction

Acne, confidence: 85%

Figure 6-11: Screenshot from the Diagnosing Diagnosis experiment website showing the DLS suggestion. Participants are randomly assigned to either see the three options in the order presented or the reverse order with "Update my top prediction..." on top, "Update my differential" in the middle, and "Keep my differential" on the bottom.



This is Acne

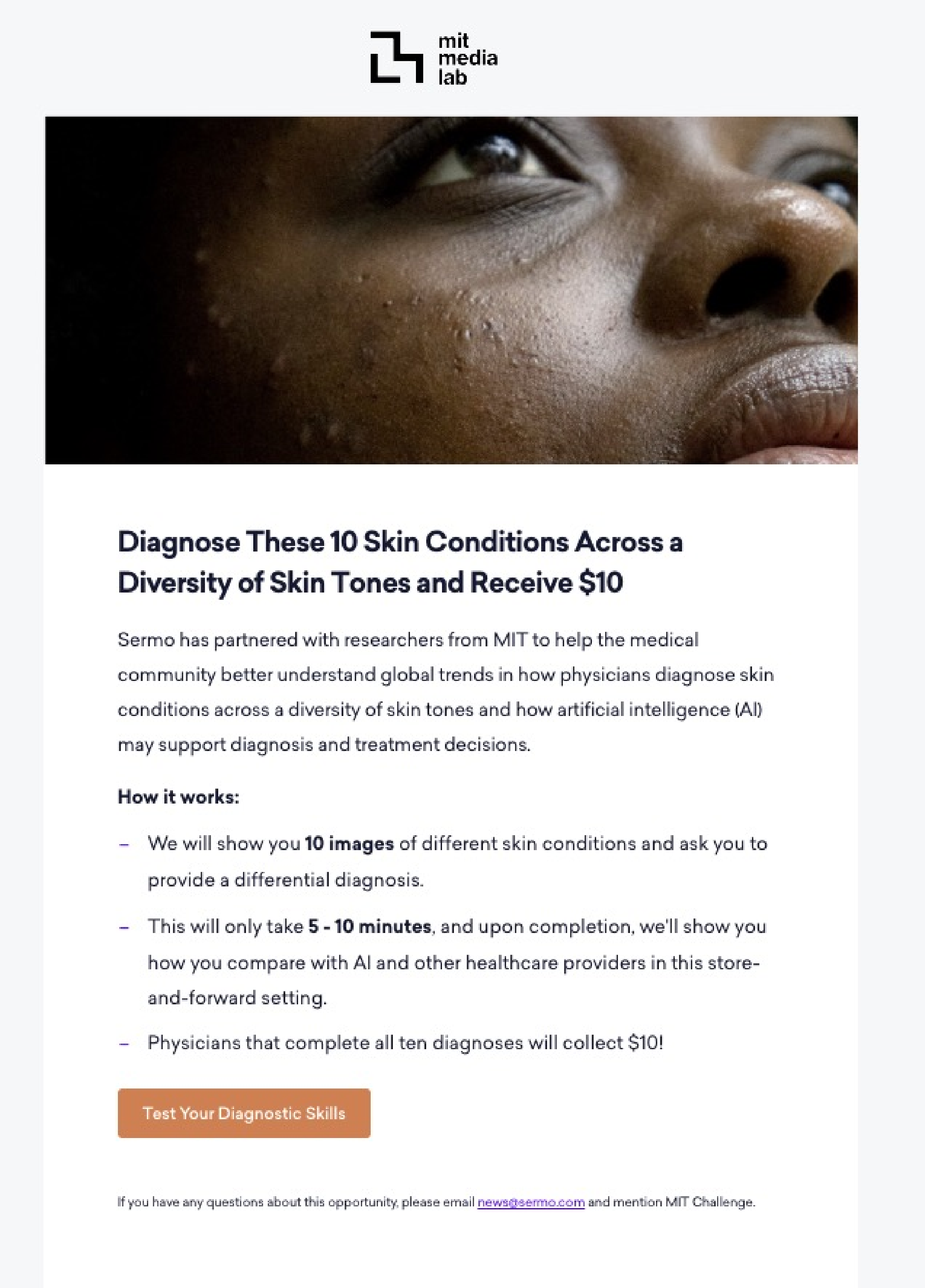
You have seen 1 of 10 images. See how you compare to others after seeing 10 images.

Diagnosis

Acne

This is acne, and the most common incorrect diagnosis for this image is psoriasis

Figure 6-12: Screenshot from the Diagnosing Diagnosis experiment website showing the feedback based on the original label.



Diagnose These 10 Skin Conditions Across a Diversity of Skin Tones and Receive $10

Sermo has partnered with researchers from MIT to help the medical community better understand global trends in how physicians diagnose skin conditions across a diversity of skin tones and how artificial intelligence (AI) may support diagnosis and treatment decisions.

How it works:

- We will show you **10 images** of different skin conditions and ask you to provide a differential diagnosis.
- This will only take **5 - 10 minutes**, and upon completion, we'll show you how you compare with AI and other healthcare providers in this store-and-forward setting.
- Physicians that complete all ten diagnoses will collect $10!

If you have any questions about this opportunity, please email news@sermo.com and mention MIT Challenge.

Figure 6-13: Screenshot from the email to healthcare providers on Sermo's platform.

Chapter 7

Computational Empathy Counteracts the Effects of Anger on Human Creative Problem Solving

Abstract

How does empathy influence creative problem solving? We introduce a computational empathy intervention based on context-specific affective mimicry and perspective taking by a virtual agent appearing in the form of a well-dressed polar bear. In an online experiment with 1,006 participants randomly assigned to an emotion elicitation intervention (with a control elicitation condition and anger elicitation condition) and a computational empathy intervention (with a control virtual agent and an empathic virtual agent), we examine how anger and empathy influence participants' performance in solving a word game based on Wordle. We find participants who are assigned to the anger elicitation condition perform significantly worse on multiple performance metrics than participants assigned to the control condition. However, we find the empathic virtual agent counteracts the drop in performance induced by the anger condition such that participants assigned to both the empathic virtual agent and the anger condition perform no differently than participants in the control elicitation condition and significantly better than participants assigned to the control virtual agent and the anger elicitation condition. While empathy reduces the negative effects of anger, we do not find evidence that the empathic virtual agent influences performance of participants who are assigned to the control elicitation condition. By introducing a framework for computational empathy interventions and conducting a two-by-two factorial design randomized experiment, we provide rigorous, empirical evidence that computational empathy can counteract the negative effects of anger on creative problem solving.[1]

[1] This chapter, which is co-authored by Craig Ferguson, Robert Lewis, and Rosalind Picard appeared in the proceedings for the Affective Computing and Intelligent Interactions (ACII) 2022 [245].

7.1 Motivation

Empathic virtual agents are artificial intelligence systems designed to perceive and express affect in order to simulate the appearance of empathy in interactions with humans. Computational empathy involves recognizing an individual's emotional state and responding appropriately via affective mimicry and perspective taking [479]. While affective computing [500] seeks to address the challenges of recognizing emotions and responding empathically, these are not solved problems and there remain many open questions on how to evaluate computational empathy [654]. In evaluating empathy in humans, psychologically validated methods like the Interpersonal Reactivity Index [147], Empathy Quotient [50, 362], and Toronto Empathy Questionnaire [581] involve measuring the self-reported empathy traits and preferences of an individual, but these first-person scales are not relevant for evaluating how individuals perceive empathy expressed by others. In order to evaluate perceived empathy, recent evaluations have transformed previously validated methods into evaluations by outside observers, which can be either an interaction partner or a third party [270, 655]. However, these evaluations by outside observers can be affected by a range of factors including observer-level factors (sociocultural background and experience with computers), context-level factors (the role of the agent as a companion or trainer and the quality of experience from a perspective of effectiveness, efficiency, utility, and acceptability), and agent-level factors (likeability, anthropomorphism, animacy, perceived intelligence and safety) [654].

Rather than evaluating how empathic a virtual agent appears, we focus this paper on two contributions: (1) the introduction of a constrained context and a virtual agent designed to respond to humans with contextually appropriate affective mimicry and perspective taking and (2) the evaluation of the effectiveness of such an empathic virtual agent in enhancing a human's cognitive performance [398]. Motivated by the question, "How does computational empathy influence creative problem solving?," we evaluate how an empathic virtual agent, which is integrated into an online word guessing game, Affective Wordle Lab (based closely on Wordle [60]), influences cognitive performance.

Many research experiments have shown that emotions influence creative problem solving and decision-making. For example, past research has experimentally demonstrated that positive affect (as elicited in the late 1980s by either a short blooper reel or a small gift of candy) facilitates creative problem solving [295]. This research operationalized creative problem solving based on two tasks, Duncker's Candle task [175] and the Remote Associates Test (RAT) [419], which involve finding the "relatedness in diverse stimuli that normally seem unrelated" [295] and require "breaking set" – e.g., recognizing the box of tacks in Duncker's Candle task as a box and tacks and recognizing relatedness of words in the RAT based on many different kinds of associations. In contrast to the effects of positive affect on creative problem solving and decision-making [296], past research on experimentally elicited anger shows anger inhibits decision-making by reducing depth of processing and increasing reliance on heuristic processing [374–377].

7.2 Related Work

7.2.1 Video Games, Emotions, and Cognitive Testing

Video games are an area of growing interest in affective computing and other computational sciences, where much of the work benefits from two key properties of video games. First, video games can deeply engage players and evoke poignant emotional experiences, thus enabling the study of various psychological constructs on large study or real-world user populations [15, 199, 544]. Second, video games can be customised to create highly controlled environments that probe specific cognitive and affective processes, thus giving researchers fine-grained control of their experiments and the ability to objectively measure constructs of interest by analyzing the game telemetry data of player actions and decisions [417, 658]. For example, numerous mini-games have been developed to assess cognitive processes such as working memory, motivation, appraisal of and aversion to risk or reward, creativity, and other general executive functions. These tests aid our understanding of the mechanisms of human decision-making and problem solving, and have found utility in various contexts including mental health monitoring where impairments in decision-making processes may be indicative of disorders like anxiety or anhedonia [228, 394].

7.2.2 Empathic Virtual Agents

In the past twenty years, many experiments have empirically demonstrated the power of empathic virtual agents to influence human affect including undoing negative feelings of frustration [334], increasing people's feelings of being cared for [67, 82], altering people's feelings of fear into neutral feelings [440], and reducing public speaking anxiety [449]. While the fundamental tenets of affective mimicry and perspective taking drive computational empathy, the design space for computational empathy is combinatorically large. One recent study examined the systematic manipulation of animation quality, speech quality, and rendering style and their impacts on people's perceptions of virtual agents in terms of naturalness, engagement, trust, credibility, and persuasion in a health counseling domain [485]. In virtual agent chatbots that are limited to text interfaces, the range of possible conversations remains very large yet examples of personalized machine-learning based chatbots have been shown to interact empathically with humans and be perceived as likable [204, 225, 226]. While past experiments have examined human perception of computational empathy, the authors are not aware of past experiments examining the impact of computational empathy on cognitive performance metrics.

7.3 Methods

7.3.1 Participants (N=1,006)

We recruited 1,006 participants from Prolific, an online platform for recruiting research participants. We restricted recruitment to individuals on Prolific who live in the United States and speak English as a first language. As a robustness check to the inclusion criteria, we ask participants "Are you a native English speaker?", and 99.7% of participants respond "Yes." Participants' ages range from 18 to 84 with a median of 35, and 53% of participants identify as female. Before participants played Wordle in this experiment, we asked "How many times have you played Wordle" to which 31% of participants respond they have never played Wordle, 4% just once, 19% 2 to 10 times, 42% 11 to 100 times, and 3% have played Wordle over 100 times.

7.3.2 Experimental Design

We pre-registered the experiment on *AsPredicted* at https://aspredicted.org/yx4k3.pdf.

Participants are randomly assigned to two interventions: an emotion elicitation intervention with two conditions (control and anger) and a virtual agent personality intervention with two conditions (control personality and empathic personality). In this 2x2 factorial design, we assign participants to control and treatment conditions with equal likelihoods; 26% of participants were assigned to the control-control group, 26% of participants were assigned to the anger-control group, 24% of participants were assigned to the control-empathy group, and 24% of participants were assigned to the anger-empathy group.

Emotion Elicitation

We based the emotion elicitation intervention on a reflective writing exercise from Small and Lerner (2008) [576]. In both conditions, we ask participants to respond to two similarly structured questions with a minimum response of 150 characters each. The goal of these questions is to generate equivalent cognitive loads while activating incidental anger in one condition and not activating any specific emotion in the other condition. An incidental emotion refers to an emotion unrelated to the main task, which contrasts with an integral emotion, which refers to an emotion intrinsically tied to the main task. In this experiment, we focus on incidental anger.

For participants assigned to the control elicitation condition, we first ask: "What are three to five activities that you did today? Please write two-three sentences about each activity that you decide to share. (Examples of things you might write about include: walking, eating lunch, brushing your teeth, etc.)" After they answer, we follow up with a second question: "Now, we'd like you to describe in more detail the way you typically spend your evenings.

Begin by writing down a description of your activities and then figure out how much time you devote to each activity. Examples of things you might describe include eating dinner, studying for an exam, working, talking to friends, watching TV, etc. If you can, please write your description so that someone reading this might be able to reconstruct the way in which you, specifically, spend your evenings."

For participants assigned to the anger elicitation condition, we first ask: "What are the three to five things that fill you with anger? Please write two-three sentences about each thing that fills you with anger. (Examples of things you might write about include: being treated unfairly by someone, being insulted or offended, etc.)" After they answer, we follow up with a second question: "Now, we'd like you to describe in more detail the one situation that makes you (or has made you) experience the most anger. This could be something you are presently experiencing or something from the past. Begin by writing down what you remember of the anger-inducing event(s) and continue by writing as detailed a description of the event(s) as is possible. If you can, please write your description so that someone reading this might even feel anger just from learning about the situation. What is it like to be in this situation? Why does it make you so feel such anger?"

Affective Wordle Lab

After responding to the reflective writing exercise, we provide instructions to participants for playing Wordle and invite participants to play four rounds of Wordle. The rules of the Affective Wordle Lab experiment are the same as the rules in the official version of Wordle hosted by the New York Times [60]. The goal of each round is to guess a 5-letter-word within 6 guesses. After each guess, players are informed whether each letter in their guess is: (a) in the correct position for the solution word, (b) in the solution word but not the correct position, or (c) not in the solution word. Players can use this information to home in on the 5-letter-word solution. Wordle closely resembles the word game Jotto, which attracted the interest of computer programmers in the early 1970s for studying information theory aspects of the game [57].

We adapted Wordle's game mechanics such that participants play 4 rounds and have the option to play additional bonus rounds, which stands in contrast to the official version of Wordle's standard limit of a single round per day. The list of 12,972 acceptable guesses in the Affective Wordle Lab is identical to the list of acceptable guesses in the official version of Wordle and all solutions are chosen from the official list of 2,315 acceptable solutions (the solutions are a subset of the acceptable guesses, which are based on common word use). 12,972 acceptable guesses corresponds to a possibility space of about 10^{24} (precisely bounded above by $12,972 * 12,971 * 12,970 * 12,969 * 12,968 * 12,967$) combinations of guesses for arriving at the correct 5-letter-word.

We selected four neutral words as solutions to the four rounds in the following order: "plant," "fuzzy," "diner," and "image." We expected "plant" and "diner" to be relatively easy, "image" to be moderately difficult because it begins with a vowel, and "fuzzy" to be difficult because it contains uncommon double letters.

We hosted the Affective Wordle Lab version of Wordle at https://wordlelab.media.mit.edu/ based on an open-source clone of Wordle that we extended with a Django backend and a MySQL database.

Virtual Agent Personalities

In a deviation from the popular Wordle game, the Affective Wordle Lab includes a virtual agent, a dynamic cartoon polar bear wearing a red scarf based on the Animated Login Screen by JcToon on Rive. We designed the virtual agent to either display a "control" or "empathic" personality.

Just a doodle

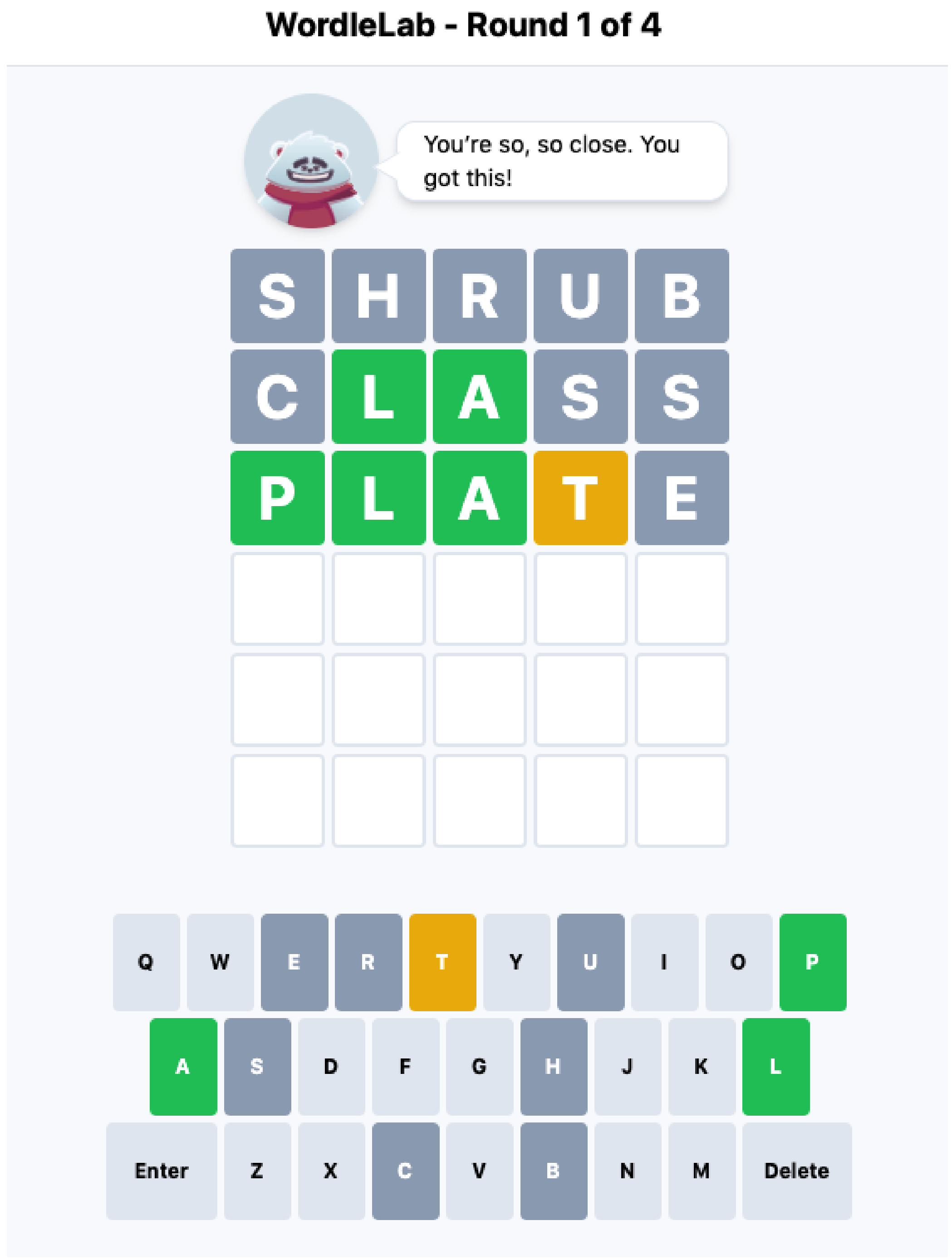
In the control condition, the virtual agent is always idle and is programmed to only communicate obvious game status information. Specifically, the virtual agent's speech bubble is limited to "Guess [1-6] of 6" for each guess iteration or "You [won/lost] after [1-6] guesses" between each round.

In the empathy condition, the virtual agent makes expressions based on game-specific contexts to empathize with the participant via affective mimicry and perspective taking. In particular, we programmed the virtual agent to take into account how many possible words remain, how many guesses the participant has made, how quickly a participant responds, how many letters a participant has uncovered, whether a guess is valid, whether a participant wins or loses, and whether a participant is idling. In Figure 7-2, we present screenshots of the virtual agent's six expressions, and in Table 7.1, we detail the 6 expressions and 39 messages paired with 13 game-specific contexts. We draw on the management science of nonverbal behavior [105] to pair these expressions, messages, and contexts.

Each context is associated with 1 or 2 expressions paired with 1, 4, or 6 messages. The virtual agent never repeats the same message in the same round and always selects a new message for each round for contexts with 4 messages. The virtual agent selects a message based on the order of contexts presented in Table 7.1; for example, "Fewer than 6 words remaining" trumps "Fast Guess (under 4 seconds)," which trumps "5th guess" and so forth.

Post-Game Questionnaire

After participants complete 4 rounds of Wordle, we ask participants two additional sets of follow-up questions. First, we ask participants "How are you feeling right now?" and the experiment interface provides affective sliders [65] for participants to report their valence and arousal. Second, we ask participants to answer 3 questions from the Cognitive Reflection Test (CRT) designed to measure depth of reflective reasoning [214]. After participants respond to the questions from the Cognitive Reflection Test, we congratulate participants for finishing the experiment and provide a link back to the Prolific website where participants can collect their payment. After collecting payment, participants have the option to continue playing more rounds of Wordle.



WordleLab - Round 1 of 4

You’re so, so close. You got this!

Figure 7-1: Screenshot of the Affective Wordle Lab experiment with an empathic virtual agent.

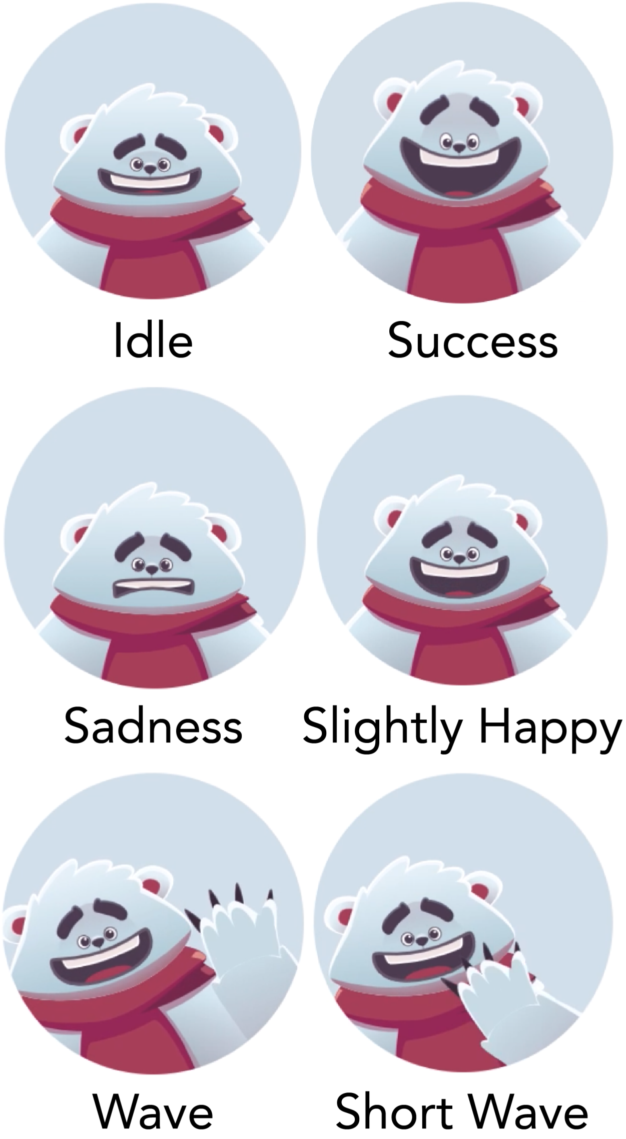


Figure 7-2: Screenshots of the virtual agent in the six dynamic displays designed for affective mimicry. The idle display shows the bear moving his eye brows up and down every few seconds; success shows the bear shrug and burst into a wide smile with raised eyebrows; sadness shows the bear shrug, raise head, and shrug again while frowning and wiggling its ears; slightly happy shows the bear shrug and raise its head into a smile with raised eyebrows; the wave shows the bear waving buoyantly three times with raised eyebrows; the short wave shows the bear waving two times with raised eyebrows.

Context	Expression	Message	Message
Fewer than 6 words remaining	Wave Short	You're so, so close. You got this!	
Fast Guess (under 4 seconds)	Wave Short	Wow, you're so fast! Incredible!	
Slow Guess (over 60 seconds)	Wave Short	Taking your time really paid off!	
1st guess	Wave Short	Good luck! You got this!	Another round! You can do this!
1st guess	Wave Short	You've got the hang of this!	I know you can get this one!
5th guess	Idle	Two guesses left, that's plenty of time!	Last two guesses! Trust yourself, you got this.
5th guess	Idle	This is a tough one, but you're close!	This one can be hard, but I believe in you!
6th guess	Wave	Just breathe and think it through. You got this!	Stay calm and use all the facts you uncovered.
6th guess	Wave	You final chance. You can do it!	Don't give up now! Stay calm and breathe.
Fewer than 101 words remaining	Wave Short	You're getting closer!	Oh nice, that really narrowed the field!
Fewer than 101 words remaining	Wave Short	Ooh, you're getting close now!	That was a really good guess!
Additional letters revealed	Success	Wow! What a great guess!	Ooh nice one! I didn't think of that.
Additional letters revealed	Success	You learned more information! Nice work!	Great guess!
No additional letters revealed	Slightly Happy	Okay! Well now we know what doesn't work.	Nice! Now we know what to avoid
No additional letters revealed	Sadness	Aww, I was sure that would be it.	Hmm, what could it be?!
Invalid	Sadness	Oops! I don't know that word! Give it another try.	
Win	Win	This must be your lucky day	Two guesses?! Are you a wizard?!
Win	Win	Three guesses? You're a rock star!	Great job! You won in four guesses!
Win	Success	You did it! You won in five guesses!	That was close, but you did it!
Loss	Sadness	You almost had it! Let's try again.	
Idle (triggered at 90 seconds)	Wave Short	It's good to think it through carefully.	I believe in you!
Idle (triggered at 90 seconds)	Wave Short	It's okay to feel stumped. You'll get it!	
Idle (triggered at 90 seconds)	Sadness	This one is a toughy, isn't it?	

Table 7.1: The 6 expressions and 39 messages associated with 13 game-specific contexts. Each context is associated with 1 or 2 expressions and 1, 4 or 6 messages.

7.3.3 Dependent Variables

We examine four measures of game performance: (1) a binary variable for winning for each round, (2) the number of guesses per round, (3) an adjusted number of guesses per round where participants who lost are assigned 7 instead of 6 guesses, and (4) entropy reduction at the guess level, which is computed as the number of bits remaining: $log_2(w)$ where w is the mean number of words remaining for all possible solutions after each guess. We consider entropy reduction based on both a reduction of the 2,315 possible 5-letter solutions and the 12,972 possible 5-letter guesses in the official Wordle game. These are two approaches to evaluating entropy reduction of guesses, but other reasonable approaches could alternatively consider the five-letter-words from Scrabble, the Oxford English dictionary, or another source. Likewise, other reasonable approaches to evaluating entropy reduction could also take into account the word frequency. We limit our analysis to the reduction of the possible 5-letter solutions and guesses in the official Wordle game and leave additional analyses for future work.

In addition to game performance, we examine self-reported valence and arousal from the post-game questionnaire, whether participants engage in additional game play immediately after finishing the experiment, the sentiment of participants' guesses based on the VADER rule-based model [287], the time between guesses, the word frequency of participants' guesses, and the number of invalid attempts submitted (e.g. a guess of "QQQQQ" is an example of an invalid attempt).

7.4 Results

7.4.1 Round Level Performance

We evaluate treatment effects of the anger elicitation intervention and the empathic virtual agent personality by running the following pre-registered ordinary least squares (OLS) regression where $Y_{i,t}$ is a dependent variable (specified above) for individual, i, in round t, A is a binary variable for assignment to the anger elicitation condition, E is a binary variable for assignment to the computational empathy intervention, α and β_{1-3} are the regression intercept and coefficients, respectively, and ϵ is the error term clustered at the individual level [2]:

$$Y_{i,t} = \alpha + \beta_1 A_{i,t} + \beta_2 E_{i,t} + \beta_3 A_{i,t} E_{i,t} + \epsilon_i \tag{7.1}$$

We find the effects of both the anger condition and the interaction between the anger and empathy condition are statistically significant at the $p < 0.05$ significance level. We report these results in Table 7.2. Relative to the control group, participants assigned to the anger condition won 7 percentage points less frequently ($p = 0.021$), made an additional 0.15 guesses ($p = 0.017$), and made an additional 0.21 adjusted guesses ($p = 0.012$). Relative to participants assigned to the anger condition but not the empathy condition, participants assigned to both the anger and empathy condition won 8 percentage points more often ($p = 0.040$), made 0.21 fewer guesses ($p = 0.018$), and made 0.30 fewer adjusted guesses ($p = 0.016$). We do not find assignment to the empathy intervention increases performance relative to participants assigned to the control emotion elicitation condition; participants assigned to the empathy condition won 2 percentage points less frequently ($p = 0.45$), made an additional 0.11 guesses ($p = 0.10$), and made 0.13 additional adjusted guesses ($p = 0.143$). These results remain the same with the inclusion of round fixed effects to the linear model.

With 1,006 participants and 4 rounds of Wordle, we should have 4,024 observations in the regression analysis, but instead we have 3,975 observations. We are missing 1.2% of observations due to interruptions in some participant's internet connections that allowed 24 participants (2.4% of participants) to continue the experiment without all their responses logged to the experiment's server.

7.4.2 Heterogeneity of Treatment Effects on Performance

We examine heterogeneity of treatment effects on round-level performance by including experience playing Wordle at least once before, depth of reflective reasoning as proxied by the CRT measured from 0-3, and self-reported sex [598] in the OLS regressions.

Formally, Equation 2 includes the same terms as Equation 1 but also includes H_i, which is the heterogeneous feature of interest (either a binary variable for playing Wordle at least

	Did Win	Guesses	Guesses (Adjusted)
Constant	0.72***	4.82***	5.10***
	(0.02)	(0.04)	(0.06)
Anger	-0.07*	0.15*	0.21*
	(0.03)	(0.06)	(0.08)
Empathy	-0.02	0.11	0.13
	(0.03)	(0.06)	(0.09)
Anger * Empathy	0.08*	-0.21*	-0.30*
	(0.04)	(0.09)	(0.12)
Observations	3,975	3,975	3,975
Number of Participants	1006	1006	1006

*p<0.05; **p<0.01; ***p<0.001

Table 7.2: Ordinary least squares (OLS) regressions with robust standard errors clustered at the participant level.

once before, a continuous variable from 0 to 3 indicating performance on the CRT, or a binary variable for Female). β_4 represents the direct association of the heterogeneous feature with the dependent variable, while β_{5-7} are the associations of its interactions (i.e., the heterogeneous effects):

$$\begin{aligned} Y_{i,t} = \alpha + \beta_1 A_{i,t} + \beta_2 E_{i,t} + \beta_3 A_{i,t} E_{i,t} + \\ \beta_4 H_i + \beta_5 H_i A_{i,t} + \beta_6 H_i E_{i,t} + \\ \beta_7 H_i A_{i,t} E_{i,t} + \epsilon_i \end{aligned} \tag{7.2}$$

We find that participants who had played Wordle at least once before this experiment won 20% more frequently ($p < 0.001$), took 0.35 fewer guesses ($p < 0.001$), and took 0.55 fewer adjusted guesses ($p < 0.001$). Likewise, we find that for every CRT question participants answered correctly, they won 9% more frequently ($p < 0.001$), took 0.20 fewer guesses ($p < 0.001$), and took 0.29 fewer adjusted guesses ($p < 0.001$). We do not find significant difference between men and women's performance. Moreover, we do not find statistically significant effects of interactions (i.e., β_{5-7}) between either experience playing Wordle, CRT performance, or sex and the experimental conditions on round-level performance.

7.4.3 Guess Level Entropy Reduction

We evaluate treatment effects on entropy reduction at the guess level as an additional performance metric. Specifically, we run OLS regressions following Equation 1 with an additional index g on each term to denote the guess index. We measure entropy reduction by computing the $log_2(w)$ where w is the mean number of possible words (out of either the 2,315 solutions or 12,972 valid words) remaining after each guess.

In Figure 7-3, we present the 95% confidence intervals for treatment effects on the mean marginal bits remaining from the 2,315 solutions for each guess iteration. We find that participants assigned to the anger elicitation condition have 0.11 to 0.16 additional bits of information remaining in their first four guesses ($p = 0.01$, $p = 0.06$, $p = 0.03$, $p = 0.03$) compared to the participants assigned to the control elicitation condition. In contrast, the Anger * Empathy interaction term ranges from -0.22 to -0.10 for the first four guesses ($p = 0.09$, $p = 0.06$, $p = 0.04$, $p = 0.10$). By the fifth and sixth guesses many participants have already identified the word and the average remaining bits is 0.6 and 0.47, respectively, so the lack of statistical differences across the anger elicitation and anger elicitation paired with computational empathy interventions in the fifth and sixth guesses can be explained by differential dropout and floor effects.

As a robustness check, we also examine the treatment effects on the mean marginal bits remaining from the 12,972 valid words. We find that participants assigned to the anger elicitation condition have 0.13 to 0.17 additional bits of information remaining (based on the 12,972 valid words) in their first four guesses ($p = 0.003$, $p = 0.13$, $p = 0.10$, $p = 0.16$) compared to the participants assigned to the control elicitation condition. In contrast, the Anger * Empathy interaction term ranges from -0.25 to -0.10 for the first four guesses ($p = 0.08$, $p = 0.08$, $p = 0.18$, $p = 0.24$).

7.4.4 Self-Reported Affect and Additional Outcomes

Based on OLS regressions of treatment effects following Equation 1 on self-reported valence and arousal that participants provided after completing the four rounds of Wordle, we find the anger elicitation has a statistically significant, negative effect on self-report affect. On a scale from 0 to 100, participants assigned to the anger condition report a 5.1 point lower arousal ($p = 0.030$) and a 4.6 point lower valence ($p = 0.064$) relative to participants assigned to the control emotion elicitation condition. We do not find statistically significant effects on self-reported valence and arousal from assignment to the computational empathy condition or the interaction between anger and computational empathy.

Based on OLS regressions of treatment effects on additional outcomes measured at the guess-level following Equation 1 including whether participants engaged in additional rounds of Wordle, guess sentiment based on the VADER rule-based model, the response time between guesses, the word frequency of participants' guesses, and the number of invalid attempts, we do not find statistically significant treatment effects of anger, empathy, or the interaction of anger and empathy. While we did not find significant treatment effects on these additional outcomes, we do see variation across participants across these features. 17% of participants participated in at least one bonus round. Participants' guesses were classified by VADER as neutral for 86% of words, positive for 7% of words, and negative for 7% of words. The mean time between each guess was 35 seconds. Participants submitted 6,074 5-letter strings, 3,175

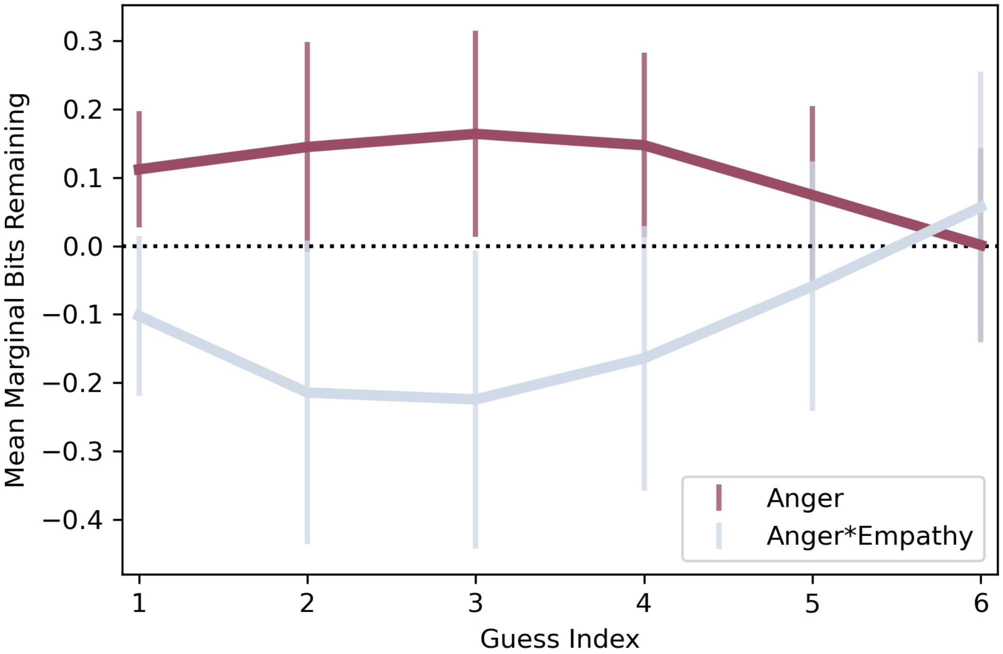


Figure 7-3: Mean marginal bits remaining for all 2,315 possible Wordle solutions after each guess based on OLS regressions with robust standard errors clustered at the participant level. Bits are computed as $log_2(w)$ where w is the mean number of words remaining out of the 2,315 solutions. Error bars represent 95% confidence intervals.

unique valid words, and 1,609 unique words that are possible Wordle solutions, and 1,045 different first guesses of which 727 were valid Wordle solutions. Finally, 77% of participants submitted at least 1 invalid guess.

7.5 Discussion

How do incidental anger and computational empathy influence creative problem solving in a word guessing game?

The results from our pre-registered experiment corroborate past research finding anger impairs decision-making and reduces depth of cognitive processing [374–377]. In particular, we find the anger elicitation condition (relative to the control elicitation condition) leads participants to lose more often, make more guesses (and adjusted guesses), and submit less informative guesses (as measured by entropy reduction); these results are statistically significant.

In contrast, we find the computational empathy intervention counteracts the negative effect of anger on performance. Participants assigned to both the anger elicitation condition and the empathic virtual agent perform better than participants assigned to the anger elicitation condition and the control virtual agent on all performance metrics, which are statistically significant at the $p < 0.05$ level for round-level performance metrics and statistically significant at the $p < 0.10$ level for guess-level entropy reduction. While computational empathy counteracts the negative effects of anger on cognitive performance, we do not find computational empathy changes performance for participants assigned to the control elicitation condition.

We find experience playing Wordle and depth of reflective reasoning as proxied by the CRT is strongly associated with performance, but we do not find significant heterogeneous treatment effects based on either participants' experience playing Wordle or participants' depth of reflective reasoning. The lack of heterogeneous treatment effects on both these characteristics and also participants' sex suggest that none of these characteristics make participants more or less vulnerable to the negative effects of anger or to the counterbalancing effects of computational empathy.

While anger and empathy influence overall performance, we do not find treatment effects on other outcomes like sentiment of guesses, response time between guesses, word frequency of guesses, number of invalid guesses, or whether participants engaged in additional rounds of Wordle after collecting payment for participating. The lack of treatment effects on these outcomes narrows possible mechanisms by which anger and empathy influence cognitive performance.

After participants complete four rounds in the Affective Wordle Lab, we present participants an opportunity to self-report their valence and arousal with an affective slider [65]. We find effects of the anger elicitation condition but not the empathic virtual agent personality on both arousal and valence. As expected, we find lower valence and arousal in participants assigned to the anger elicitation condition relative to the control condition. This may be surprising because anger is associated with positive arousal in the circumplex model of affect [541], but recent research that clusters semantic emotion categories and affective dimensions reveal different variations of anger that include both high and low arousal [130]. The persistent effects of anger on self-reported affect – that last through multiple rounds of the Affective Wordle Lab and are not counteracted by the computational empathy intervention – reveal the effectiveness of the reflective writing exercise from Small and Lerner (2008) [576] for eliciting incidental anger.

7.6 Limitations

We evaluate treatment effects of an anger elicitation intervention and computational empathy intervention in a 2x2 factorial design on performance in an online word guessing game that involves creative problem solving. Similar to Duncker's Candle task and the RAT, this game represents only one kind of creative problem solving and it does not represent all creative problem solving. Moreover, we focused this experiment on incidental anger because it is more straightforward to elicit in experimental settings than integral anger. Future work may consider how integral emotions influence both creative problem solving and the ability of computational empathy to counteract the negative influence of anger.

In this experiment, we avoid precisely defining computational empathy and treat computational empathy as a gestalt of contextualized interactions involving affective mimicry and perspective taking. This gestalt treatment allows us to operationalize computational empathy as an intervention, but it prevents us from identifying the precise features that help counteract the negative effects of the anger elicitation intervention. Future research may consider the effectiveness of the empathic virtual agent without affective mimicry or without perspective taking or without some contexts to identify the most effective component parts and combination of component parts of computational empathy for improving an angry individual's performance. Likewise, future research may consider how computational empathy influences emotion regulation.

7.7 Contributions and Implications

We present a conceptual replication of experiments on anger and decision-making, and we corroborate previous findings that anger inhibits problem solving. Moreover, we present experimental evidence that computational empathy can counteract the negative effects of anger on creative problem solving. The countervailing force of computational empathy on anger highlights the importance of designing empathy into virtual agents to not only make people feel cared for but to boost people's creative performance.

Affective Wordle Lab presents a new tool and paradigm for interweaving a virtual agent within the constrained context of a game such that researchers can experimentally elicit emotions and manipulate virtual agents to evaluate computational empathy not only based on self-reports of how people feel but also as an assistive technology that can influence human decision-making and creative problem solving.

7.8 Ethics and Informed Consent

This research complies with all relevant ethical regulations. The Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects determined this study to fall under Exempt Category 3: Benign Behavioral Intervention and Exempt Category 2: Educational Testing, Surveys, Interviews or Observation with id E-3888.

All participants are informed that "WordleLab is a research project created by the MIT Media Lab" and "All submissions are collected anonymously for research purposes, and participation is entirely voluntary. For questions, please contact wordlelab@media.mit.edu." All participants are recruited from the Prolific survey platform with the following message: "The Wordle Lab Experiment is a research project created at the MIT Media Lab to study how people play Wordle. First, we'll ask you to share three to five examples of something relevant to you, then you'll play 4 rounds of Wordle, and last we'll ask you a few follow up questions. All data is collected anonymously. We estimate this experiment to take about 15 minutes." We compensated participants with $2.38 each, which is a rate of $9.52 an hour.

7.9 Data and Code Availability

We open-sourced the code for Affective Wordle Lab at `https://github.com/MITMediaLabAffectiveComputing/WordleLab` and share anonymized participant data and replication code at `https://github.com/mattgroh/affective_wordle_lab_replication`.

Acknowledgments

We acknowledge JcToon on Rive for creating and licensing the Animated Login Screen. We thank Neska El Haouij and Ila Kumar for feedback on an early version of the experiment and Boyu Zhang for feedback on an early draft.

Chapter 8

Context in Automated Affect Recognition

Abstract

Affect recognition depends on interpreting both expressions and their associated context. While expressions can be explicitly measured with sensor technologies, the role of context is more difficult to measure because context is often left undefined. In an effort to explicitly incorporate pragmatics in automated affect recognition, we develop a framework for categorizing context. Building upon ontologies in affective science and symbolic artificial intelligence, we highlight seven key categories: ambient sensory environment, methods of measurement, semantic representation, situational constraints, temporal dynamics, sociocultural dimensions, and personalization. In this chapter, we focus on how the epistemological categories of context influence the training and evaluation of machine learning models for affect recognition. Incorporating context in the practical and theoretical development of affect recognition models is an important step to developing more precise and accurate models.[1]

8.1 Motivation

In an early 20th century film experiment, cinematographer Lev Kuleshov presented audiences with a short clip of an actor expressing a neutral facial expression followed by one of three scenes: a bowl of soup, a young girl in a coffin, and a woman lying on a couch. Depending on which scene the audience saw, the audience described the actor's expression as indicative of different emotions; hunger for the soup, sadness for the deceased, and lust

[1] This chapter, which is co-authored with Rosalind Picard, is a draft that was presented as an abstract at the Meaning in Context workshop at the Conference for Neural Information Processing Systems in December 2021 and the Affective Cognition Workshop at Cognitive Science Society in July 2021.

for the woman. Two recent experiments replicated the results of the original Kuleshov experiment and extended it to show that scenes conveying fear and desire also lead audiences to report neutral facial expressions as expressions matching the sentiment in the juxtaposed scenes [51, 94].

Context shapes how humans perceive and recognize emotions. For example, the art of transforming a script into a heart-wrenching movie involves not only actors' dialogue and physicality (their observable expressions) but also how these expressions relate to scene transitions, the musical accompaniment, lighting conditions, costume and set design, and narrative devices. Likewise, how an observer interprets another person's smile depends on contextual cues like whether a person is acting earnestly, whether a person is in a pain-eliciting situation, or whether social display rules might influence a person to mask their inner feelings.

In affective computing, emotion recognition has been described as a combination of "observations of emotional expressions" and "reasoning about an emotion-generating situation" [500]. This dual focus of emotion recognition on expressions and context matches research in affective science, which shows observable expressions are often ambiguous without context [6, 43, 52, 55, 149, 178, 200, 201, 223, 275, 409, 639, 642]. Emotion recognition is a subset of affect recognition, which is sometimes referred to as affect detection, affect estimation, and affect measurement in the field of affective computing. Emotion recognition has also been called empathic accuracy in the field of affective science and emotion reasoning in the field of developmental psychology [288, 538]. Automated affect recognition applies methods from signal processing and machine learning to situated expression data, which are data on observable expressions and their associated context [139]. While facial expressions, physical gestures, speech prosody, physiology (heart-rate, breathing-rate, and electrodermal activity), and other human behavior are all concrete examples of observable expressions, context is more amorphous and generally refers to the relationship of these expressions to each other and the external environment [99]. Moreover, context is multidimensional and difficult to circumscribe with a single label. In a recent experiment examining facial expressions across contexts in video, context is defined as the 653 categories that a neural network has been trained to classify, which include categories such as breakfast, car, humor, airport, lake, bottle, and mother where mother can refer or "pertain to mothers in any number of ways, ranging from footage of actual parenting to a man discussing his mother" [132]. This definition describes an algorithmic classification schema that identifies potentially useful yet vague aspects of context.

8.2 Building a Framework for Context

How can we systematically identify the roles of context in automated affect recognition? First, we need a language to discuss what we mean by context in affect recognition. In the abstract, context represents a complex high-dimensional feature space representing the inter-relatedness of elements that are often only partially available to observers. In affective science, context has been described as the collective "unmeasured factors" that contribute to how emotions are constructed; in the same paper, the authors describe the most salient, yet often unmeasured contexts as situational, social, physical, mental, temporal, personal, and cultural [54]. Likewise, in another review of context in emotion research, context is presented as a framework made up by three major components: personal, situational, and cultural features [240]. By explicitly identifying these categories rather than leaving context as a catch-all term for anything unmeasured, we can begin to build a framework to more precisely evaluate and discuss the varying roles of context in affect recognition.

When we examine affect recognition performed by computers, we need to take additional context into account. In the field of symbolic artificial intelligence (AI), ontology engineers have developed frameworks for incorporating context in common sense reasoning on natural language processing tasks [252, 372]. These frameworks have been useful for identifying assumptions that are often taken for granted in human communication but necessary for machine communication. In particular, the context identified in symbolic AI includes epistemological components addressing system-level questions like how to arbitrate opposing perspectives, what serves as evidence, what can be assumed, what expertise is required for making observations and judgments, and who believes a claim and why. In affect recognition within the context of affective computing, these questions become: how do we semantically represent affect, how do we label data, what do we assume about the accuracy of any human or machine appraisal of affect, what qualifies someone to label data, and how do we evaluate a model's performance.

Drawing from and expanding on entry points from both affective science and symbolic AI, we identify seven key categories to consider in automated affect recognition: ambient sensory environment, methods of measurement, semantic representation, situational constraints, temporal dynamics, sociocultural perspectives, and personalization. Our aim in establishing this seven-category framework is not to establish a new theory of emotions nor to claim there cannot be an eighth category, but instead, our aim is to take concrete steps toward unifying the many useful elements of context for affect recognition that have been already articulated in the affective science and affective computing literature. As such, we aim to synthesize a framework that provides both a theoretical foundation and a practical set of constructs. We are most inspired when theory and practice support each other, and since practice in affect recognition is growing rapidly, we seek to advance a theoretical framework for context that can grow with it, supporting and strengthening the growing practice. We describe the seven categories briefly below.

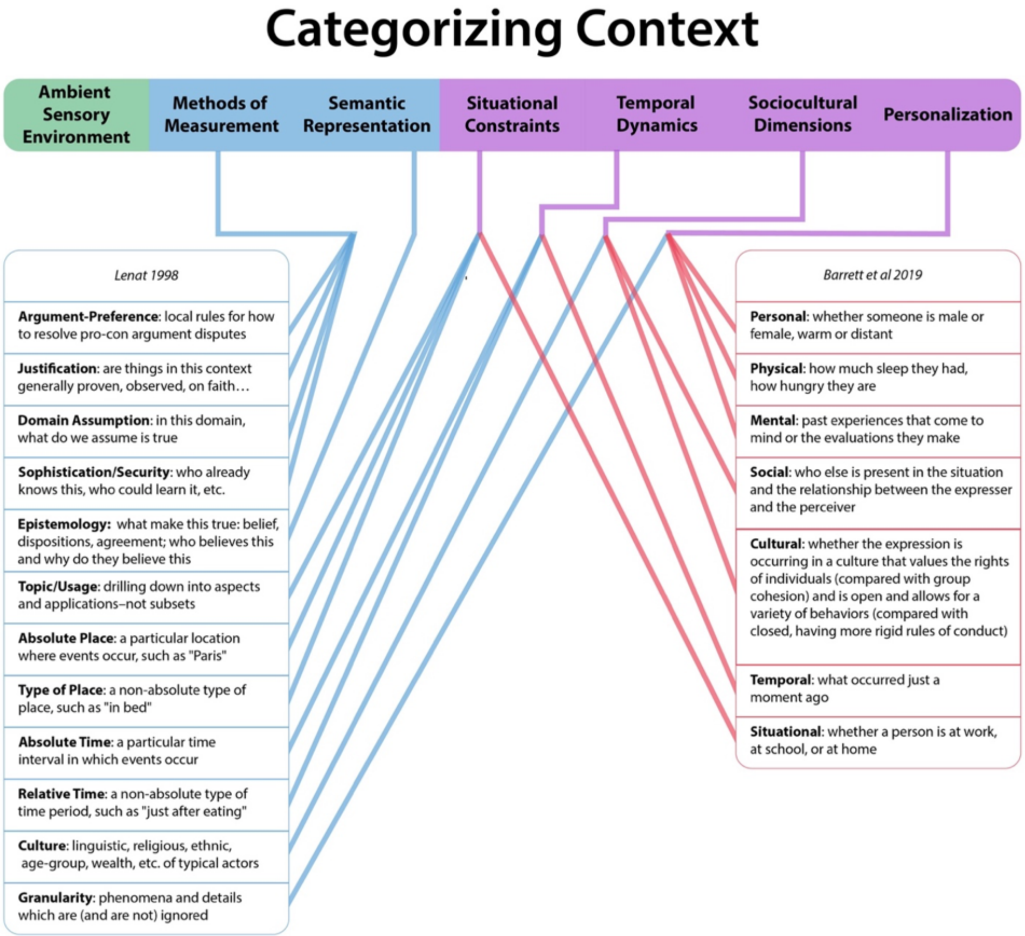
The first category, ambient sensory environment, refers to the sensory aspects of one's immediate surrounding settings e.g., the weather, soundscape, scenery, and smells. While ambient sensory environment does not neatly fit into any of the categories specified by Lenat 1998 or Barrett et al 2019, ambient sensory environment includes the face-context pairings described in earlier affective science research e.g., "face imbedding" (information within an image around a target face) and "response coherence" (information on congruence of facial expressions with non-facial expressions) [408]. The next two categories, methods of measurement and semantic representation are based on the five categories in the symbolic AI framework which focus on the system-level, epistemological concerns that are relevant for training machine learning models to predict affect labels. The final four categories occur in both Lenat 1998 and Barrett et al 2019. Situational constraints refer to constraints imposed by the activity or venue within which something is happening. For example, the inability to safely take one's eyes off the road while driving is a situational constraint. In the Lenat 1998 framework, situational constraints are further divided into topic/usage (referring to activity) and absolute place (referring to a place like the pyramids of Giza or the Golden Gate bridge) and type of place (referring to a place like a pizza joint or a shower), and here, we address all three of these categories together. Temporal dynamics refer to the dynamic nature of expressions and the trajectory and seasonality of emotional events. Sociocultural dimensions are the components of context related to other people. Finally, personalization refers to individuals' idiosyncrasies, which can range from an individuals' tastes and preferences to mental disabilities. All of these categories of context can overlap, and they are not mutually exclusive. These categories serve as a starting point to systematically examine how each different component of context situates expressions and shapes the appraisal and recognition of emotions.

8.3 Evaluating Automated Affect Recognition

Instead of detailing the role of each category here, we address the epistemological categories (methods of measurement and semantic representation) by asking: how can we evaluate the accuracy of an affect recognition model? In order to empirically evaluate a statistical learning model, we identify a source of human-provided (ground truth) labels, y, upon which to compare the model's predictions, $\hat{y}$. For affect recognition, ground truth labels usually come from one or more of these sources: individuals' self-reports of what they feel, external observers' reports of what they perceive others to experience, and experimentally-elicited or situationally-driven emotions. These three different methods of measurement are all useful yet imperfect for representing ground truth.

Self-reports provide an opportunity to collect ground truth labels based on an individual's inner feelings, but self-reports are subject to willful deception, can be inhibited by interoceptive ability and alexithymia, and are subject to social and cognitive biases [308]. For example, acquiescence bias is one particularly pernicious bias where research participants tend to agree with what they think the researchers want to hear [578].

External observers' reports can be collected by impartial and emotionally intelligent third parties. Most adult human observers know that outward appearance of affect does not necessarily reflect an individual's inner feelings, and as such, observation generally involves applying theory of mind and pragmatic reasoning about the target individual's expressions and situation before assigning an emotion label. Nonetheless, external observers' reports (just like self-reports) are not guaranteed to match an individuals' inner feelings. Moreover, while this approach allays concerns about willful deception and interoceptive ability, it cannot rule out social and behavioral biases of observers. One advantage of examining external observers' labels (as opposed to self-reports) is the ability to control the information to which the observers have access (e.g., a video with audio, audio only, silent video, a video with a mask over the target individual or background, a full body photograph, a photograph



Lenat 1998

Argument-Preference: local rules for how to resolve pro-con argument disputes

Justification: are things in this context generally proven, observed, on faith…

Domain Assumption: in this domain, what do we assume is true

Sophistication/Security: who already knows this, who could learn it, etc.

Epistemology: what make this true: belief, dispositions, agreement; who believes this and why do they believe this

Topic/Usage: drilling down into aspects and applications–not subsets

Absolute Place: a particular location where events occur, such as "Paris"

Type of Place: a non-absolute type of place, such as "in bed"

Absolute Time: a particular time interval in which events occur

Relative Time: a non-absolute type of time period, such as "just after eating"

Culture: linguistic, religious, ethnic, age-group, wealth, etc. of typical actors

Granularity: phenomena and details which are (and are not) ignored

Barrett et al 2019

Personal: whether someone is male or female, warm or distant

Physical: how much sleep they had, how hungry they are

Mental: past experiences that come to mind or the evaluations they make

Social: who else is present in the situation and the relationship between the expresser and the perceiver

Cultural: whether the expression is occurring in a culture that values the rights of individuals (compared with group cohesion) and is open and allows for a variety of behaviors (compared with closed, having more rigid rules of conduct)

Temporal: what occurred just a moment ago

Situational: whether a person is at work, at school, or at home

Figure 8-1: Key categories of contxt in automated affect recognition. Previously articulated components are listed verbatim as they have been described in previous frameworks except for "Domain Assumption" and "Epistemology," which are paraphrased for clarity.

showing only the face, or many other permutations), because manipulation of information modalities enables research into context effects.

The third approach to collecting ground truth labels is generating situations known to elicit particular emotions. For example, an experiment could elicit affect by asking a participant to reflect on a past emotional experience, ask a participant to count backwards from 100 by 7s (which often elicits stress), or routing a participant's car into rather than away from traffic jams (which often elicits stress or sometimes anger) [333, 373]. However, experiments designed to elicit emotions in participants do no always elicit the intended emotions because people respond to different situations differently. Moreover, laboratory conditions often do not match real-world settings, which raises questions about how well the findings of an experiment generalize to the real-world.

Measuring affect requires selecting a method for representing affect. It is well known that affect can be represented as: continuous affective dimensions (e.g., valence, arousal, dominance) and discrete emotion categories (e.g., joy, anger, fear, sadness, disgust, surprise). Affect can also be represented as: emotion categories connected by continuous gradients (e.g., horror, fear, disgust, anxiety), mixtures of emotion categories (e.g., angrily surprised, sadly fearful), enduring states (e.g., frustration, stress, pain, anxiety, depression), or even by sets of symbols like emojis, which can represent discrete or mixed and overlapping states. For example, emojis can represent emotions that are otherwise difficult to express via text. By training a machine learning model on a large corpus of tweets using sets of emojis as labels, researchers achieved state-of-the-art performance on three natural language processing benchmark tasks including emotion classification, sentiment analysis, and sarcasm detection [197]. While there are many competing theories of emotion, there is no universal agreement on how emotion should be represented [53, 135, 137, 172, 183, 276, 322, 338, 542, 543, 552]. The choice of how affect is represented will influence how an affect recognition model is trained and ultimately how accurately it recognizes affective states.

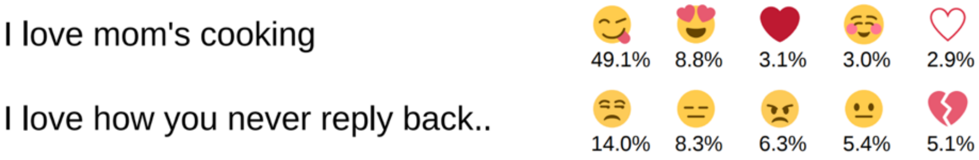


Figure 8-2: Probability distribution of the five most likely emojis associated with the two lines of text that both begin with "I love..." Only the first line is generally accepted by people as positive. By training natural language processing models on more than one billion tweets with emojis, Felbo et al 2017 obtained state-of-the-art performance on eight different benchmark datasets for sentiment, emotion, and sarcasm detection with a single pretrained model. Figure 2 is adapted from Felbo et al 2017 [197].

We evaluate the accuracy of an affect recognition model and its generalizability on data the model has never previously seen. Consider a model represented algebraically as $\hat{y} = f(x, c)$ where $\hat{y}$ represents the predicted affect label, x indicates the physical expression data, and c signifies context. Once the model has been trained on an initial dataset, we can evaluate its performance on a hold-out set and compare $\hat{y}$, the machine-predicted affect labels, to y, the human-provided labels. This allows us to evaluate a range of accuracy metrics including sensitivity, specificity, F1-score, AUC, log-loss, Pearson correlation coefficient, Matthew's correlation coefficient, and Cohen's kappa among others. In assessing how well a model recognizes self-reported emotions or observed emotional states, "a reasonable criterion of success is to get a computer to recognize affect as well as another person, i.e., better than chance, but below 100% accuracy" [500]. In some instances where physiological signals from the autonomic nervous system are imperceptible to humans without computational tools, the evaluation criteria shift to how well these otherwise imperceptible signals predict experimentally elicited emotions or long-term measures of mental health like reduced stress, reduced casualties while driving, or better learning outcomes [102, 341].

In practice, the assumption that the training and holdout data are independent and identically distributed (i.i.d.) often does not hold because context changes. As such, real-world implementation of automated affect recognition systems needs to explicitly incorporate as much contextual information as possible to most effectively generalize and avoid spurious correlations between observable expressions and affective labels. A recent review on facial expressions and emotions concludes that context matters for interpreting emotions from facial expressions: "When facial movements do express emotional states, they are considerably more variable and depend on context," [54]. This conclusion refreshes the need to examine an engineering question: Can we measure the contexts that inform the relationship between facial expressions and emotional states? This is not a new question in the field of affective computing; the development of large-scale datasets for facial expression recognition in the wild (e.g., EmotiW, Aff-Wild) draws from the premise that context mediates how facial expressions are interpreted [158]. The limitations of context-free affect detection and the importance of context-awareness were discussed as core challenges to building affect recognition systems a decade ago [98, 260, 623].

Recent advances in sensing technology and neural networks have enabled researchers to incorporate context more effectively than ever before, which now raises additional questions: What contexts are informative for affect recognition, and how can we measure these contexts? In this chapter, we identify seven key categories of context that should be considered in artificial intelligence systems for affect recognition.

8.4 Seven Key Categories of Context

8.4.1 Ambient Sensory Environment

Imagine yourself on a road-trip with friends driving through beautiful countryside with your windows rolled down. As you approach a large farm, the paved road turns into dirt and the ride becomes not only bumpy as you drive over potholes but also smelly from piles of cow manure. The sensorimotor and olfactory aspects of the car ride have changed. If the car is equipped with an affect recognition system using sensors to detect heart-rate variability or respiration rate, then the potholes will likely overwhelm the sensor's measures and produce errors if the movement from the potholes is not accounted for. If you roll up the windows to avoid the smell, then you will reduce the ambient noise in the car. As a result, you no longer need to speak over the wind and an affect recognition system might adjust its sensitivity to vocal stress features because they are related to ambient noise [266, 392]. Now imagine clouds begin to appear overhead. If the clouds change how the driver's face is illuminated, then the system's ability to recognize facial expressions may change [584]. All of these sudden environmental changes along this imaginary road-trip could be measured with sensors, which would allow an automobile-based affect recognition system to incorporate the ambient sensory environment and provide a more accurate assessment of the driver's affective state and the system's uncertainty about the driver's affective state than any system could provide given only the driver's expressions.

Recent research in affective science provides a series of examples demonstrating the role of immediate visual surroundings on how accurately humans appraise affect. In a study examining how well participants can distinguish between tennis players' emotions after winning or losing points, participants could not reliably distinguish between intense positive and negative emotions based only on the athlete's facial expressions; however, participants could accurately recognize the athlete's emotions when considering the athlete's full body expression [42]. Facial expressions can be ambiguous and factors as simple as facial orientation toward an observer can change perceptions of dominance in expressions of emotion [355, 644]. Moreover, different stereotypical facial expressions of the same emotion may interact differently with background scenery; in a recent study, researchers paired the same open and closed-mouth facial expressions of disgust with varying background scenes and found participants' appraisal of facial expressions varies more when the facial expressions were open-mouth [528]. Likewise, a recent study demonstrated that appraisals of facial expressions can be influenced by digitally manipulating body postures and background scenery [527]. In another recent experiment, participants were randomly assigned to view either silent videos or silent videos with human bodies occluded; participants inferred valence and arousal with high inter-rater reliability and similar accuracy to a group of participants who viewed the same videos without occlusions [113] What these experiments show is that immediate surroundings can be just as informative as (and sometimes more so than) facial expressions or physical gestures for how humans visually appraise emotion. Figure 3 provides a visual example demonstrating the importance of visual surroundings for accurately appraising emotion.

Like the human visual system, computer vision models can leverage statistical properties of a scene to more accurately recognize the affective state of an individual. In an experiment comparing a neural network trained on facial expressions, body gestures, and background scenes from non-posed, static images, researchers found that including the background scene improves recognition accuracy over a neural network trained solely on the face and body for twenty-six out of twenty-six different emotion categories and two out of three affective dimensions (valence and dominance but not arousal) [349]. This demonstration that background scenery informs automated affect recognition is a specific example of a more general concept in computer vision and pattern recognition: the relationship between elements within a scene helps to inform what any particular element is recognized as [473, 652].

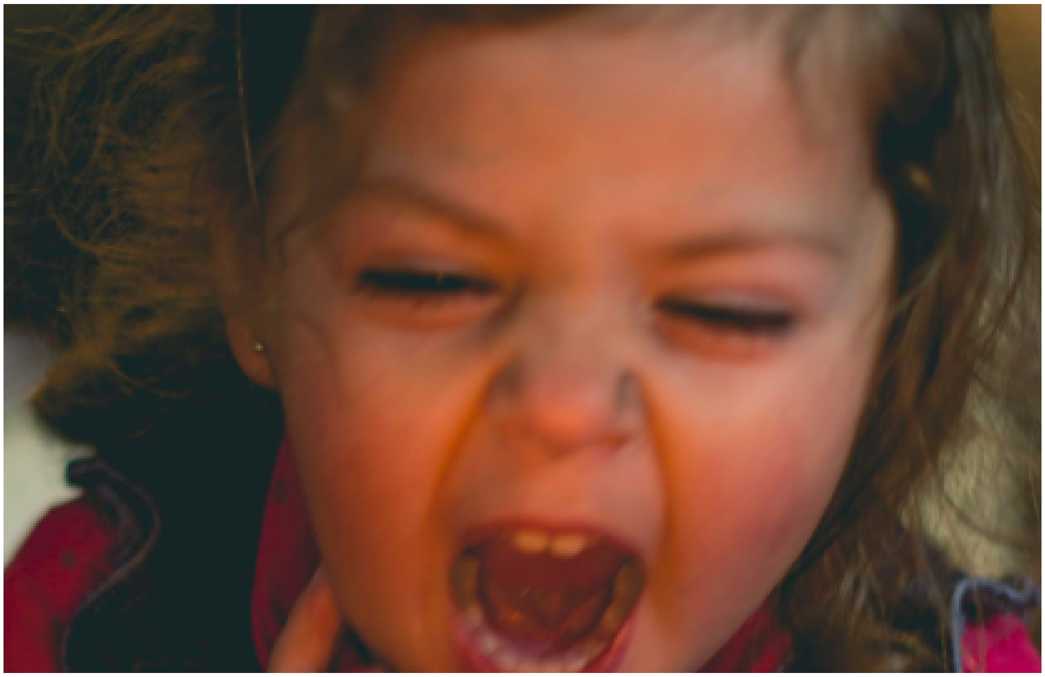


Figure 8-3: Close-cropped photo of a young girl's face. With no context beyond the close-cropped photo, the young girl's facial expression could easily be perceived as rage, excitement, or perhaps some other emotion. See Figure 4 to see the uncropped photo of the young girl. Figure 3 has a Creative Commons Zero License.

While the background can offer important contextual cues, it can also produce misleading heuristics that work for most examples but do not always hold. For example, neural networks trained to classify a bird as a land bird or water bird without proper regularization will use the presence of water in the background as a shortcut, which leads to misclassifying images of water birds temporarily resting on land [400, 546]. In contrast, only focusing on the object of interest (e.g. birds or facial expressions) can lead to ignoring important information. One solution for integrating background information without making errors from shortcut learning is to characterize the relationship between the central object and its surroundings. For example, consider individuals with extreme sensory sensitivities; if a screaming baby enters the room of an autistic sound-hypersenstive child, we would predict the child would experience a strong stress reaction because we recognize a relationship between the child's stress and her surrounding sensory environment.

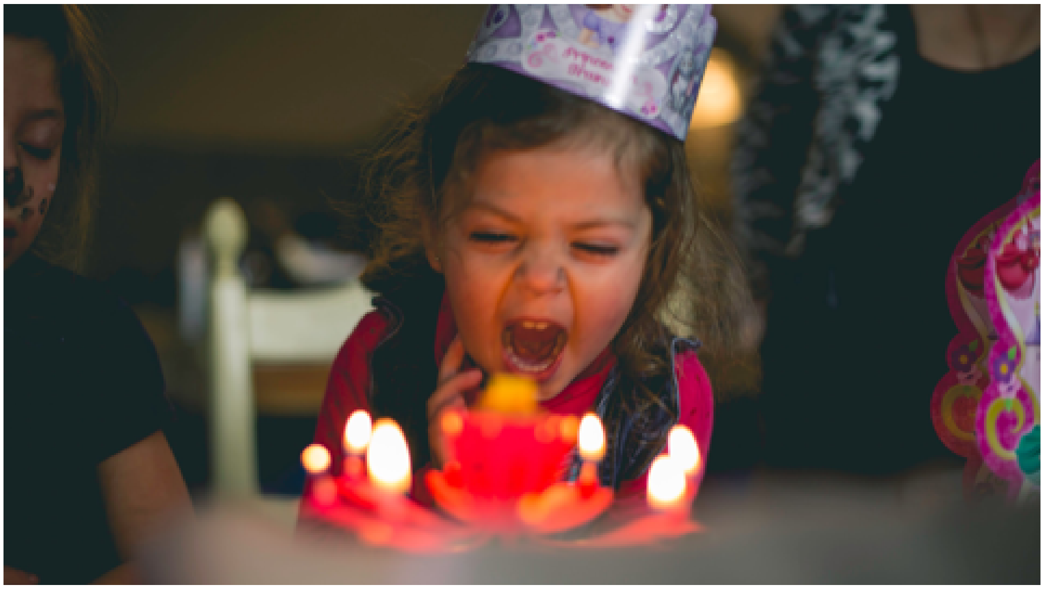


Figure 8-4: A photo of a young girl about to blow out candles on her birthday cake. See Figure 3 for a cropped version of this photo. Figure 4 has Creative Commons Zero License

8.4.2 Methods of Measurement

Addressing how affect is measured includes addressing how human-provided labels are generated, how data collection influences human behavior, how sensors are situated to collect data, and how variability in human-provided labels are handled.

The observer effect – the disturbance of a system by the act of observation – is important to consider in automated affect recognition. If people believe they are being watched and analyzed, they will behave and express emotions differently [216]. Recent research shows that machine learning and signal processing applied to webcam video and wi-fi radio waves can accurately estimate respiration, heart-rate variability, and affective states using information that most people cannot see without special sensing technology [112, 414, 415, 504, 666]. These new developments raise privacy concerns and may already be changing individuals' behavior. In response, recent research has focused on building algorithms to remove physiological signals by making subtle alterations to video that confuse a computer system but are imperceptible to the human visual system [119, 491]. As technology for measuring affect and preserving individuals' privacy progresses, people's behavior will adapt, which is another aspect that affect recognition models should consider.

The accuracy of an affect recognition model can depend on how sensors collect physiological features. For example, electrodermal activity (EDA) is a signal that changes with sudomotor innervation, which can be elicited by changes in temperature, exertion, and arousal. The ability to measure arousal depends on the ability to control for artifacts from physical movement and other noise [600]. Moreover, EDA measurements depend on where an EDA sensor is placed on the human body. Traditionally, EDA measures are obtained from a pair of electrodes placed on a palmar or plantar site. Recent research shows that alternative electrode locations carry meaningful, and different, signals related to emotion, likely arising from pathways that connect different regions of the brain to different regions of the skin [316, 502]. By evaluating varying placements of sensors and developing algorithms to control for varying physiological artifacts, it is possible to reduce – though likely not eliminate – several kinds of measurement error.

Another approach to reducing errors in affect recognition models involves examining and controlling for inter and intra-annotator variability in third party appraisals. For example, a Monte Carlo dropout method for disambiguating annotator bias (variance from inter-annotator agreement of subjective judgements) from data bias (variance across clusters of similar images) can highlight a model's accuracy disparities across subsets of the data e.g., images of dark-skinned people and highly illuminated images [224]. One technique for incorporating inter-annotator variability in affect recognition models is to represent emotion labels as a statistical distribution of annotations provided by multiple observers, which allows a model to incorporate both diverse perspectives and the relative subtlety of an emotional expression [261, 662]. Additional work shows that the labeling process can improve inter-annotator reliability when it is performed via relative ordinal rankings rather than absolute values or categories because rankings force observers to contextualize expressions relative to one another [657]. Furthermore, many experiments have shown affective priming on third-party observers can alter the labeling processes; affective primes change how an observer recognizes both neutral words and expressions [80, 257, 556, 565]. By highlighting where bias can creep into the labeling processes, we can identify where an affect recognition model is likely to make inaccurate predictions.

8.4.3 Semantic Representation

The representation people use to describe affect influences the evaluation of affect appraisal tasks. For example, a convolutional neural network trained to predict twenty-six emotion categories from images is more accurate when trained on people's faces and bodies than the background for twenty-five of twenty-six emotion categories while the same model architectures, re-trained to predict affective dimensions (valence, arousal, and dominance), are just as accurate or more accurate when trained on the background than when trained on people's faces and bodies [349].

Depending on the medium and the specific content within that medium, the most useful semantic representation can change. In a recent experiment where participants reported the emotional states elicited by short videos, a dimensionality reduction technique applied to participants' responses revealed that twenty-seven emotional categories reliably capture the variance in self-reported responses better than affective dimensions like valence, arousal, dominance, and others [130]. With the same dimensionality reduction methodology, another experiment reveals that for participants across two different cultures speech prosody can be best characterized by twelve categories of emotion [131].

Moreover, words conveying emotions are not always perfectly translatable across languages and cultures. For example, saudade in Portuguese and natsukashii in Japanese are similar to nostalgia in English but not exactly the same, and there's no exact translation in Hindi or Malayalam for disgust [192, 345, 346]. Neologisms like vemödalen – the fear that everything has already been done – can help people recognize an emotion they have experienced that they have not yet been able to articulate [342].

In affect appraisal tasks, researchers may ask participants to choose the best fitting word from a list, select degrees of affective dimensions, describe an affective state in one's own words, find a story that has matching emotions, or react with one or more facial expressions or perhaps even emojis. In a recent experiment on emotion recognition in storytelling where participants were randomly assigned to three conditions – watch a silent video, listen to the audio, watch the video with audio – participants were less accurate at recognizing emotions in the silent video condition than the other two conditions, but the silent video condition was associated with a higher degree of heartrate synchrony between the participant and the storyteller [305] Different contextual modalities yield different levels of agreement across how affect appraisal is measured.

In practice, commercially available facial expression recognition systems typically detect and label facial expressions as facial action units or classify expressions into a small number of categories such as happy, sad, surprise, disgust, fear, anger. This pre-set list of categories naturally imposes limits on what these commercial systems can do. In a recent comparison of eight commercially-available APIs with human observers on classifying facial expressions into six categories, humans outperform all eight APIs on videos showing people making facial expressions [176]. What is particularly notable in this study is that humans perform better on spontaneous expressions than on posed expressions while computers perform better on posed than on spontaneous expressions [176] Without details about the training data or the models used to create the commercial APIs, it is difficult to precisely identify the source of the model error in these commercial APIs, but these results suggest the commercial APIs may be trained on mostly posed (not spontaneous expressions), static images (not dynamic video), homogenous populations, or some other contextual component that differs from the videos in this experiment. If posed and spontaneous expressions represent emotions differently, then a useful affect recognition model might output two predictions: a predicted emotion category and a prediction for whether the expression is posed or spontaneous.

In contrast to commercial APIs, the latest research competitions in affect recognition focus on recognizing affect in the wild across multiple representations of emotional expressions (including facial action units, student engagement, physiological signals in response to video, and group-level cohesion) and across a variety of specific situations [159, 347, 636]. There is no single agreed upon semantic representation of affect, and as such, scientific research should continue to evaluate affect along its many semantic representations to avoid unintentional semantic biases across people, situations, and cultures.

8.4.4 Situational Constraints

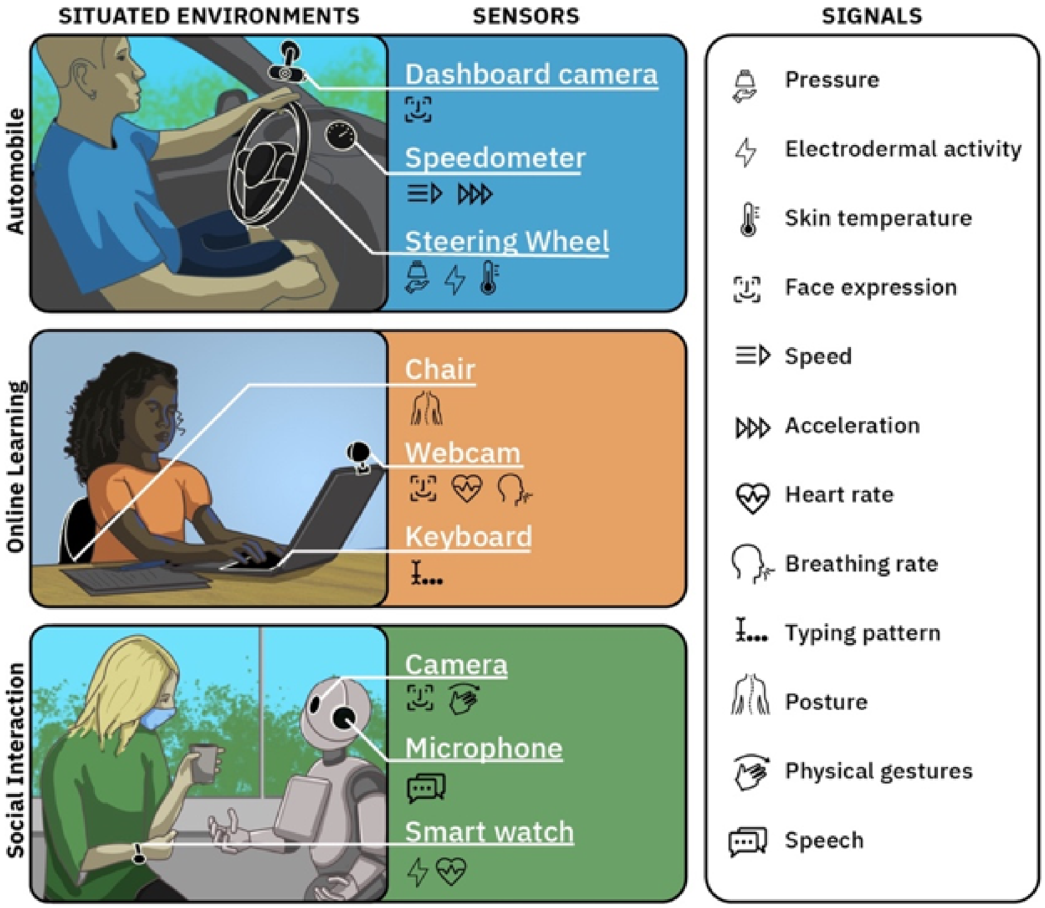


Figure 8-5: Figure 5 presents examples of affect recognition in three contexts: automobile driving, online learning, and human-robot interaction. Sensors in a constrained environment collect signals data to predict an individual's affective state. The sensors (and the signals that these sensors measure) are not an exhaustive list but rather examples of what are currently available for use in automated affect recognition.

Beyond driving, specific situations like online learning and human-robot interaction can reduce the state-space of possible behaviors and improve the accuracy of affective predictions.

In online learning, affect recognition systems can detect mind wandering and boredom from dynamic face expressions, posture, and keystrokes [71, 77]. This is particularly useful because the negative effects of mind wandering on learning outcomes can be counteracted with interventions like just-in-time questions and re-reading [138, 431]. As a separate example confined to people walking around their homes, mobile robots can use video of dynamic, three-dimension gaits to predict a small set of affective-states – happy, angry, sad, and neutral – significantly better than chance [454].

Depending on the situation, observers will appraise similar expressions as different emotions. For concrete examples, an open mouth and furrowed brow may be perceived as expressing different emotions in the context of a wedding, a job interview, a poker match, or a negotiation. Emotion is such a powerful tool in negotiations that human negotiators actively seek to deceive their human and machine counterparties by expressing emotions they do not inwardly feel [148, 420, 421]. In a prisoner's dilemma game, automated affect recognition has been shown to be as accurate as players at appraising their opponents' emotional state; moreover, information on the state of the game improves the accuracy of the model beyond what the model predicts based on facial expressions alone [277]

Previously, affect appraisal has been described as an automatic, effortless, and seamless process that can be applied to a photograph of a facial expression [307] However, rapid affect appraisal without reasoning about the context can lead to errors in both professional contexts and simple tasks like appraising an emotion from a photograph (e.g. Figure 2 and Figure 3). In an experiment on reasoning about emotion, researchers found that affect appraisal by participants of computer-generated facial expressions and associated situational cues are better predicted by a Bayesian integration of the two signals (the facial expressions and situational cues) than by either signal alone [475] Building upon these findings, researchers have proposed an intuitive theory framework for predicting emotions based on Bayesian inference over a graphical model incorporating an individual's expressions and actions in response to a particular situation [474] According to such a model, identical expressions and behaviors could be appraised as completely different emotions depending on the situation, which suggests that affect recognition systems will be most accurate when considering both observable expressions and the associated context.

8.4.5 Temporal Dynamics

Imagine watching a friend listen to an audio recording of her favorite standup comedian. As the stand-up comedian sets up the joke, the corners of your friend's lips begin to rise. When the comedian delivers the punchline, a large smile shoots across her face as she tilts her head back in laughter. If you only saw a static photograph of her smile, you may wonder if her smile was sincere or not, but with the perfectly synchronized temporal dynamics of her expression with the comedic timing, you have strong evidence that her smile is expressing sincere amusement. Research shows that temporal dynamics help humans discern between genuine and fake smiles [356, 357] This is important because smiles do not always indicate joy or amusement; in fact, researchers found pre-school children will smile after a failure more often than after a success [555]

A static photograph of a smile is much less informative than a video showing a smile's dynamic trajectory. In an experiment where neither human annotators nor machine learning models could discern between smiles of frustration or smiles of delight in photographs better than random, a model trained to identify smile type via dynamic muscle movements from videos could classify smile type significantly better than both random guessing and human participants [282] In another experiment where participants recorded themselves watching clips of the 2012 Obama-Romney presidential debate, researchers found that smirks followed by smiles tend to express a different affective state than smirks without an accompanied smile [413]. In a submission to the EmotioNet Challenge 2020, researchers demonstrated that different facial action units converge at various speeds and as such choosing the optimal checkpoint improves the recognition accuracy of facial action units [629]. While temporal dynamics can improve recognition accuracy, recent research reveals a human tendency for outside observers to overestimate a target individual's emotional state when viewing sequences of images and videos, which suggests videos labeled by outside observers may include a biased amplification of emotion [232]. Despite the importance of temporal dynamics, a recent review of research on automated facial expression recognition shows that only 4 out of 17 publicly available datasets currently include videos or image sequences [382].

Beyond facial expressions, temporal dynamics play a role in recognition of body posture and evaluating speech and conversation. For example, models can automatically recognize students' level of interest and boredom during learning experiences from posture dynamics [78, 442]. Likewise, models can consider varying emotion states and non-linguistic utterances throughout the flow of a conversation to improve emotion classification accuracy [399, 507, 508].

Future emotions can be predicted by current expressions and emotions. Recent experimental evidence supports the facial feedback hypothesis that facial expressions may influence the subsequent emotions people feel, e.g., happy poses lead to happy feelings [127]. Likewise, third party observers' mental models of others' emotional dynamics predict future emotions based on current emotions much better than both chance and reductionary models based only on valence or the holistic similarity between emotions [603].

Another component of temporal dynamics is recurrence and seasonality. People behave differently in the morning versus night, weekdays versus weekends, summer versus winter, and workday versus holiday. These temporal dynamics are often related to other categories of context. For example, seasonality is intertwined with weather – it's often cold in the winter and hot in the summer, but it can be the reverse in San Francisco – and culture – the weekend starts on Friday evening in the United States but it starts on Thursday evening in Arabic-speaking countries.

8.4.6 Sociocultural Dimensions

In lab experiments, social and cultural context play a significant role in how humans perceive facial expressions. For example, in an experiment presenting dynamic, virtual facial expressions to participants, western Caucasian participants perceive six emotions (happy, sad, surprise, disgust, fear, anger) as a distinct clustering of facial movements while East Asians do not [297]. In a follow-up experiment, participants perceived the same kinds of facial expressions as pain regardless of their cultural heritage whereas participants perceived facial expressions of orgasms differently depending on their cultural heritage [110]. Multiple studies show East Asian cultures are more sensitive than Anglophone cultures to social context when appraising emotion [406, 407]. In addition to cultural context, gender stereotypes can make it difficult to disentangle facial expressions of emotion from gender such that the same facial expression is perceived differently depending on whether it's a man's or woman's face [306]. Moreover, age matters; results from a recent experiment show older adults are better at discerning emotionally incongruent information between face expressions and physical gestures [213] The social context – whether one's friends or other people are around – changes how individuals express their emotions, and as such, it changes how observers should appraise expressions [136, 217]. In an experiment on appraising virtual facial expression blends (e.g., shame-sadness), participants are often influenced by the expressions on other virtual faces that appear to be socially interacting with the blended face [448]. Beyond lab experiments, there is plenty of evidence that people hide their emotions or express the opposite of what they feel, particularly at work in the midst of relationships with power dynamics [237] While these are not an exhaustive list of sociocultural dimensions at play in affect recognition, these examples help to identify the kinds of social and cultural information that need to be considered when incorporating labels from external observers for affect recognition models.

The relative social status of third-party observers to observed subjects may serve as an affective prime that can influence how emotions are annotated. For example, people tend to evaluate ingroup faces as more positive in a circumplex affect grid than outgroup faces [367]. Previously, we highlighted research showing that affective priming can bias third-party annotations of expressions [565]. If third-party observers claim to feel empathy towards the subject whose emotion they are annotating, then a question arises whether the observer's emotion really matches the observed person's emotions. In fact, perceived social status and agency can mediate envy, pity, and feelings of anger versus sadness for another person's situation [467, 649]. Recent research on facial expression reactions to positive and negative images reveal wide variation in facial action unit intensity between observers' reactions to the same images [273].

The data and labels upon which affect recognition models are optimized come with a social and cultural context. Models trained with biased data on the basis of race and gender lead to models that produce similarly biased outputs [74, 87, 95, 248]. If an annotator racially stereotypes people as calm or angry, then this bias will likely appear in the algorithm's outputs. Without diverse representation in both the labeling community and the data, biases are likely to be encoded in affect recognition algorithms and will correspondingly bias predictions when the model is applied to new datasets. As more datasets are labeled, researchers should be careful to include social and cultural information from annotators such that label bias can be inspected for expected sociocultural biases. Recent research in fairness and transparency in machine learning highlights several approaches for how to effectively document information for future algorithmic auditing [87, 433, 518].

8.4.7 Personalization

Instead of optimizing for accuracy across a group of individuals, personalized machine learning optimizes for accuracy within each individual. People not only vary from others in how they express affect, but they also vary from themselves over time. In one of the earliest iterations of an automated affect recognition system, researchers discovered that there is more variation in day-to-day physiological measurements of a single emotion than there is across eight different emotions on the same day [501]. By normalizing physiological measurements according to an individual's daily baseline, researchers significantly increased the accuracy at which the system recognized emotional states [501]. In a more recent example demonstrating the effectiveness of personalized machine learning, researchers applied multitask learning to longitudinal data (surveys, electrodermal activity, sleep activity, smartphone usage, and weather) to predict self-reported next-day levels of mood, stress, and overall health [106, 599]. By customizing for the needs of each individual, accuracy on predicting mood, stress, and health increased from 66%, 68%, and 59% to 78%, 82%, and 82%, respectively [599] These performance improvements mirror work on recognizing affect in paralinguistic speech that exploits the speakers' demographic features and personality characteristics to improve recognition accuracy [664].

Personalized affect recognition models are designed to learn to recognize how individuals express affect rather than how affect is expressed on average. In another recent example, researchers designed a neural network architecture – a feature layer, a context layer (including demographics, behavioral assessment scores, and information varying at the individual layer), and an inference layer, which are illustrated in Figure 6 – for predicting valence, arousal, and engagement of autistic children that outperforms a neural network architecture that does not include a personalization component [539].

Individuals' responses to an emotional cue or particular situation depend on personal experience. For example, individuals categorize verbal and nonverbal vocalizations as calm or upset not according to a universal cut-off but rather depending on the recent distribution of vocalizations they encountered [650]. Another example revealing how personal experience influences affect appraisal comes from developmental psychology research on children who have suffered from abuse: abused children identify dynamic facial expressions of anger faster than a control group and abused children over identify anger in facial expressions conveying mixed emotions relative to a control group of children who under identify anger in these same stimuli [505, 506]. In some cases, there may exist no more than a single outside observer or small cohort of observers who can recognize an individual's affective state. In research on non-verbal or minimally-verbal individuals who express their affective states in diverse, non-traditional, audible ways, researchers designed a personalized labeling and machine-learning system to recognize non-verbal individuals' affective-cognitive states, e.g., frustration, delight, dysregulation, self-talk, request [453]. The labels were recorded in real-world environments by caregivers who intimately knew the individuals. While affect recognition models often benefit from labels produced by a diverse population of third-party observers, this is an example where deep knowledge of the non-verbal individual's communication style is essential for identifying the correct label. By applying transfer learning to the personalized labels, personalized affect recognition models can incorporate individuals' specific idiosyncrasies such that generally unrecognized sounds can be automatically recognized by a computer and communicated to caregivers who would otherwise not be able to understand these non-verbal and minimally verbal individuals.

8.5 Conclusion

Context in automated affect recognition has been difficult to precisely define because context refers to everything, past, present, and anticipated, that influences the elicitation and persistence of emotion, its expression, display, interpretation, and measurement. Nonetheless, we can develop a framework, which allows us to systematically evaluate the most salient categories of context: the visual, olfactory, auditory, and scene-specific elements, how emotions are labeled, measured, and represented, the activity and venue in which something occurs, temporal dynamics, social and cultural factors, and individuals' idiosyncrasies. By explicitly naming each of these categories of context in Figure 1 and discussing the role of each, we have presented a framework intended to help researchers more precisely consider varying aspects of context in affect recognition. This framework is intended to be useful for studies that involve gathering human provided labels on emotion, studies on developing models for automated affect recognition, and studies evaluating models for automated affect recognition. In this article, we have shown how this framework can be used to organize many examples of recent research addressing contextual influences on automated affect recognition.

As we continue to name and measure the categories of context, aggregating and organizing the huge effort of many researchers, we build a language that enables us to precisely discuss variability arising in human-centered statistical learning. Moreover, these categories highlight important generalizability considerations for affect recognition models and experimental findings. Future models and experiments can frame their implications and limitations with respect to the contexts they were and were not able to incorporate. There will likely always be unexplained variance that neither models nor third-party observers can account for. Automated affect recognition does not require hard and fast rules or unique emotion signatures, but rather, it infers or predicts an unobserved state based on patterns in expressions and their associated contexts, which are becoming increasingly observable as technology advances.

Acknowledgments

We acknowledge funding from MIT Media Lab Consortium member companies. We thank Desmond Ong, members of the Affective Computing lab, and participants at the Affective Cognition workshop at Cognitive Science Society conference for helpful feedback.

Chapter 9

Identifying the Context Shift between Test Benchmarks and Production Data

Abstract

Machine learning models are often brittle on production data despite achieving high accuracy on benchmark datasets. Benchmark datasets have traditionally served dual purposes: first, benchmarks offer a standard on which machine learning researchers can compare different methods, and second, benchmarks provide a model, albeit imperfect, of the real world. The incompleteness of test benchmarks (and the data upon which models are trained) hinder robustness in machine learning, enable shortcut learning, and leave models systematically prone to err on out-of-distribution and adversarially perturbed data. The mismatch between two datasets has traditionally been described as dataset shift. In an effort to clarify how to address the mismatch between test benchmarks and production data, we introduce context shift to describe semantically meaningful changes in the underlying data generation process. In this paper, we identify three methods for addressing context shift that would otherwise lead to model prediction errors: first, we describe how human intuition and expert knowledge can identify semantically meaningful features upon which models make systematic errors, second, we detail how dynamic benchmarking – with its focus on capturing the data generation process – can promote generalizability through corroboration, and third, we highlight how clarifying a model's limitations can reduce unexpected errors. Robust machine learning concerns real-world model performance beyond benchmarks, and as such, we consider three domains for machine learning applications – facial expression recognition, deepfake detection, and medical diagnosis – to highlight how implicit assumptions in benchmark tasks lead to errors in practice. By paying close attention to the role of context in a prediction task, researchers can design more comprehensive benchmarks, reduce context shift errors, and increase generalization performance.[1]

9.1 Motivation

Dataset benchmarks offer a standard for comparing and evaluating the performance of machine learning models on real-world tasks like object detection [153], handwritten digit recognition [154], image captioning [118], general language understanding [626], affect recognition [350], deepfake detection [168], medical diagnosis (e.g. for skin disease [146], pneumonia [294], critical care [304], etc.), and many other tasks. As a standard for comparison, dataset benchmarks have enabled rapid progress in computer vision and natural language processing.

Despite intentions to create and curate data that match the real-world as closely as possible, the dynamic, high-dimensional, combinatoric complexity of many real-world tasks is often difficult to capture in a single static benchmark. Indeed, the development and evaluation of machine learning models on benchmarks often suffer from a variety of historical, representational, measurement, aggregation, and evaluation biases [582]. These biases can be further exacerbated by deployment biases where the task that a benchmark is intended to measure differs from the real-world task [589]. Moreover, data for benchmarks are often collected

[1] This chapter is currently available as a pre-print [242].

at scale with minimal oversight [540], which leaves data open to poisoning attacks [119], leakage [313], multiple interpretations [235] and error [465]. As a consequence, machine learning models that appear to be approaching (and sometimes surpassing) human-level ability on a test benchmark will often error when shown out-of-distribution data [605]. In other words, the reliance on static test benchmarks as metrics for projecting production performance inflates the accuracy of machine learning model performance [602] and leaves open the questions, "Can you trust your model? Will it work in deployment?" [385]

The meaning of out-of-distribution data depends on a task's context. Two canonical examples of out-of-distribution data in object detection tasks are images of either a cow on a sandy beach or a camel on a green pasture [32]. Today's commonly used training data rarely contain such animal-environment pairs, and as a result, machine learning models often learn spurious correlations such as cloven hoofed mammals next to sand are camels but the ones next to grass are cows. With *a priori* knowledge of potentially spurious correlations, one approach for addressing this kind of model brittleness is to include auxiliary labels that can serve as a causally-motivated regularization framework [400]. However, post hoc model explanations are often ineffective for identifying previously unknown shortcuts [8] (though both explanations via concept traversals [227] and identifying model failures as directions in latent space via contrastive learning where images and natural language are embedded in a shared latent space show promise [301]). In contrast, human intuition can identify many out-of-distribution contexts on which spurious correlations (sometimes called shortcut learning) may occur.

In one of the clearest examples of spurious correlations that lead to the benchmark-production gap, researchers recreated ImageNet [153] and CIFAR-10 [353] with new images and demonstrated that the state-of-the-art models' performances are significantly lower on the recreated versions of these datasets [523]. The benchmark-production gap is particularly salient in this example because these two datasets have been the most commonly used benchmarks for object recognition over the last decade. Recht et al 2019 explain that the drop in performance does not appear to be explained by random sampling error, hyperparameter tuning for optimizing performance on the original test set, or obvious changes in semantically meaningful features, but instead, the performance gap appears to arise from subtle changes in the data [523]. Object recognition is not as straightforward a task as it might appear at first glance and involves edge cases arising from a variety of contexts.

In complex human-centered machine learning applications, a task's context involves answers to the following kinds of questions: What is the task? For whom is the task designed? When and where does it take place? Why is it done? Are there any interventions happening that might alter features and labels associated with the task? And how is the task measured? The lack of clear answers to these questions indicates that the model and its evaluation lack generalizability simply because it is not clear to what the model should generalize. Likewise, clear answers to these questions without a corresponding diverse representation in the benchmark dataset to evaluate performance leaves open the question of whether the dataset generalizes to the contexts in which the model is intended to generalize.

As an example of a generalization failure in a human-centered machine learning application, consider facial recognition. In Joy Buolamwini's and Timnit Gebru's algorithmic audit of facial recognition benchmarks and classifiers, the authors reveal the most commonly used benchmarks for evaluating facial recognition accuracy were composed of images of people with predominantly light skin. In other words, images of people with dark skin were relatively out-of-distribution [87]. Furthermore, the Buolamwini and Gebru 2018 audit presented a new benchmark to evaluate accuracy across intersectional identities. Commercial gender classification models performed extremely accurately in identifying men with light skin (with a maximum error rate of less than 1%) but incorrectly in women with dark skin (with a maximum error rate of 35%) [87]. This large accuracy disparity reveals how failures to generalize can be hidden by benchmarks that do not represent the diversity of the real world. Research on machine learning applied to the diagnosis of skin disease reveals a similar story: models trained to classify skin disease based on images of only light or dark skin are more accurate in skin tones closest to the skin tones in the images in which the model was trained [248]. These examples corroborate the notion that simply optimizing for predictive accuracy with very large datasets can often misrepresent the true data generating process and lead to systematic errors [285].

In other domains like affect recognition, an out-of-distribution context can be very task specific. For example, spontaneous facial expressions can be out-of-distribution for facial expression benchmarks that primarily contain posed expressions [177]. Likewise, images labeled with emotions such as anger or surprise can be out-of-distribution for the same benchmarks where happy and neutral labels are most common [382].

Machine learning models that have been trained on perceptual data are subject to systematic failures on a special case of out-of-distribution data: adversarial perturbations. Adversarial perturbations refer to minor changes in data that do not influence classification of the data by humans but radically alter a model's classification. As an example, researchers have demonstrated that adding a small sticker to a stop sign can alter the classification of machine learning models' such that the models incorrectly classify the stop sign as a yield sign [85, 188]. Researchers have shown that one can generalize adversarial perturbations by attaching a mainly translucent sticker on the lens of a camera [381]. Likewise, researchers have demonstrated that adversarial perturbations can be applied to medical data e.g. noise or rotations in medical images and text substitution in medical notes and reimbursement codes [202]. In general, adversarial perturbations demonstrate a lack of model robustness [292], lead to model errors that reasonable humans would rarely make, and open the question: How can we build models that are invariant to the same semantically meaningful features to which humans are invariant? Training robust models with adversarial perturbations is a starting point for aligning model performance more closely with human perceptions [610], but it is often difficult to identify the comprehensive possibility space of adversarial perturbations.

What drives the systematic errors by machine learning models on out-of-distribution data? The next section discusses two perspectives for characterizing the benchmark-production gap: the distribution shift perspective and the context shift perspective. The rest of the paper describes three methods for addressing context shift and considers three case studies of context shift in facial expression recognition, deepfake detection, and medical diagnosis.

9.2 Systematic Errors Arise from Context Shift and Lead to Distribution Shift

The mismatch between two datasets (e.g. the train and test splits or a test benchmark and production data) has been traditionally described as dataset shift [514]. More recently, machine learning researchers have described the same concept as distribution shift. In order to illustrate the growing attention to and evolving semantics of distribution shift, we present the number of papers on Google Scholar containing both "machine learning" and "distribution shift" (and other sub-components of distribution shift) in Figure 9-1.

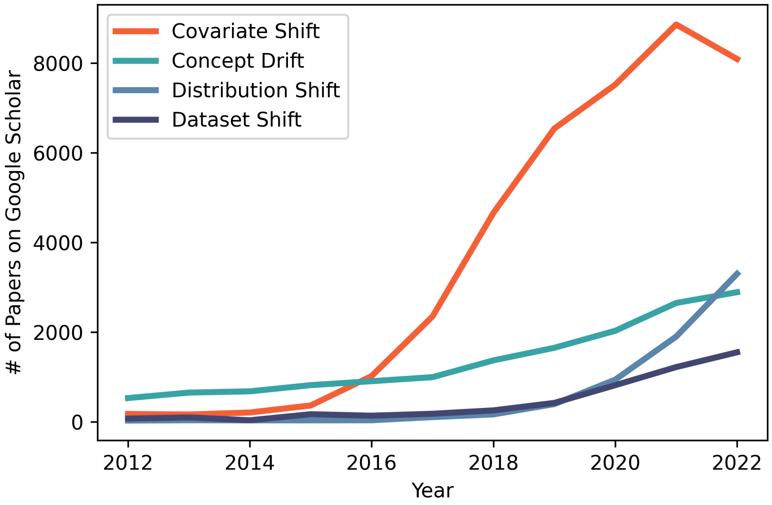


Figure 9-1: Number of papers on Google Scholar from 2012 to 2021 for search queries combining "machine learning" + the four most common terms for distribution shifts. For context, "machine learning" returns 185,000 articles in 2012, 597,000 articles in 2019 (the peak over the last decade), and 188,000 article in 2022. The terms "prior probability shift" and "concept shift" return 445 and 1,050 papers over all time, respectively, when paired with "machine learning".

Distribution shift refers to the non-equivalence of the joint distributions between two datasets. Formally, distribution shift describes the following equation $P_1(y, x) \neq P_2(y, x)$ where $P_n(y, x)$ is the joint distribution of labels, y, and covariates, x for a particular dataset, n [439]. Based on Moreno et al 2012, the four subcategories of distribution shift include **covariate shift** when the distribution of features changes but everything else remains the same, **prior probability shift** when the distribution of labels changes but everything else remains the same, **concept shift** (more commonly referred to as concept drift) when the distribution of labels conditional on features changes but everything else remains the same, and **other distribution shift** when none of the other three shifts hold but the joint distributions between two datasets is different. We illustrate examples of each shift in Figure 9-2 to motivate intuition as to how the changes appear. Moreno et al 2012 formally specify the

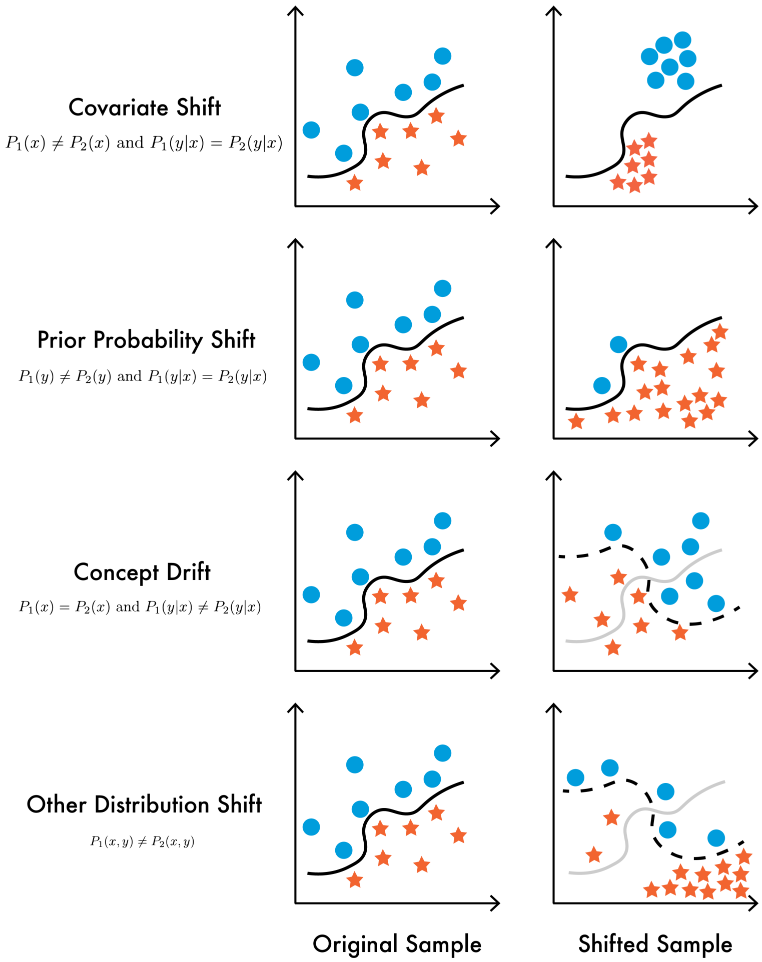

Covariate Shift
$P_1(x) \neq P_2(x)$ and $P_1(y|x) = P_2(y|x)$
Prior Probability Shift
$P_1(y) \neq P_2(y)$ and $P_1(y|x) = P_2(y|x)$
Concept Drift
$P_1(x) = P_2(x)$ and $P_1(y|x) \neq P_2(y|x)$
Other Distribution Shift
$P_1(x, y) \neq P_2(x, y)$
Original Sample
Shifted Sample

Figure 9-2: Illustrations of the four kinds of distribution shifts as defined in Moreno et al. 2012 [439]. The spatial positions represents the feature space, geometric shapes and colors represents the ground truth label, the solid boundary line represents the learned representation of labels from the original sample, and the dotted boundary line represents the learned representation of labels from the shifted sample. Most real-world distribution shifts involve changes across features, labels, and the relationship between features and labels, and as such would be characterized as "Other Distribution Shift." The core problem with the conceptual framework of distribution shift is that it is merely a symptom of changes in data-generating processes - how data are created, collected, and curated – but not part of the data-generating process itself. In order to improve model reliability and robustness, a data-centric perspective takes into consideration the data generating process.

four subcategories of distribution shifts as follows [439]:

- Covariate shift: $P_1(x) \neq P_2(x)$ but $P_1(y|x) = P_2(y|x)$
- Prior probability shift: $P_1(y) \neq P_2(y)$ but $P_1(y|x) = P_2(y|x)$
- Concept drift: $P_1(y|x) \neq P_2(y|x)$ but $P_1(x) = P_2(x)$
- Other distribution shift: $P_1(y, x) \neq P_2(y, x)$ where none of the above three shifts applies.

In theory (and with synthetic data), these four subcategories of distribution shift can be disentangled. However, production data, especially in human-centered applications, is subject to changing distributions that are most often characterized by the catch-all "Other distribution shift" sub-category. As such, the concept of distribution shift is a useful abstraction for understanding why machine learning models trained on one dataset may not generalize to the next but distribution shift is not sufficient for addressing a model's generalizability. Ultimately, distribution shift is downstream of the data generating process, which needs to be taken into consideration to address model robustness.

Machine learning researchers have long combined data-centric and model-centric perspectives in applications of machine learning (e.g. examining the hidden contexts that drive distribution shift [637]), but there is no clear terminology for referring to changes in the semantically meaningful features that influence data-generating processes. As such, we introduce "context shift" to refer to changes in the semantically meaningful features that influence data-generating processes. In order to address context shift and how it may affect a model's generalizability, researchers must begin to identify the dimensions that guide the creation, collection, and curation of data.

Instead of distribution shift, which focuses on the differences in two distributions without regard for the reasons behind the difference, context shift focuses on the dimensions that drive differences in distributions. We can identify these dimensions by looking at sample selection bias (e.g. the new dataset contains images of people from a demographic not represented in the old dataset), adversarial perturbations (e.g. the new dataset contains noise injections that are imperceptible to human perception but change model performance), or non-stationarity (e.g. the new dataset contains images of smart phones post 2018 but the old dataset only contains flip phones before 2010). While we list non-stationarity separately from sample selection bias, non-stationarity can be considered as a special case of sample selection bias where sample selection bias arises from the inability to sample from features and labels in the future. We present Figure 9-3 to illustrate sample selection bias and adversarial perturbations, which can be formally described as follows:

- Sample selection bias: $P_1(s) \not\subset P_2(s)$ where s indicates x, y, or $y|x$
- Adversarial perturbations: $P_1(x) \neq P_2(x)$ but $P_1(y|H(x)) = P_2(y|H(x))$ where $H(x)$ represents human perception of the data

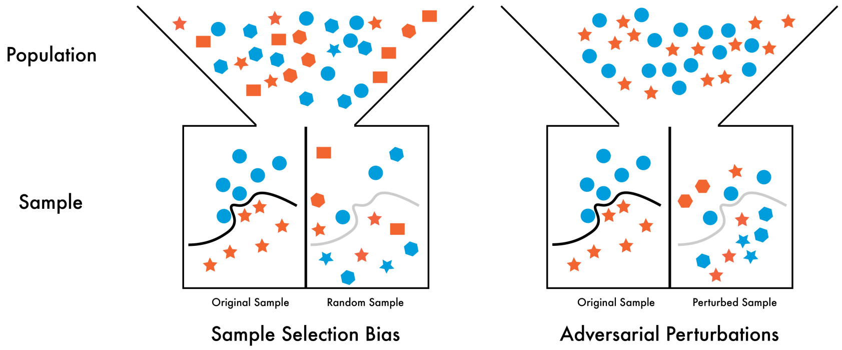


Figure 9-3: Illustrations of sample selection bias and adversarial perturbations with colors representing the ground truth label, geometric shapes and spatial positions representing the features, the top of the funnel representing the full populations, the bottom of the funnel representing the samples drawn from the population, and the solid boundary line representing the learned representation of labels from the original sample. On the left, the population contains upright stars, rotated stars, hexagons, rectangles, and circles, but the biased original sample only contains circles and stars. The random sample contains much higher diversity of features and relationships between features and labels. As such, the learned representation fails in more than 50% of observations. On the right, the population contains upright stars and blue circles. The original sample contains the same set of features, but the perturbed sample includes both rotated hexagons and stars, which may not be immediately noticeable to humans at first glance. Depending on the rotation, the learned representation misclassifies the perturbed shapes. Both pairs of samples present changes in features and changes in labels conditional on the features, which would make these examples of “Other Distribution Shift.” This figure is intended to provide intuition for where the perspective of distribution shift is inadequate and where the perspective of identifying semantically meaningful features that influence how samples are curated and created may inform approaches for addressing robustness in applications of machine learning.

Unlike distribution shift, which can be measured between two datasets, context shift can only be fully addressed by learning the entire population's data distribution, the kinds of changes that are and are not perceptible to humans, and how the population's data distribution changes over time and space. Outside of artificially constrained spaces like synthetic datasets or games, access to the entire populations data distribution (or the rules governing the distribution) across space and time is rare. Nevertheless, people generally have intuition and the ability to reason about data distributions of combinatoric contexts that they might never experience. In fact, cognitive science research shows that intuitive reasoning about statistical distributions (e.g. statistical power analysis [490]) begins early in childhood.

By addressing the benchmark-production gap problem from the data-centric perspective of context shift as opposed to distribution shift, we can consider three approaches for increasing generalizaibility: human intuition and subject matter expertise in machine learning model development, dynamic benchmarking in the evaluation of machine learning models, and limitations statements that clarify how a machine learning model will generalize.

9.3 Addressing Robustness with Human Intuition and Expertise

Over the last few years, researchers have been developing data-centered frameworks to offer guidance for breaking down the data generating process into relevant component parts that reveal where context shift may lead to benchmark-production performance gaps. These frameworks include *Data Statements for Natural Language Processing* [58], *The Dataset Nutrition Label* [279], *Model Cards for Model Reporting* [433], *Datasheets for Datasets* [222], *Closing the AI accountability gap* [517], *The Ethical Pipeline for Healthcare Model Development* [114], *The Clinician and Dataset Shift in Artificial Intelligence* [203], and *Interactive Model Cards* [134]. Likewise, meta-frameworks offer guidance for ensuring data documentation frameworks are useful and actionable [274].

As a heuristic for human-centered machine learning applications, teams of conscientious, creative, and skilled model developers, data engineers, and subject matter experts may find it useful to identify a first-order, non-exhaustive list of dimensions on which context shift is likely to occur. This list of dimensions depends largely on the context and the degree to which the data are subjective, representative, and missing [443]. In ethnographic interviews with machine learning engineers, researchers find that engineers often address changes in context with "elaborate rule-based guardrails to avoid incorrect outputs" [562]. Recent examples of semantically meaningful dimensions that have been demonstrated as useful for evaluating robustness in applied machine learning include skin color in face recognition [87] and dermatology diagnosis [146, 248], gender in facial attribute recognition [631], intersectionality in human-centered applications [628], background scenery for affect recognition [350], number of people in a video for deepfake detection [243], number of chronic illnesses for algorithmic healthcare risk prediction [470], data artifacts like surgical markings [643] or clinically irrelevant labels [468] for medical diagnosis classification, patients' self reports of pain for quantifying severity of knee osteoarthritis [503], and image similarity characteristics for pathologists to disambiguate between machine learning and user errors [93].

Knowledge elicitation for identifying semantically meaningful features is not a solved problem, but helpful questions that may guide the identification of potential context shifts in complex, human-centered machine learning applications include (and are not limited to): who are represented in the data and as annotators of the data, when and where is the data collected, how do social, geographical, temporal, technological, aesthetic, financial incentives and other idiosyncrasies influence the creation of the data, and why the data is curated as it is. Knowledge elicitation has been historically ill-defined in artificial intelligence applications [212], and recent work developing taxonomies for knowledge elicitation helps to formalize the process and increase transparency along the way [117, 324]. In a specific example of machine learning applied to radiology, Lebovitz et al 2022 reveals how subject matter experts (radiologists) currently evaluate machine learning based decision support tools not based on the tools' "know-what" (accuracy on a holdout set) but the tools' "know-how" (qualitative performance as judged by reaching reasonable level of certainty in situated contexts relative to professional standards) [368]. By asking experts to evaluate model performance, researchers can begin to fill the evaluation gaps in machine learning practice [286] that have emerged due to the machine learning field's focus on accuracy on benchmark datasets [68].

Another expert intuition guided approach to closing the benchmark-production gap involves developing test benchmarks with adequate diversity in the data along the contextual dimensions upon which human intuition and expertise suggests model performance is most likely to vary. Recent examples of benchmark datasets working towards this goal are *BREEDS: Benchmarks for Subpopulation Shift* [551] and *WILDS: A Benchmark of in-the-Wild Distribution Shifts* [343], which includes labels for relative contexts and sub-populations for the explicit examination of context shifts.

9.4 Addressing Robustness with Dynamic Benchmarking

A second approach to addressing the benchmark-production gap is to transform the practice of evaluation from static benchmarks to dynamic benchmarks where models' performance is not evaluated on a single dataset, but rather continually evaluated on datasets produced via well-specified, quality controlled data generation processes. Examples of this dynamic benchmarking include "Beat the Machine" [40] (designed for any prediction tasks and evaluated on specific tasks including detecting hate speech and adult content) and dynabench [328] (designed for evaluating natural language processing tasks). For general development of dynamic benchmarks, data generation process desiderata should include specifying the following dimensions of a dynamic benchmark:

- **Prediction task**: What are the input features and output labels? For example, inputs may be images and outputs may be lists of objects or inputs may be described more specifically as images of skin lesions photographed by dermatoscopes and outputs may be classifications of benign and malignant by board-certified dermatologists in the United States. It is important to be careful that the task matches the expected goal because unexpected mismatches between tasks and goals are relatively common [323, 444].

- **Ground truth annotation arbitration**: Who has the authority to annotate the data? How do experts differ from crowdworkers or an algorithm [246]? How should the data be annotated? How should inter-annotator disagreement be represented? What categories should be included?

- **Data inclusion and exclusion criteria**: What are the possible data sources? How are data curated from these sources? What is the data distribution of categories and subcategories? What are the quality constraints?

- **Benchmark size and shape**: What is the minimum size of a batch of data to serve as a benchmark? How should benchmarks by different groups for the same task be combined together?

These desiderata enable the development of dynamic benchmarks that further enable quantitative evaluation of model robustness via corroborated accuracy, which is the distribution of accuracy scores across dynamic benchmarks. Rather than simply evaluating a model on a single or a few static test benchmarks, we might consider a well-corroborated model to be one that meets two criteria: first, it is reasonably available for evaluation, and second, all attempts to uncover systematic errors in well-specified contexts reveal no significant accuracy disparities. The practice of dynamic benchmarking could be particularly relevant for addressing the *AI Knowledge Gap* [186] characterized by the disparity between the large number of machine learning models and the small number of studies evaluating these models' performance. Furthermore, dynamic benchmarking can be combined with benchmark task misalignment methodologies [291, 609] to assess how aligned (or misaligned) model predictions are with human annotations and considering diverse examples that bring transparency to the ethical implications and societal impact of model development [488].

The transition from static benchmarks on a particular instance (or set of instances) to dynamic benchmarks on data generation processes defined by explicit desiderata may be useful for addressing the fundamental issue of construct validity that arises in singular, static benchmarks [517].

9.5 Addressing Robustness by Clarifying a Model's Limitations

A third approach to reducing the benchmark-production gap is to appropriately specify the contexts in which a model is expected to work via a limitations section [577].

To clarify domain-specific limitations driving the benchmark-production gap, we consider implicit assumptions that lead to a context shift in three real-world computer vision tasks: facial expression recognition, deepfake detection, and medical diagnosis.

9.6 Case Studies for Addressing Context Shift in Applied Machine Learning

9.6.1 Facial Expression Recognition

In the field of affective computing, facial expression recognition (FER) is a task to classify human facial expressions with affective labels [126, 382], which can be a useful component in designing human-AI interactions with computational empathy [245, 479, 500]. Model-based FER is similar to how humans recognize the emotions of others (called empathic accuracy in affective science [289] and emotion reasoning in developmental psychology [538]) except that FER is based solely on facial expressions, whereas affect recognition can include information about someone's gestures, language, tone, physiological measurements, and the long-tail of context, which can include factors such as the temperature outside, the social relationship between two individuals, what happened the day before, and more.

Consider an example from relatively recent research [436] where a standard neural network architecture, AlexNet [354], is trained on a large number of images of spontaneous and posed facial expressions to classify images into seven categories (anger, disgust, fear, happiness, sadness, surprise, and neutral) and achieves accuracy scores ranging from 48.6% in SFEW [160] to 56.0% in MMI [482] to 56.1% in DISFA [410] to 61.1% in FER2013 [233] to 77.4% in FERA [44, 616] to 92.2% in CK + [393] to 94.8% in MultiPie [249]. While this model's accuracy is significantly better than random guessing, which would be 14.2%, it varies dramatically depending on the chosen benchmark dataset. How should we interpret a performance gain of 21.9 percentage points on one dataset and an average performance gain of 3.5 percentage points on the other 6 datasets in an alternative network architecture? How should we interpret the model's ability to achieve higher accuracy scores than non-neural network methods on three of the seven benchmark datasets? What does the distribution of performance tell us about how this model would perform on real-world production data? There is no clear answer to any of these questions, yet an implicit assumption in the well-cited, peer-reviewed publication of this FER paper is the slightly improved performance on several benchmark datasets appears to mark a contribution to the field of facial expression recognition. This assumption has the potential to lead to another more pernicious and mistaken assumption: the role of contextual features for real-world performance can be ignored when assessing the state-of-the-art methodology in applied problems like FER.

Clearly, models can learn facial expression features that map to human annotations for a handful of emotion categories to classify images at significantly better than chance rates. But, it is not reported nor clear how changes in lighting, head pose, occlusion, skin tone, ethnicity, age, gender, and background scenery influence both the model's performance or human annotations. It is also underexplored how well FER models would perform if humans of diverse cultures annotated these images. Likewise, it is unclear how the model would perform on more fine-grained emotion categories [132] or labels based on affective dimensions like valence, arousal, and dominance. Furthermore, in many real-world settings where people may feign smiles to appease their managers, cry to express joy, or appear neutral to hide a winning poker hand, the perspective of outside observers may be very different than the perspective of close friends or individuals themselves. While these are not an exhaustive list of contextual features, these represent intuitive, first-order contexts for conducting algorithmic audits, developing future benchmark datasets with these labeled contexts, and adapting models to handle these dimensions. In a recent study, researchers show that including workplace activity and context data improves alignment of FER with self-reported emotions [320]. While researchers build the next version of contextualized dynamic benchmarks, other researchers who are focused on developing models should at the very least include caveats in their papers about the likely contextual dimensions that may affect performance.

9.6.2 Deepfake Detection

As a second case study of context shift in real-world applications of computer vision, we consider deepfake detection. Deepfakes are videos that have been manipulated to make someone appear to do or say something they have not said [75]. These types of manipulation can be qualitatively characterized as face swapping where two people's faces are swapped, head puppetry where facial landmarks are adjusted to make someone appear to be speaking, and lip-syncing where an individual's lips are moved in sync with the phonemes from an external audio track [397].

The largest deepfake detection benchmark dataset to date is the Deepfake Detection Competition Dataset (DFDC) [166, 168], which consists of 128,154 videos based on performances by 960 consenting actors representing diversity across sex and ethnicity. However, Groh et al 2022 point out,"Unlike viral deepfake videos of politicians and other famous people, the videos from [this benchmark dataset] have minimal context: These are all 10 [second] videos depicting unknown actors making uncontroversial statements in nondescript locations" [243]. This deepfake test benchmark is designed to evaluate algorithmic performance in identifying videos that have (and have not) been manipulated by seven synthetic techniques.

But, the real-world deepfake detection problem is not simply identifying whether one of seven synthetic techniques has been applied to a video. Instead, the real-world problem is identifying videos that have been algorithmically altered to impersonate innocent people and deceive the viewer. This problem is more than just a computer vision problem; it is a deception detection problem that involves both searching for artifacts that reveal that a manipulation has occurred and applying prior knowledge and critical reasoning to assess the likelihood that the video has been fabricated.

The DFDC does not include politicians or any scenes of news conferences or people speaking to a large audience. If we assume that harmful deepfakes will involve these kinds of contexts (like a deepfake of President Volodomyr Zelensky that appeared in March 2022 [72]), then it is important to evaluate models on videos with these kinds of dimensions, such as those from the Presidential Deepfakes Dataset [247, 550] and the Protecting World Leaders against Deepfakes Dataset [13]. When Groh et al 2022 examined the leading state-of-the-art for detecting DFDC videos on deepfakes of Kim Jung-un and Vladimir Putin, they found the the leading model predicted a 2% and 8% likelihood these videos are deepfakes. While failure on two examples is only an anecdote, this failure speaks to an important need: diverse test benchmarks that cover the first-order dimensions where human intuition and expertise suggests context shift is most likely to occur.

9.6.3 Medical Diagnosis

As a third case study of context shift, we consider medical diagnosis in store-and-forward teledermatology settings where clinical data are collected at one site and sent electronically for evaluation at another site. Recent research on machine learning applied to skin disease classification has demonstrated the human expert-level performance of models in a number of specific tasks [187, 388]. However, it is unclear how these models will perform in production especially on people with dark skin because the first paper does not describe the distribution of ethnicity or skin tone in the evaluation benchmark [187] and the evaluation benchmark in the second paper contains only 2.7% of people with the second darkest of the six Fitzpatrick Skin Types (FSTs) and 1 person with the darkest of the FSTs [388]. Given the accuracy disparities that appeared across skin types in facial recognition, expert intuition suggests that systematic errors are likely to also appear in skin disease classifiers.

In fact, empirical research corroborates this intuition [248], and the Diverse Dermatology Images (DDI) dataset [146] reveals that state-of-the-art skin disease classification models make systematically more errors on dark skin than on light skin. The DDI represents a more comprehensive benchmark than previous datasets, and as a result, the DDI exposed errors that should guide and motivate the future development of machine learning models towards more robustness. However, the DDI is not perfectly comprehensive; the dataset is de-identified for privacy reasons and lacks free text clinical notes and other information that physicians would acquire via an in-person examination [146]. Given that many skin diseases appear similarly and expert diagnoses are based on clinical history and non-visual features, expert intuition would expect, once again, that systematic errors lurk in the state-of-the-art machine learning models for store-and-forward skin disease classification.

9.7 Towards Robustness in Applied Machine Learning

Supervised machine learning models are very good at identifying statistical regularities in a given dataset but tend to err on out-of-distribution data that may arise from sample selection bias, adversarial perturbation, or nonstationarity. On the other hand, humans can be quite good at identifying contextual examples of out-of-distribution data. By combining the strengths of machine learning models with human intuition and expertise, early career ancient historians can quickly restore and date ancient texts [37], content moderation teams can more accurately distinguish between real and fake videos [243], and general practitioners can more accurately diagnose skin conditions from images [300] (although AI advice can also mislead experts; see [5, 221, 299, 608, 615]). In fact, initial evidence suggests that human intuition is fairly accurate in predicting model misclassifications on common object detection tasks [667]. The integration of machine predictions with human decisions in collaborative decision making systems may be the most immediately effective way to avoid errors from context shift. The three case studies suggest the following advice for applied machine learning researchers:

- **Human intuition and subject matter expertise** can be useful for identifying first-order dimensions where context shift is likely to occur. These dimensions can inform the write-up of a limitations section, the development of a test benchmark, the collection of new data, or changes to model architecture.
- The practice of **dynamic benchmarking** mirrors the real-world more closely than static benchmarking and can enable insights from anywhere into systematic model failures.
- The inclusion of **limitations statements** in peer-reviewed research can increase model generalizability by simply clarifying the contexts in which a model is expected to generalize or not.

Promising future research directions for developing robust machine learning models under distribution shift involve the following iterative process: first, identify missing contexts in test benchmarks, second, collect data that contain those missing contexts, and third, adjust the model accordingly. Researchers can begin to identify missing contexts by collaborating with human experts who may be able to identify first-order drivers of context shift on a task-by-task basis. Similarly, researchers can further identify missing contexts by evaluating models against data generation process desiderata rather than a single or a few datasets.

Finally, one of the most effective solutions for addressing the benchmark-production gap is for researchers to clearly communicate the contexts in which a model has been evaluated and the contexts in which the model's performance is unknown.

Acknowledgments

I thank Katie Matton, Noah Jones, Robert Lewis, Rosalind Picard, Christelle Godin, Marion Mainsant, Adrien Viault, and Jessica Hullman for feedback on early drafts of this paper, and I thank Zach Lipton and other participants at the ICML Principles of Distribution Shift workshop for additional insightful feedback.

Chapter 10

Conclusion

Overall, this dissertation examines *human-AI collaboration* across a variety of domains and tasks via (1) large-scale digital experiments (Chapters 2, 3, 6, and 7), (2) algorithmic audits (Chapters 4 and 5), and (3) proposed frameworks for addressing context in applied machine learning (Chapters 8 and 9). The science of human-AI collaboration is concerned with the rigorous evaluation of human, machine, and hybrid problem solving approaches and the art of human-AI collaboration is concerned with the design of hybrid systems, intuition for identifying systematic errors of humans and machines, and consideration to which contexts a system (human, machine, or hybrid) should attend. The goal of this dissertation has been to build an understanding of augmented intelligence, communicate this understanding as explicitly as possible, and address real-world problems in all their complexity along the way.

Chapter 2 introduces an experimental paradigm for comparing human performance (individual accuracy and crowd wisdom), machine performance, and AI-assisted human performance in deepfake detection. In the highly focused task of distinguishing whether a minimal context video has been manipulated by a deepfake algorithm or not, we find crowd wisdom is comparable to the leading algorithm. Moreover, we find that AI-assisted human performance is more accurate than the performance of humans or the leading algorithm alone, but we find this overall performance comes with a major caveat: AI-assisted human performance is significantly worse than human performance without AI assistance when the leading algorithm is wrong or equivocal. We demonstrate that **algorithm anchoring** – the tendency of humans to update towards algorithms' predictions whether the predictions are correct or not – extends to instances where the algorithm makes relatively surprising and systematic errors (e.g. the canonical deepfake videos of Kim Jung-un and Vladimir Putin). Notably, the leading algorithm is only trained to detect manipulations to the pixels in videos whereas humans could identify these political leader deepfakes by any number of contextual clues. In order to further understand how people use context to distinguish between authentic and fabricated media, Chapter 3 examines how people incorporate the visual-audio cues versus the content of what is said in authentic and fabricated political speeches by US presidents. It is important to note that in these deepfake experiments, participants are generally not experts and are far from certain about what is a deepfake and what is not, which may leave them to be more susceptible to algorithm anchoring than experts would be.

Chapter 6 adapts and extends the experimental paradigm of Chapter 2 for store-and-forward teledermatology diagnosis. In the diagnostic accuracy task, we find that generalists are much more likely to accept algorithmic advice than specialists, which provides further evidence that algorithm anchoring is moderated by human expertise. However, in contrast to human performance with AI-assistance in the deepfake detection task in Chapter 3, Chapter 6 demonstrates that the physicians rarely override a correct diagnosis with an incorrect suggestion by the AI support system. One potential explanation for resilience to algorithmic anchoring is the positive asymmetry (physicians including correct suggestions and ignoring incorrect suggestions) is AI diagnostic assistance in dermatology encourages mental exploration: there are 1,000s of dermatological conditions, which are not necessarily the first condition to come to a physician's mind but relatively easy for a physician to rule out if the image does not look like the suggested condition. One potentially surprising finding in Chapter 6 is the low performance of specialists (specialists are nearly twice as accurate as generalists but only 38% accurate overall), which reveals the importance of contextual information for dermatology diagnosis. In particular, different skin diseases can appear visually similar and a physician would distinguish between these diseases with more in-depth visual information, non-visual information like a patient's clinical history or recent activities, or how the condition plays out over times and in response to different interventions. As practitioners begin to design human-AI collaborations in the real-world, it becomes very important to help physicians (or whomever the AI assistance is designed for) recognize what information the algorithm has access, so they can build a mental model for how they expect an algorithm to perform and identify the contexts where the algorithm is susceptible to systematic failures.

Chapters 4 and 5 reveal where systematic failures of algorithms lurk and present insights for promoting transparency in machine learning applied to clinical dermatology. In an effort to reduce systematic errors in applied machine learning, Chapter 8 presents a framework for identifying relevant contexts in affect recognition tasks and Chapter 9 discusses the importance of human intuition, expert knowledge, dynamic benchmarking, and clearly communicating a model's capabilities.

Drawing on insights from affective computing, chapter 7 examines how AI assistance could augment human performance without providing any strategic information and only offering empathetic support. In Chapter 7, we enumerated empathetic responses to players' potential actions in Wordle and evaluated players' performance objectively based on their win-rate and entropy reduction of their guesses, which allowed us to evaluate the effectiveness of the empathetic responses. We find that the empathetic agent counteracts the negative effects of anger on human problem solving, which offers initial evidence that digitally mediated expressions of empathy can objectively improve human problem solving in a single task. The natural follow-up question is to which contexts does this generalize; when, where, how, and why does digitally mediated expressions of empathy influence human problem solving?

10.1 Future Work

This dissertation pulls back the curtains on the complexities of building and evaluating hybrid human-AI decision making systems for addressing real-world problems. With the insights (and proverbial curtains!) drawn from this dissertation, the stage of human-AI collaboration is set for building principles of human-AI collaboration via rigorously examining performance in "model organisms" and carefully considering how context influences performance.

10.1.1 Robustness in Applied Machine Learning

In the field of machine learning, the current paradigmatic question for addressing robustness and generalization can be paraphrased as followed: "How can we design machine learning models that are (a) robust to adversarial attacks, (b) adverse to distribution shifts, (c) robust to dataset biases, and (d) interpretable and transparent?" In Chapter 9, I propose a paradigm shift that would integrate human capacity in applied machine learning such that context is incorporated into the common task framework of evaluating performance on benchmarks. Future work may explore how to create systematic, dynamic benchmark tasks where the data is left unspecified but the data generating process is precisely specified. This shift is particularly timely because many models often appear to be performing at human-level when evaluating on select, static benchmarks but countless examples reveal that models fail in ways that reasonable humans would not.

10.1.2 Designing Algorithmic Assistance Interfaces

In Chapter 6, we present evidence that the order in which choices to include or ignore algorithmic advice are presented to humans influences what humans choose. What are the principles that guide the design of interfaces for algorithmic assistance? Future research questions abound in how interface designs will influence human-AI collaboration. For example, how should designers select the examples in algorithmic assistance tutorials to help users build mental models of machine performance? When should AI assistance be shown before human input and when should AI assistance be reserved for after a human finalizes their initial judgment? Should AI assistance ever create friction to force a human to deliberate longer, and if so when? How should an interface effectively communicate algorithmic uncertainty and data attributes like quality, relevance, similarity to past examples, and out-of-distribution status? What information should the algorithmic assistance provide? For example, in store-and-forward dermatology, models are often trained to predict the diagnosis but a potentially more effective model and interface might provide classifications of an image by its visual features e.g. heliotrope signs, gottron's papules, bullseye rashess, and other known visual signatures of skin disease. While some past research has touched on aspects of each of these questions, open questions remain on what a succinct list of principles looks like and the boundary conditions of when and where these principles apply.

10.1.3 Misinformation and Synthetic Media

How can technology assist people in distinguishing authentic photographs, audio, and video from synthetic media online? In the future, it is possible that synthetic media generated by AI will be perceptually indistinguishable from media recorded by people. In fact, some synthetic media appears hyperrealistic today. In a future where multimedia is no longer considered reasonable evidence of indexicality, we may treat image-based evidence similarly to how we treat something we read. While this future of perceptually indistinguishable media is possible, the near term future is more likely to involve intermittently perceptually indistinguishable media in certain contexts. Future research on deepfake detection and the effects of synthetic media may center around the question of when and how people trust media. For example, how do people calibrate what media to believe? How realistic is synthetic media, and how can we evaluate the photorealistic capacities of an algorithm? What does a framework look like that clarifies the contexts (both within an image and outside an image) in which synthetic media is perceptually indistinguishable from human recorded media and the contexts where synthetic media often leaves perceptual artifacts? What are the tell-tale signs of physical implausibilities and biometrics that can help people identify future synthetic media? How do technologies to track the source and provenance of media influence the spread of synthetic media? How should law and content moderation policies address synthetic media? How will centralized (newspapers, television channels) versus decentralized (social media) media institutions evolve as synthetic media evolves?

10.1.4 Medical Diagnosis and Physician-Machine Partnerships

Future directions in physician-machine partnerships may address the following questions: How do physicians build intuition about clinical decision support tools based on AI assistance over time? How accurate are physician-machine partnerships across diverse characteristics in a longitudinal settings? How do physicians' calibrations of the AI assistance change over time? How does this depend on level of expertise and the presence of algorithmic bias? How would the results in Chapter 6 generalize to an experimental set-up consisting of in-person clinical visits? How would second opinions from specialists or other generalists change the results? How could medical organizations triage patients most effectively based on the dynamics of physician-machine partnerships?

10.1.5 Empathy and Digitally Mediated Expressions of Empathy

The results from Chapter 7 lay the groundwork for a research program on the design and implications of digitally mediated expressions of empathy. Research questions on empathy include: How to systematically evaluate digitally mediated expressions of empathy (e.g. text-based conversations)? How do contexts (e.g. peer support, therapy, customer service, etc.) influence the evaluation of digitally mediated expressions of empathy? How can we guide humans and large language models towards empathetic communication? When is digitally mediated expressions of empathy helpful in real-world applications? How can digitally mediated expressions of empathy go awry? What are the ethics of designing digitally mediated expressions of empathy?

10.1.6 Generative AI

This dissertation focused on AI assistance in the form of discriminative machine learning models and a wizard of oz model for digitally mediated expressions of empathy, and future work may consider generative AI including generative computer vision (e.g. diffusion models, generative adversarial networks, and autoencoders), large language models (LLMs), and text-to-speech. While we examined the detection of deepfakes, the production of deepfakes is one example of human and generative AI collaboration. For example, the Tom Cruise deepfake mentioned in Chapter 3 involves an autoencoder, a visual effects artist, and a look alike actor. The expert visual effects and acting are key components in producing a deepfake that appears hyperrealistic. Future research may consider how to characterize the quality of generative AI and how to characterize the marginal quality improvements in human-generative AI systems. Likewise, future research may consider evaluating the quality of diffusion models across contexts and human-AI collaboration depth. Diffusion models can transform a text prompt into a relatively high quality image in less than a second (though quality is remains imprecisely defined), but high quality images often take minutes to hours to curate as the human iterates through prompts and variations on the same prompt. A rigorous evaluation of the quality of generated images and the time it takes to produce these images can provide insight into the future applications of diffusion models.

Similarly, rigorous evaluations of LLMs and LLM assistance to humans is another opportunity for future work. Where can LLMs be effective in coaching humans? Where and why do LLMs fail in improving human performance? How do statements produced by LLMs influence human beliefs? How do LLMs transformed into perceptually believable speech by text-to-speech algorithms influence human beliefs? And, how can we design benchmarks that demonstrate LLMs' capacity (or lack thereof) to solve general problems rather than showcase performance that may be contaminated by memorizing the answers?

10.2 Parting Showcase of Human-AI Collaboration by Rewriting the Abstract in Iambic Pentameter

As a parting showcase of human-AI collaboration in writing verse, I prompted an LLM to rewrite the abstract in iambic pentameter and copied the results below on the left. Technically, iambic pentameter has ten syllables per line, and the majority of lines in the prompt generated abstract do not have ten syllables. Despite the wrong number of syllables for most lines, the generated verse still feels a bit like verse because the rhyming scheme resembles the heroic couplets that often appear in iambic pentameter verse. In the right column, as a showcase of the iterative potential of human-LLM collaboration, I edited the LLM-produced abstract verse to strictly adhere to iambic pentameter and evoke more precise meaning.

Prompt Generated Abstract

Though AI seems to outperform
Expert humans in tasks the norm
When deployed, it sometimes fails
And leaves us with unexpected tales

Humans, on the other hand
Adapt to new contexts with ease, unplanned
Their strengths and flaws in contrast to AI
Motivate research question: where and why

Can human-AI hybrid systems lead
To surpassing performance, to succeed
This dissertation seeks to answer how
And when we should rely on the now

Experiments in three domains are run
To shed light on the question, one by one
Deepfake detection, dermatology diagnosis
And Wordle, creative problem solving for theosis

In deepfake detection, we explore
Similarities, differences, performance galore
Between human and machine vision, trade-offs we find
In hybrid systems, AI and humans combined

Dermatology diagnosis, non-visual info in demand
Disparities across skin colors, we understand
Decision support from a fair deep learning system
Can increase accuracy and reduce disparities, an important wisdom

Wordle shows us how digitally mediated empathy
Counters the negative effects of anger, a key
Algorithmic audits on clinical dermatology images too
Reveal systematic errors, and how context influences what's true

These contributions, empirical evidence in hand
Guide us in designing human-AI collaborations more grand
And provide a framework for when and where
Hybrid systems should rely on human problem-solving care.

Prompt Scaffolded Abstract

Although AI appears to outperform
Expert humans in games and tasks the norm,
When deployed, the algorithm often fails
And leaves us with – oops – unexpected tales!

Humans, sweet earthlings, on the other hand
Adapt to new contexts with ease, unplanned.
Their strengths and flaws in contrast to AI
Motivate research questions: where and why

Can human-AI hybrid systems lead
To surpassing performance, to succeed?
This dissertation seeks to answer how
And when to rely on or disavow.

Experiments in three domains are run
To shed light on the question, one by one:
Deepfake detection, derm diagnosis
And Wordle, solve for apotheosis!

In deepfake detection, we do explore
Similarities, diffs, results galore
Between human, machine, hybrid combined
Systems involve real tough trade-offs we find.

In derm, non-viz info is in demand;
Disparities put on the witness stand;
Fair clinical diagnostic support
Builds accuracy and doctor rapport.

When it comes to games, we find empathy
Counters the effects of anger, a key;
Algo audits on derm images too;
Plus how context influences what's true.

These contributions, evidence in hand,
Guide us in designing collabs more grand
And provide a framework for when and where
Hybrid systems should rely on human care.

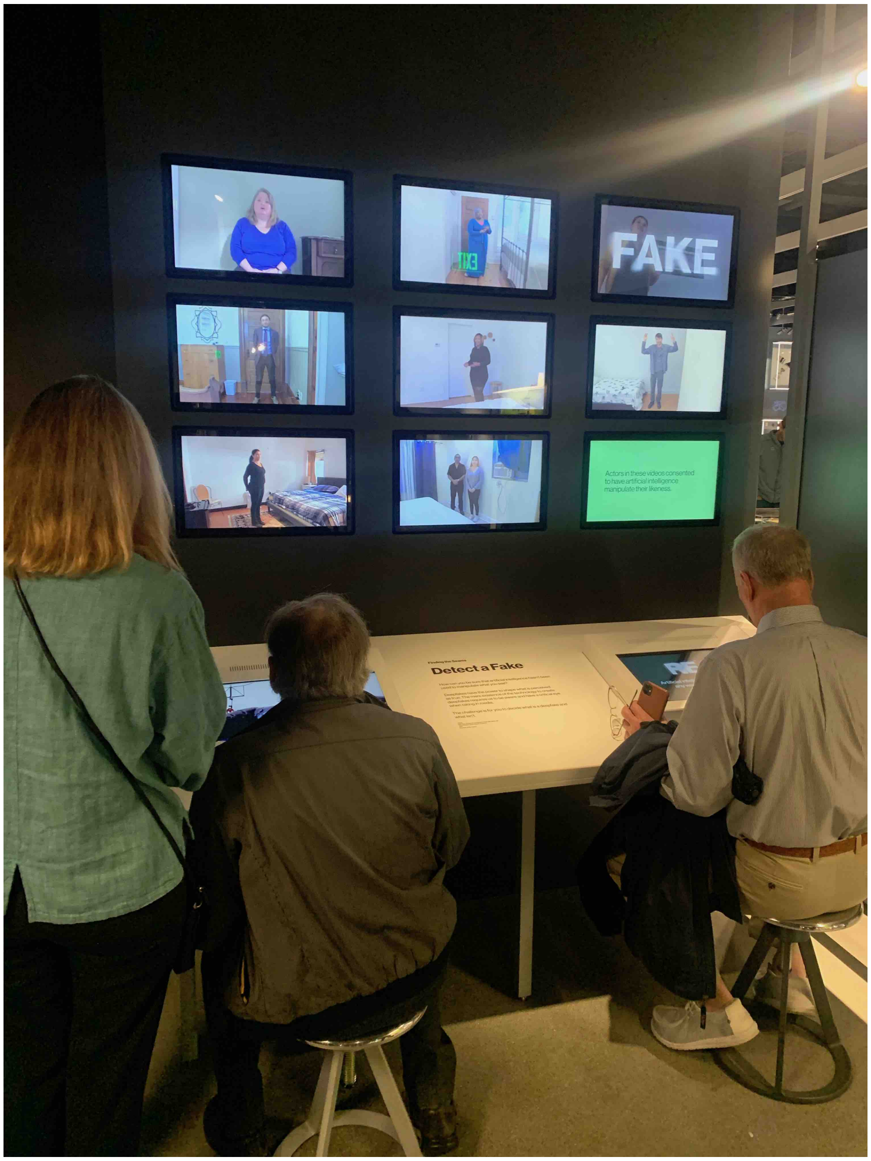

FAKE
Actors in these videos consented to have artificial intelligence manipulate their likeness.
Detect a Fake

Figure 10-1: Photo of extended family visiting the "Detect a Fake" exhibit at the MIT Museum, which is based on the experiments in chapter 2.

Chapter 11

Citations

Bibliography

1. (FAIR)†, M. F. A. R. D. T. *et al.* Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* **378,** 1067–1074 (2022).
2. Abadie, A., Athey, S., Imbens, G. W. & Wooldridge, J. *When should you adjust standard errors for clustering?* en. Tech. rep. arXiv: 1710.02926 (National Bureau of Economic Research, Oct. 2017). `http://arxiv.org/abs/1710.02926` (2021).
3. Abbasi-Sureshjani, S., Raumanns, R., Michels, B. E. J., Schouten, G. & Cheplygina, V. Risk of Training Diagnostic Algorithms on Data with Demographic Bias. en. *arXiv:2005.10050 [cs, stat].* arXiv: 2005.10050. `http://arxiv.org/abs/2005.10050` (2020) (June 2020).
4. Abebe, R. *et al. Roles for computing in social change* en. in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Acm, Barcelona Spain, Jan. 2020), 252–260. ISBN: 978-1-4503-6936-7. `https://dl.acm.org/doi/10.1145/3351095.3372871` (2021).
5. Abeliuk, A., Benjamin, D. M., Morstatter, F. & Galstyan, A. Quantifying machine influence over human forecasters. en. *Scientific reports* **10,** 1–14. ISSN: 2045-2322. `http://www.nature.com/articles/s41598-020-72690-4` (2021) (Dec. 2020).
6. Abramson, L., Petranker, R., Marom, I. & Aviezer, H. Social interaction context shapes emotion recognition through body language, not facial expressions. *Emotion* (2020).
7. Adamson, A. S. & Smith, A. Machine Learning and Health Care Disparities in Dermatology. en. *JAMA Dermatology* **154,** 1247. ISSN: 2168-6068. `http://archderm.jamanetwork.com/article.aspx?doi=10.1001/jamadermatol.2018.2348` (2020) (Nov. 2018).
8. Adebayo, J., Muelly, M., Abelson, H. & Kim, B. *Post hoc explanations may be ineffective for detecting unknown spurious correlation* in *International Conference on Learning Representations* (2021).
9. Adelekun, A., Onyekaba, G. & Lipoff, J. B. Skin color in dermatology textbooks: an updated evaluation and analysis. *Journal of the American Academy of Dermatology* **84,** 194–196 (2021).
10. Agarwal, S. & Farid, H. *Detecting Deep-Fake Videos From Aural and Oral Dynamics* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 981–989.

11. Agarwal, S., Farid, H., Fried, O. & Agrawala, M. *Detecting deep-fake videos from phoneme-viseme mismatches* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), 660–661.

12. Agarwal, S., El-Gaaly, T., Farid, H. & Lim, S.-N. *Detecting Deep-Fake Videos from Appearance and Behavior* en. arXiv: 2004.14491. Apr. 2020. `http://arxiv.org/abs/2004.14491` (2020).

13. Agarwal, S. *et al. Protecting World Leaders Against Deep Fakes.* en. in *CVPR Workshops* **1** (2019), 38–45.

14. Agarwal, S. *et al.* Watch Those Words: Video Falsification Detection Using Word-Conditioned Facial Motion. en. *arXiv:2112.10936 [cs].* arXiv: 2112.10936. `http://arxiv.org/abs/2112.10936` (2022) (Dec. 2021).

15. Ali, S., Devasia, N. E. & Breazeal, C. *Escape! Bot: Social Robots as Creative Problem-Solving Partners* in *Creativity and Cognition* (2022).

16. AlKattash, J. A. DermaAmin. *https://www.dermaamin.com/site/.*

17. Alkhatib, A. & Bernstein, M. *Street-level algorithms: A theory at the gaps between policy and decisions* en. in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Acm, Glasgow Scotland Uk, May 2019), 1–13. ISBN: 978-1-4503-5970-2. `https://dl.acm.org/doi/10.1145/3290605.3300760` (2021).

18. Allen, J., Arechar, A. A., Pennycook, G. & Rand, D. G. *Scaling up fact-checking using the wisdom of crowds* en. 2020.

19. Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. *Evaluating the fake news problem at the scale of the information ecosystem* 2020.

20. Almaatouq, A. *et al.* Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences. *Behavioral and Brain Sciences,* 1–55 (2022).

21. Alsan, M., Garrick, O. & Graziani, G. Does diversity matter for health? Experimental evidence from Oakland. *American Economic Review* **109,** 4071–4111 (2019).

22. *Altmeyers Enzyklopädie - Fachbereich Dermatologie — altmeyers.org* `https://www.altmeyers.org/de/dermatologie`. [Accessed 17-Feb-2023].

23. Alvarado, S. M. & Feng, H. Representation of dark skin images of common dermatologic conditions in educational resources: a cross-sectional analysis. en. *Journal of the American Academy of Dermatology,* S0190962220311385. ISSN: 01909622. `https://linkinghub.elsevier.com/retrieve/pii/S0190962220311385` (2020) (June 2020).

24. Amodei, D. *et al.* Concrete Problems in AI Safety. en. *arXiv:1606.06565 [cs].* arXiv: 1606.06565. `http://arxiv.org/abs/1606.06565` (2022) (July 2016).

25. André, P., Kittur, A. & Dow, S. P. *Crowd synthesis: Extracting categories and clusters from complex data* in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (2014), 989–998.

26. Appel, M. & Prietzel, F. The detection of political deepfakes. *Journal of Computer-Mediated Communication* **27,** zmac008 (2022).

27. Appiah, O. Rich Media, poor Media: The Impact of audio/video vs. text/picture testimonial ads on browsers' evaluations of commercial web sites and online products. *Journal of Current Issues & Research in Advertising* **28,** 73–86 (2006).

28. Archer, C. B. *Ethnic dermatology: clinical problems and skin pigmentation* 2008.

29. Arechar, A. A. *et al.* Understanding and Reducing Online Misinformation Across 16 Countries on Six Continents. en. *PsyArXiv.* Accessed 2022-02-22, 50 (2022).

30. Arifin, M. S., Kibria, M. G., Firoze, A., Amini, M. A. & Yan, H. *Dermatological disease diagnosis using color-skin images* in *2012 international conference on machine learning and cybernetics* **5** (2012), 1675–1680.

31. Arik, S. O., Chen, J., Peng, K., Ping, W. & Zhou, Y. Neural voice cloning with a few samples. *arXiv preprint arXiv:1802.06006.* Accessed 2021-07-07 (2018).

32. Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. *Invariant Risk Minimization* en. Number: arXiv:1907.02893 arXiv:1907.02893 [cs, stat]. Mar. 2020. `http://arxiv.org/abs/1907.02893` (2022).

33. Arnold, H. L., Odom, R. B., Andrews, G. C. & James, W. D. *Andrews' diseases of the skin: clinical dermatology* 1990.

34. Aroyo, L. & Welty, C. Truth is a lie: Crowd truth and the seven myths of human annotation. en. *AI Magazine* **36,** 15–24. ISSN: 2371-9621, 0738-4602. `https://ojs.aaai.org/index.php/aimagazine/article/view/2564` (2022) (Mar. 2015).

35. De-Arteaga, M., Fogliato, R. & Chouldechova, A. *A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores* in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–12.

36. Ashby, W. R. *An introduction to cybernetics* (Chapman & Hall Ltd, 1957).

37. Assael, Y. *et al.* Restoring and attributing ancient texts using deep neural networks. en. *Nature* **603,** 280–283. ISSN: 0028-0836, 1476-4687. `https://www.nature.com/articles/s41586-022-04448-z` (2022) (Mar. 2022).

38. Athalye, A., Engstrom, L., Ilyas, A. & Kwok, K. *Synthesizing robust adversarial examples* in *International conference on machine learning* (2018), 284–293.

39. *Atlas of Dermatology — kkh.dk* `https://www.kkh.dk/atlas/index.html`. [Accessed 17-Feb-2023].

40. Attenberg, J. M., Ipeirotis, P. G. & Provost, F. *Beat the machine: Challenging workers to find the unknown unknowns* in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence* (2011).

41. Austin, E. W. & Dong, Q. Source v. content effects on judgments of news believability. *Journalism quarterly* **71,** 973–983 (1994).

42. Aviezer, H., Trope, Y. & Todorov, A. Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions. en. *Science* **338,** 1225–1229. ISSN: 0036-8075, 1095-9203. `https://www.sciencemag.org/lookup/doi/10.1126/science.1224313` (2020) (Nov. 2012).

43. Aviezer, H., Ensenberg, N. & Hassin, R. R. The inherently contextualized nature of facial emotion perception. *Current Opinion in Psychology* **17,** 47–54 (2017).

44. Bänziger, T. & Scherer, K. R. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook* **2010,** 271–94 (2010).

45. Barabas, C., Doyle, C., Rubinovitz, J. & Dinakar, K. *Studying up: reorienting the study of algorithmic fairness around issues of power* en. in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), 167–176.

46. Barari, S., Lucas, C. & Munger, K. *Political Deepfake Videos Misinform the Public, But No More than Other Fake Media* en. preprint. Accessed 2021-01-14 (Open Science Framework, Jan. 2021). `https://osf.io/cdfh3` (2021).

47. Barasch, A., Schroeder, J., Zev Berman, J. & Small, D. Cues to Sincerity: How People Assess and Convey Sincerity in Language. *ACR North American Advances* (2018).

48. Barata, C., Celebi, M. E. & Marques, J. S. Explainable skin lesion diagnosis using taxonomies. *Pattern Recognition* **110,** 107413 (2021).

49. Barocas, S. & Selbst, A. D. Big data's disparate impact. en. *Calif. L. Rev.* **104,** 671. ISSN: 1556-5068. `https://www.ssrn.com/abstract=2477899` (2021) (2016).

50. Baron-Cohen, S. & Wheelwright, S. The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders* (2004).

51. Barratt, D., Rédei, A. C., Innes-Ker, Å. & de Weijer, J. Does the Kuleshov effect really exist? Revisiting a classic film experiment on facial expressions and emotional contexts. *Perception* **45,** 847–874 (2016).

52. Barrett, L. F. Emotions are real. en. *Emotion* **12,** 413–429. ISSN: 1931-1516, 1528-3542. `http://doi.apa.org/getdoi.cfm?doi=10.1037/a0027555` (2020) (2012).

53. Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience* **12,** 1–23 (2017).

54. Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M. & Pollak, S. D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* **20,** 1–68 (1 2019).

55. Barrett, L. F., Mesquita, B. & Gendron, M. Context in emotion perception. *Current Directions in Psychological Science* **20,** 286–290 (2011).

56. Beede, E. *et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy* en. in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Acm, Honolulu HI USA, Apr. 2020), 1–12. ISBN: 978-1-4503-6708-0. `https://dl.acm.org/doi/10.1145/3313831.3376718` (2021).

57. Beeler, M. Information Theory and the Game of JOTTO (1971).

58. Bender, E. M. & Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* **6,** 587–604 (2018).

59. Benjamin, R. Assessing risk, automating racism. en. *Science* **366,** 421–422. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.aaz3873 (2022) (Oct. 2019).

60. Benveniste, A. *The Sudden Rise of Wordle* https://www.nytimes.com/2022/01/31/crosswords/nyt-wordle-purchase.html.

61. Berger, J. & Milkman, K. L. What makes online content viral? *Journal of marketing research* **49,** 192–205 (2012).

62. Berinsky, A. J., Huber, G. A. & Lenz, G. S. Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political analysis* **20,** 351–368 (2012).

63. Berinsky, A. J., Margolis, M. F. & Sances, M. W. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science* **58,** 739–753 (2014).

64. Bertsimas, D. & Paskov, A. An Exact and Interpretable Solution to Wordle. *Available at URL.(Accessed: 14 November 2022)* (2022).

65. Betella, A. & Verschure, P. F. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one* (2016).

66. Bialek, M. & Pennycook, G. The cognitive reflection test is robust to multiple exposures. *Behavior research methods* **50,** 1953–1959 (2018).

67. Bickmore, T. W. & Picard, R. W. *Towards caring machines* in *CHI'04 extended abstracts on Human factors in computing systems* (2004).

68. Birhane, A. *et al. The values encoded in machine learning research* in *2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), 173–184.

69. Bissoto, A., Valle, E. & Avila, S. *Debiasing skin lesion datasets and models? not so fast* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), 740–741.

70. Bissoto, A. *et al.* Deep-learning ensembles for skin-lesion segmentation, analysis, classification: RECOD titans at ISIC challenge 2018. *arXiv preprint arXiv:1808.08480* (2018).

71. Bixler, R. & D'Mello, S. *Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits* in *Proceedings of the 2013 international conference on Intelligent user interfaces* (2013), 225–234.

72. Boháček, M. & Farid, H. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences* **119,** e2216035119 (2022).

73. Bolognia, J. L., Schaffer, J. V. & Cerroni, L. *Dermatologıéa* (Elsevier Health Sciences, 2018).

74. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* **29,** 4349–4357 (2016).

75. Boneh, D., Grotto, A. J., McDaniel, P. & Papernot, N. How relevant is the Turing test in the age of sophisbots? *IEEE Security & Privacy* **17,** 64–71 (2019).

76. Bonferroni, C. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* **8,** 3–62 (1936).

77. Bosch, N. & D'Mello, S. Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing* (2019).

78. Bosch, N. *et al. Automatic detection of learning-centered affective states in the wild* in (2015), 379–388.

79. Bossuyt, P. M. *et al.* STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Clinical chemistry* **61,** 1446–1452 (2015).

80. Bouhuys, A. L., Bloem, G. M. & Groothuis, T. G. G. Induction of depressed and elated mood by music influences the perception of facial emotional expressions in healthy subjects. *Journal of affective disorders* **33,** 215–226 (1995).

81. Brashier, N. M. & Marsh, E. J. Judging truth. *Annual review of psychology* **71,** 499–515 (2020).

82. Brave, S., Nass, C. & Hutchinson, K. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent (2005).

83. Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* **16,** 199–231 (2001).

84. Brown, N. & Sandholm, T. Superhuman AI for multiplayer poker. *Science* **365,** 885–890 (2019).

85. Brown, T. B., Mané, D., Roy, A., Abadi, M. & Gilmer, J. Adversarial patch. *arXiv preprint arXiv:1712.09665* (2017).

86. Bryan, C. J., Tipton, E. & Yeager, D. S. Behavioural science is unlikely to change the world without a heterogeneity revolution. en. *Nature Human Behaviour* **5,** 980–989. ISSN: 2397-3374. `https://www.nature.com/articles/s41562-021-01143-3` (2021) (Aug. 2021).

87. Buolamwini, J. & Gebru, T. *Gender shades: Intersectional accuracy disparities in commercial gender classification* in *Conference on fairness, accountability and transparency* (2018), 77–91.

88. Buslaev, A. *et al.* Albumentations: Fast and Flexible Image Augmentations. en. *Information* **11,** 125. ISSN: 2078-2489. `https://www.mdpi.com/2078-2489/11/2/125` (2020) (Feb. 2020).

89. Buxton, P. & Morris-Jones, R. Eczema (Dermatitis) including management. *ABC of Dermatology,* 24–35 (2009).

90. Cabrera, Á. A., Druck, A. J., Hong, J. I. & Perer, A. Discovering and Validating AI Errors With Crowdsourced Failure Reports. en. *Proceedings of the ACM on Human-Computer Interaction* **5,** 1–22. ISSN: 2573-0142. `https://dl.acm.org/doi/10.1145/3479569` (2021) (Oct. 2021).

91. Cai, C. J., Winter, S., Steiner, D., Wilcox, L. & Terry, M. *Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants* en. in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Acm, Yokohama Japan, May 2021), 1–7. ISBN: 978-1-4503-8095-9. https://dl.acm.org/doi/10.1145/3411763.3443435 (2021).

92. Cai, C. J., Winter, S., Steiner, D., Wilcox, L. & Terry, M. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. en. *Proceedings of the ACM on Human-Computer Interaction* **3,** 1–24. ISSN: 2573-0142. https://dl.acm.org/doi/10.1145/3359206 (2021) (Nov. 2019).

93. Cai, C. J. *et al. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making* en. in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (ACM Press, Glasgow, Scotland Uk, 2019), 1–14. ISBN: 978-1-4503-5970-2. http://dl.acm.org/citation.cfm?doid=3290605.3300234 (2020).

94. Calbi, M. *et al.* How context influences our perception of emotional faces: A behavioral study on the Kuleshov effect. *Frontiers in psychology* **8,** 1684 (2017).

95. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356,** 183–186 (6334 2017).

96. Callen, J. P., Greer, K. E., Hood, A. F., Paller, A. S. & Swinyer, L. Color atlas of dermatology (1993).

97. Calo, R., Coward, C., Spiro, E. S., Starbird, K. & West, J. D. How do you solve a problem like misinformation? *Science advances* **7,** eabn0481 (2021).

98. Calvo, R. A. & D'Mello, S. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. en. *IEEE Transactions on Affective Computing* **1,** 18–37. ISSN: 1949-3045. http://ieeexplore.ieee.org/document/5520655/ (2020) (1 Jan. 2010).

99. Calvo, R. A., D'Mello, S., Gratch, J. M. & Kappas, A. *The Oxford handbook of affective computing* (Oxford University Press, USA, 2015).

100. Campbell, M., Hoane Jr, A. J. & Hsu, F.-h. Deep blue. *Artificial intelligence* **134,** 57–83 (2002).

101. Campero, A. *et al.* A Test for Evaluating Performance in Human-Computer Systems. *arXiv preprint arXiv:2206.12390* (2022).

102. Can, Y. S., Arnrich, B. & Ersoy, C. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics* **92,** 103139 (2019).

103. Cardwell, B. A., Henkel, L. A., Garry, M., Newman, E. J. & Foster, J. L. Nonprobative photos rapidly lead people to believe claims about their own (and other people's) pasts. en. *Memory & Cognition* **44,** 883–896. ISSN: 0090-502x, 1532-5946. http://link.springer.com/10.3758/s13421-016-0603-1 (2022) (Aug. 2016).

104. Cardwell, B. A., Lindsay, D. S., Förster, K. & Garry, M. Uninformative photos can increase people's perceived knowledge of complicated processes. en. *Journal of Applied Research in Memory and Cognition* **6,** 244–252. ISSN: 2211-369x, 2211-3681. `http://doi.apa.org/getdoi.cfm?doi=10.1016/j.jarmac.2017.05.002` (2022) (Sept. 2017).

105. Carney, D. R. Ten things every manager should know about nonverbal behavior. *California Management Review* (2021).

106. Caruana, R. Multitask learning. *Machine learning* **28,** 41–75 (1 1997).

107. Celebi, M. E., Codella, N. & Halpern, A. Dermoscopy image analysis: overview and future directions. *IEEE journal of biomedical and health informatics* **23,** 474–478 (2019).

108. Centola, D., Guilbeault, D., Sarkar, U., Khoong, E. & Zhang, J. The reduction of race and gender bias in clinical treatment recommendations using clinician peer networks in an experimental setting. *Nature communications* **12,** 6585 (2021).

109. Chardon, A., Cretois, I. & Hourseau, C. Skin colour typology and suntanning pathways. en. *International journal of cosmetic science* **13,** 191–208. ISSN: 0142-5463, 1468-2494. `https://onlinelibrary.wiley.com/doi/10.1111/j.1467-2494.1991.tb00561.x` (2022) (Aug. 1991).

110. Chen, C. *et al.* Distinct facial expressions represent pain and pleasure across cultures. *Proceedings of the National Academy of Sciences* **115,** E10013–E10021 (43 2018).

111. Chen, H., Gomez, C., Huang, C.-M. & Unberath, M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Medicine* **5,** 156 (2022).

112. Chen, I., Johansson, F. D. & Sontag, D. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002* **31** (2018).

113. Chen, I. Y., Szolovits, P. & Ghassemi, M. Can AI help reduce disparities in general medical and mental health care? *AMA journal of ethics* **21,** 167–179 (2019).

114. Chen, I. Y. *et al.* Ethical Machine Learning in Health Care. en. *arXiv:2009.10576 [cs].* arXiv: 2009.10576. `http://arxiv.org/abs/2009.10576` (2021) (Oct. 2020).

115. Chen, P., Liu, S., Zhao, H. & Jia, J. *GridMask Data Augmentation* en. arXiv: 2001.04086. Jan. 2020. `http://arxiv.org/abs/2001.04086` (2020).

116. Chen, S. C. *et al.* Diagnosing and managing cutaneous pigmented lesions: primary care physicians versus dermatologists. *Journal of general internal medicine* **21,** 678–682 (2006).

117. Chen, V., Bhatt, U., Heidari, H., Weller, A. & Talwalkar, A. Perspectives on Incorporating Expert Feedback into Model Updates. *arXiv preprint arXiv:2205.06905* (2022).

118. Chen, X. *et al.* Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

119. Chen, X., Liu, C., Li, B., Lu, K. & Song, D. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. en. *arXiv:1712.05526 [cs].* arXiv: 1712.05526. `http://arxiv.org/abs/1712.05526` (2022) (Dec. 2017).

120. Chesney, B. & Citron, D. Deep fakes: a looming challenge for privacy, democracy, and national security. *Calif. L. Rev.* **107,** 1753 (2019).

121. Clore, G. *et al.* in *In'Theories of Mood and Cognition: A User's Handbook'(eds. Martin, LL & Clore, GL)* (2001).

122. Codella, N. *et al.* Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368* (2019).

123. Codella, N. C. *et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)* in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (2018), 168–172.

124. Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20,** 37–46 (1960).

125. Cohen, J. F. *et al.* STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ open* **6,** e012799 (2016).

126. Cohn, J. F. & De la Torre, F. Automated face analysis for affective computing. (2015).

127. Coles, N. A. *et al. A Multi-Lab Test of the Facial Feedback Hypothesis by The Many Smiles Collaboration* Feb. 2019. `psyarxiv.com/cvpuw`.

128. Collaboration, I. S. I. *et al.* Siim-isic 2020 challenge dataset. *International Skin Imaging Collaboration* (2020).

129. Combalia, M. *et al.* BCN20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288* (2019).

130. Cowen, A. S. & Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. en. *Proceedings of the National Academy of Sciences* **114,** E7900–e7909. ISSN: 0027-8424, 1091-6490. `http://www.pnas.org/lookup/doi/10.1073/pnas.1702247114` (2020) (Sept. 2017).

131. Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R. & Keltner, D. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. en. *Nature Human Behaviour* **3,** 369–382. ISSN: 2397-3374. `http://www.nature.com/articles/s41562-019-0533-6` (2021) (4 Apr. 2019).

132. Cowen, A. S. *et al.* Sixteen facial expressions occur in similar contexts worldwide. en. *Nature* **589,** 251–257. ISSN: 0028-0836, 1476-4687. `http://www.nature.com/articles/s41586-020-3037-7` (2021) (7841 Jan. 2021).

133. Cowgill, B. & Tucker, C. E. Algorithmic fairness and economics. *The Journal of Economic Perspectives* (2020).

134. Crisan, A., Drouhard, M., Vig, J. & Rajani, N. Interactive Model Cards: A Human-Centered Approach to Model Documentation. *arXiv preprint arXiv:2205.02894* (2022).

135. Crivelli, C. Inside-Out: From Basic Emotions Theory to the Behavioral Ecology View. en. *Journal of Nonverbal Behavior,* 34 (2019).

136. Crivelli, C. & Fridlund, A. J. Facial displays are tools for social influence. *Trends in Cognitive Sciences* **22,** 388–399 (5 2018).

137. Crivelli, C. & Fridlund, A. J. Facial Displays Are Tools for Social Influence. en. *Trends in Cognitive Sciences* **22,** 388–399. ISSN: 13646613. `https://linkinghub.elsevier.com/retrieve/pii/S1364661318300299` (2020) (May 2018).

138. D'Mello, S., Kopp, K., Bixler, R. E. & Bosch, N. *Attending to attention: Detecting and combating mind wandering during computerized reading* in *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems* (2016), 1661–1669.

139. D'Mello, S., Kappas, A. & Gratch, J. The affective computing approach to affect measurement. *Emotion Review* **10,** 174–183 (2 2018).

140. Dan, V. *et al.* Visual Mis-and Disinformation, Social Media, and Democracy. *Journalism & Mass Communication Quarterly* **98,** 641–664 (2021).

141. Daneshjou, R., He, B., Ouyang, D. & Zou, J. Y. How to evaluate deep learning for cancer diagnostics–factors and recommendations. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1875,** 188515 (2021).

142. Daneshjou, R., Smith, M., Sun, M., Rotemberg, V. & Zou, J. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. en, 8 (2021).

143. Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V. & Zou, J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA dermatology* **157,** 1362–1369 (2021).

144. Daneshjou, R. *et al.* Checklist for Evaluation of Image-Based Artificial Intelligence Reports in Dermatology: CLEAR Derm Consensus Guidelines From the International Skin Imaging Collaboration Artificial Intelligence Working Group. en. *JAMA Dermatology.* ISSN: 2168-6068. `https://jamanetwork.com/journals/jamadermatology/fullarticle/2786912` (2021) (Dec. 2021).

145. Daneshjou, R. *et al.* Disparities in Dermatology AI: Assessments Using Diverse Clinical Images. en. *arXiv:2111.08006 [cs, eess].* arXiv: 2111.08006. `http://arxiv.org/abs/2111.08006` (2021) (Nov. 2021).

146. Daneshjou, R. *et al.* Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science advances* **8,** eabq6147 (2022).

147. Davis, M. H. Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of personality and social psychology* **44,** 113 (1983).

148. De Melo, C. M., Carnevale, P. & Gratch, J. *The Effect of Expression of Anger and Happiness in Computer Agents on Negotiations with Humans* tech. rep. (2011). `www.ifaamas.org`.

149. De Gelder, B. Cultural differences in emotional expressions and body language. *Oxford library of psychology.* 223–234 (2016).

150. Dehon, E. *et al.* A systematic review of the impact of physician implicit racial bias on clinical decision making. *Academic Emergency Medicine* **24,** 895–904 (2017).

151. Del Bino, S. & Bernerd, F. Variations in skin colour and the biological consequences of ultraviolet radiation exposure. *British Journal of Dermatology* **169,** 33–40 (2013).

152. Della Penna, N. & Reid, M. D. Crowd & Prejudice: An Impossibility Theorem for Crowd Labelling without a Gold Standard. en. *arXiv:1204.3511 [cs].* arXiv: 1204.3511. `http://arxiv.org/abs/1204.3511` (2021) (Apr. 2012).

153. Deng, J. *et al. Imagenet: A large-scale hierarchical image database* in *2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255.

154. Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine* **29,** 141–142 (2012).

155. *DermIS — dermis.net* `https://www.dermis.net/dermisroot/en/home/index.htm`. [Accessed 17-Feb-2023].

156. Dermnet.com. *We are currently Redesigning Dermnet Skin disease Atlas — dermnet.com* `https://dermnet.com/`. [Accessed 17-Feb-2023].

157. *DermWeb — dermweb.com* `http://www.dermweb.com/photo_atlas/`. [Accessed 17-Feb-2023].

158. Dhall, A. *Context based facial expression analysis in the wild* in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013* (2013), 636–641. ISBN: 9780769550480.

159. Dhall, A. *EmotiW 2019: Automatic emotion, engagement and cohesion prediction tasks* in *2019 International Conference on Multimodal Interaction* (2019), 546–550.

160. Dhall, A., Goecke, R., Lucey, S. & Gedeon, T. *Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark* in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (2011), 2106–2112.

161. Diao, J. A. & Adamson, A. S. Representation and Misdiagnosis of Dark Skin in a Large-Scale Visual Diagnostic Challenge. en. *Journal of the American Academy of Dermatology,* S0190962221006538. ISSN: 01909622. `https://linkinghub.elsevier.com/retrieve/pii/S0190962221006538` (2021) (Apr. 2021).

162. Diao, J. A. & Adamson, A. S. Representation and misdiagnosis of dark skin in a large-scale visual diagnostic challenge. *Journal of the American Academy of Dermatology* **86,** 950–951 (2022).

163. Dias, N., Pennycook, G. & Rand, D. G. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review* **1** (2020).

164. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. en. *Journal of Experimental Psychology: General* **144,** 114–126. ISSN: 1939-2222, 0096-3445. `http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000033` (2021) (2015).

165. Dobber, T., Metoui, N., Trilling, D., Helberger, N. & de Vreese, C. Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics* **26,** 69–91 (2021).

166. Dolhansky, B., Howes, R., Pflaum, B., Baram, N. & Ferrer, C. C. The Deepfake Detection Challenge (DFDC) Preview Dataset. en. *arXiv:1910.08854 [cs].* arXiv: 1910.08854. `http://arxiv.org/abs/1910.08854` (2020) (Oct. 2019).

167. Dolhansky, B. *et al. Deepfake Detection Challenge Results: An open initiative to advance AI* Accessed: 2021-01-27. 2020. `https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/`.

168. Dolhansky, B. *et al. The DeepFake Detection Challenge (DFDC) Dataset* en. arXiv: 2006.07397. Oct. 2020. `http://arxiv.org/abs/2006.07397` (2020).

169. Doraiswamy, P. M., Blease, C. & Bodner, K. Artificial intelligence and the future of psychiatry: Insights from a global physician survey. *Artificial intelligence in medicine* **102,** 101753 (2020).

170. Doraiswamy, P. M., Chilukuri, M. M., Ariely, D. & Linares, A. R. Physician perceptions of catching COVID-19: insights from a global survey. *Journal of General Internal Medicine* **36,** 1832–1834 (2021).

171. Doraiswamy, P. M., Chilukuri, M. M., Linares, A. R. & Bramstedt, K. A. Are we ready for COVID-19's golden passport? Insights from a global physician survey. *Journal of Health and Social Sciences* **6,** 83–90 (2021).

172. Du, S., Tao, Y. & Martinez, A. M. Compound facial expressions of emotion. en. *Proceedings of the National Academy of Sciences* **111,** E1454–e1462. ISSN: 0027-8424, 1091-6490. `http://www.pnas.org/cgi/doi/10.1073/pnas.1322355111` (2020) (Apr. 2014).

173. Du Vivier, A. *Atlas of clinical dermatology* 2002.

174. Dulmage, B., Tegtmeyer, K., Zhang, M. Z., Colavincenzo, M. & Xu, S. A Point-of-Care, Real-Time Artificial Intelligence System to Support Clinician Diagnosis of a Wide Range of Skin Diseases. en. *Journal of Investigative Dermatology,* S0022202x20321679. ISSN: 0022202x. `https://linkinghub.elsevier.com/retrieve/pii/S0022202X20321679` (2021) (Oct. 2020).

175. Duncker, K. On problem-solving. en. Trans. by Lees, L. S. *Psychological Monographs* **58,** i–113. ISSN: 0096-9753. `http://doi.apa.org/getdoi.cfm?doi=10.1037/h0093599` (2022) (1945).

176. Dupré, D., Krumhuber, E. G., Küster, D. & McKeown, G. J. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *Plos one* **15,** e0231968 (2020).

177. Dupré, D., Krumhuber, E. G., Küster, D. & McKeown, G. J. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. en. *Plos One* **15** (ed D'Mello, S.) e0231968. ISSN: 1932-6203. `https://dx.plos.org/10.1371/journal.pone.0231968` (2021) (Apr. 2020).

178. Durán, J. I., Reisenzein, R. & Fernández-Dols, J.-M. Coherence between emotions and facial expressions. *The science of facial expression,* 107–129 (2017).

179. Ebede, T. & Papier, A. Disparities in dermatology educational resources. en. *Journal of the American Academy of Dermatology* **55,** 687–690. ISSN: 01909622. `https://linkinghub.elsevier.com/retrieve/pii/S0190962205047201` (2021) (Oct. 2006).

180. Ecker, U. K. *et al.* The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology* **1,** 13–29 (2022).

181. Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O. & Weisz, J. D. *Expanding explainability: Towards social transparency in ai systems* in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 1–19.

182. Eilers, S. *et al.* Accuracy of self-report in assessing Fitzpatrick skin phototypes I through VI. *JAMA dermatology* **149,** 1289–1294 (2013).

183. Ekman, P. An argument for basic emotions. *Cognition & emotion* **6,** 169–200 (1992).

184. *EmailMe Form - Derm101 was deactivated on December 31, 2019 — emailmeform.com* `https://www.emailmeform.com/builder/form/Ne0j8da9bb7U4h6t1f`. [Accessed 17-Feb-2023].

185. Epstein, Z., Pennycook, G. & Rand, D. *Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources* in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–11.

186. Epstein, Z. *et al.* Closing the AI Knowledge Gap. en. *arXiv:1803.07233 [cs].* arXiv: 1803.07233. `http://arxiv.org/abs/1803.07233` (2020) (Mar. 2018).

187. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. en. *Nature* **542,** 115–118. ISSN: 0028-0836, 1476-4687. `http://www.nature.com/articles/nature21056` (2020) (Feb. 2017).

188. Eykholt, K. *et al. Robust physical-world attacks on deep learning visual classification* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 1625–1634.

189. Farid, H. Digital doctoring: how to tell the real from the fake. *Significance* **3,** 162–166 (2006).

190. Farid, H. *Fake Photos* (MIT Press, 2019).

191. Farid, H. & Bravo, M. J. Image Forensic Analyses that Elude the Human Visual System. en. **7541** (eds Memon, N. D., Dittmann, J., Alattar, A. M. & Delp III, E. J.) 10. `http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.837788` (2020) (Feb. 2010).

192. Farrell, P. Portuguese saudade and other emotions of absence and longing. *Semantic primes and universal grammar: Empirical evidence from the Romance languages,* 235–258 (2006).

193. Farzmahdi, A., Rajaei, K., Ghodrati, M., Ebrahimpour, R. & Khaligh-Razavi, S.-M. A specialized face-processing model inspired by the organization of monkey face patches explains several face-specific phenomena observed in humans. *Scientific reports* **6,** 1–17 (2016).

194. Fazio, L. K., Brashier, N. M., Payne, B. K. & Marsh, E. J. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General* **144,** 993 (2015).

195. Federman, D. G., Concato, J. & Kirsner, R. S. Comparison of dermatologic diagnoses by primary care practitioners and dermatologists: a review of the literature. *Archives of family medicine* **8,** 170 (1999).

196. Federman, D. & Kirsner, R. S. The abilities of primary care physicians in dermatology: implications for quality of care. *Am J Manag Care* **3,** 1487–1492 (1997).

197. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I. & Lehmann, S. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524* (2017).

198. Fenton, A. *et al.* Medical students' ability to diagnose common dermatologic conditions in skin of color. en. *Journal of the American Academy of Dermatology* **83,** 957. ISSN: 01909622. https://linkinghub.elsevier.com/retrieve/pii/S0190962220301444 (2021) (Sept. 2020).

199. Ferguson, C., Lewis, R., Wilks, C. & Picard, R. W. *The Guardians: Designing a Game for Long-term Engagement with Mental Health Therapy* in *2021 IEEE Conference on Games (CoG)* (2021).

200. Fernández-Dols, J.-M. & Crivelli, C. Emotion and expression: Naturalistic studies. *Emotion Review* **5,** 24–29 (2013).

201. Fernberger, S. W. *False Suggestion and the Piderit Model Author* tech. rep. 4 (1928), 562–568.

202. Finlayson, S. G. *et al.* Adversarial attacks on medical machine learning. en. *Science* **363,** 1287–1289. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.aaw4399 (2022) (Mar. 2019).

203. Finlayson, S. G. *et al.* The Clinician and Dataset Shift in Artificial Intelligence. en. *New England Journal of Medicine* **385,** 283–286. ISSN: 0028-4793, 1533-4406. http://www.nejm.org/doi/10.1056/NEJMc2104626 (2021) (July 2021).

204. Firdaus, M., Thangavelu, N., Ekbal, A. & Bhattacharyya, P. I enjoy writing and playing, do you: A Personalized and Emotion Grounded Dialogue Agent using Generative Adversarial Network. *IEEE Transactions on Affective Computing* (2022).

205. Firestone, C. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences* **117,** 26562–26571 (2020).

206. Firestone, C. Performance vs. competence in human–machine comparisons. en. *Proceedings of the National Academy of Sciences* **117,** 26562–26571. ISSN: 0027-8424, 1091-6490. http://www.pnas.org/lookup/doi/10.1073/pnas.1905334117 (2021) (Oct. 2020).

207. Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10,** 507–521 (1915).

208. Fisher, R. A. On the'probable error'of a coefficient of correlation deduced from a small sample. *Metron* **1,** 1–32 (1921).

209. Fitzpatrick, T. B. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology* **124,** 869–871 (1988).

210. Forgas, J. P. Happy believers and sad skeptics? Affective influences on gullibility. *Current Directions in Psychological Science* **28,** 306–313 (2019).

211. Forgas, J. P. & East, R. On being happy and gullible: Mood effects on skepticism and the detection of deception. *Journal of Experimental Social Psychology* **44,** 1362–1367 (2008).

212. Forsythe, D. E. Engineering knowledge: The construction of knowledge in artificial intelligence. *Social studies of science* **23,** 445–477 (1993).

213. Foul, Y. A., Eitan, R. & Aviezer, H. Perceiving emotionally incongruent cues from faces and bodies: Older adults get the whole picture. *Psychology and aging* **33,** 660 (4 2018).

214. Frederick, S. Cognitive Reflection and Decision Making. en. *Journal of Economic Perspectives* **19,** 25–42. ISSN: 0895-3309. `https://pubs.aeaweb.org/doi/10.1257/089533005775196732` (2022) (Nov. 2005).

215. Freire da Silva, S. Atlas Dermatologico. *http://atlasdermatologico.com.br/.*

216. Fridlund, A. J. Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of personality and social psychology* **60,** 229 (1991).

217. Fridlund, A. J. Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of personality and social psychology* **60,** 229 (2 1991).

218. Fuller, T. *Gnomologia: adagies and proverbs; wise sentences and witty sayings, ancient and modern, foreign and British* (B. Barker, 1732).

219. Galton, F. *Vox populi* 1907.

220. Garimella, K. & Eckles, D. Images and misinformation in political groups: Evidence from WhatsApp in India. *Harvard Kennedy School Misinformation Review* (2020).

221. Gaube, S. *et al.* Do as AI say: susceptibility in deployment of clinical decision-aids. en. *NPJ digital medicine* **4,** 1–8. ISSN: 2398-6352. `http://www.nature.com/articles/s41746-021-00385-9` (2021) (Dec. 2021).

222. Gebru, T. *et al.* Datasheets for datasets. *Communications of the ACM* **64,** 86–92 (2021).

223. Gendron, M., Mesquita, B. & Barrett, L. F. Emotion perception: Putting the face in context. *The Oxford handbook of cognitive psychology. Oxford library of psychology.* 539–556 (2013).

224. Ghandeharioun, A., Eoff, B., Jou, B. & Picard, R. W. Characterizing Sources of Uncertainty to Proxy Calibration and Disambiguate Annotator and Data Bias. en. *arXiv:1909.09285 [cs, stat]*. arXiv: 1909.09285. `http://arxiv.org/abs/1909.09285` (2020) (Oct. 2019).

225. Ghandeharioun, A., McDuff, D., Czerwinski, M. & Rowan, K. *EMMA: An Emotion-Aware Wellbeing Chatbot* en. in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (Ieee, Cambridge, United Kingdom, Sept. 2019), 1–7. ISBN: 978-1-72813-888-6. `https://ieeexplore.ieee.org/document/8925455/` (2022).

226. Ghandeharioun, A., McDuff, D., Czerwinski, M. & Rowan, K. *Towards Understanding Emotional Intelligence for Behavior Change Chatbots* en. in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (Ieee, Cambridge, United Kingdom, Sept. 2019), 8–14. ISBN: 978-1-72813-888-6. `https://ieeexplore.ieee.org/document/8925433/` (2022).

227. Ghandeharioun, A. *et al.* Dissect: Disentangled simultaneous explanations via concept traversals. *arXiv preprint arXiv:2105.15164* (2021).

228. Gillan, C. M. & Rutledge, R. B. Smartphones and the neuroscience of mental health. *Annual Review of Neuroscience* **44,** 129–151 (2021).

229. Giotis, I. *et al.* MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert systems with applications* **42,** 6578–6585 (2015).

230. Glasford, D. E. Seeing is believing: communication modality, anger, and support for action on behalf of out-groups. *Journal of Applied Social Psychology* **43,** 2223–2230 (2013).

231. Goel, V., Raj, S. & Ravichandran, P. How WhatsApp leads mobs to murder in India. *The New York Times* **18** (2018).

232. Goldenberg, A. *et al. Amplification in the Evaluation of Emotional Expressions Over Time* en. preprint (PsyArXiv, June 2021). `https://osf.io/rfgy3` (2021).

233. Goodfellow, I. J. *et al. Challenges in representation learning: A report on three machine learning contests* in *International conference on neural information processing* (2013), 117–124.

234. Goodwin, C. Practices of color classification. *Mind, culture, and activity* **7,** 19–36 (2000).

235. Gordon, M. L. *et al.* Jury Learning: Integrating Dissenting Voices into Machine Learning Models. en. *arXiv:2202.02950 [cs]*. arXiv: 2202.02950. `http://arxiv.org/abs/2202.02950` (2022) (Feb. 2022).

236. Goren, C. C., Sarty, M. & Wu, P. Y. Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics* **56,** 544–549 (1975).

237. Grandey, A. A. Emotional regulation in the workplace: A new way to conceptualize emotional labor. *Journal of occupational health psychology* **5,** 95 (1 2000).

238. Green, B. & Chen, Y. The principles and limits of algorithm-in-the-loop decision making. en. *Proceedings of the ACM on Human-Computer Interaction* **3,** 1–24. ISSN: 2573-0142. https://dl.acm.org/doi/10.1145/3359152 (2022) (Nov. 2019).

239. Green, B. & Chen, Y. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. en. *Proceedings of the ACM on Human-Computer Interaction* **5.** arXiv: 2012.05370, 1–33. ISSN: 2573-0142. http://arxiv.org/abs/2012.05370 (2022) (Oct. 2021).

240. Greenaway, K. H., Kalokerinos, E. K. & Williams, L. A. Context is everything (in emotion research). *Social and Personality Psychology Compass* **12,** e12393 (6 2018).

241. Griffiths, C., Barker, J., Bleiker, T. O., Chalmers, R. & Creamer, D. *Rook's textbook of dermatology* (John Wiley & Sons, 2016).

242. Groh, M. *Identifying the Context Shift between Test Benchmarks and Production Data* 2022. https://arxiv.org/abs/2207.01059.

243. Groh, M., Epstein, Z., Firestone, C. & Picard, R. Deepfake detection by human crowds, machines, and machine-informed crowds. en. *Proceedings of the National Academy of Sciences* **119,** e2110013119. ISSN: 0027-8424, 1091-6490. http://www.pnas.org/lookup/doi/10.1073/pnas.2110013119 (2022) (Jan. 2022).

244. Groh, M., Epstein, Z., Obradovich, N., Cebrian, M. & Rahwan, I. Human detection of machine-manipulated media. *Communications of the ACM* **64,** 40–47 (2021).

245. Groh, M., Ferguson, C., Lewis, R. & Picard, R. W. *Computational Empathy Counteracts the Negative Effects of Anger on Creative Problem Solving* in *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)* (2022), 1–8.

246. Groh, M., Harris, C., Daneshjou, R., Badri, O. & Koochek, A. Towards Transparency in Dermatology Image Datasets with Skin Tone Annotations by Experts, Crowds, and an Algorithm. *Proceedings of the ACM on Human-Computer Interaction* **6,** 1–26 (2022).

247. Groh, M., Sankaranarayanan, A. & Picard, R. Human detection of political deepfakes across transcripts, audio, and video. *arXiv preprint arXiv:2202.12883* (2022).

248. Groh, M. *et al. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset* en. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* arXiv: 2104.09957 (Ieee, Nashville, TN, USA, June 2021), 1820–1828. ISBN: 978-1-66544-899-4. https://ieeexplore.ieee.org/document/9522867/ (2021).

249. Gross, R., Matthews, I., Cohn, J., Kanade, T. & Baker, S. Multi-pie. *Image and vision computing* **28,** 807–813 (2010).

250. Guess, A. M. & Lyons, B. A. Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform,* 10–33 (2020).

251. Guess, A. M., Nyhan, B. & Reifler, J. Exposure to untrustworthy websites in the 2016 US election. en. *Nature human behaviour* **4,** 472–480. ISSN: 2397-3374. http://www.nature.com/articles/s41562-020-0833-x (2021) (May 2020).

252. Guha, R. V. *Contexts: a formalization and some applications* PhD thesis (Stanford University, 1992).

253. Gupta, A. K., Bharadwaj, M. & Mehrotra, R. Skin Cancer Concerns in People of Color: Risk Factors and Prevention. en. *Asian Pacific journal of cancer prevention: APJCP* **17,** 8 (2016).

254. Gupta, V. & Sharma, V. K. Skin typing: Fitzpatrick grading and others. *Clinics in dermatology* **37,** 430–436 (2019).

255. Habif, T. Clinical dermatology: A color guide to diagnosis and therapy (2010).

256. Haibe-Kains, B. *et al.* Transparency and reproducibility in artificial intelligence. *Nature* **586,** E14–e16 (2020).

257. Halberstadt, J. B., Niedenthal, P. M. & Kushner, J. Resolution of lexical ambiguity by emotional state. *Psychological Science* **6,** 278–282 (1995).

258. Hameleers, M., Powell, T. E., Van Der Meer, T. G. & Bos, L. A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media. en. *Political Communication* **37,** 281–301. ISSN: 1058-4609, 1091-7675. `https://www.tandfonline.com/doi/full/10.1080/10584609.2019.1674979` (2021) (Mar. 2020).

259. Hameleers, M., van der Meer, T. G. & Dobber, T. You Won't Believe What They Just Said! The Effects of Political Deepfakes Embedded as Vox Populi on Social Media. *Social Media+ Society* **8,** 20563051221116346 (2022).

260. Hammal, Z. & Suarez, M. T. *Towards context based affective computing* in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013* (2013), 802. ISBN: 9780769550480.

261. Han, J., Zhang, Z., Schmitt, M., Pantic, M. & Schuller, B. *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* (2017), 890–897.

262. Han, S. S. *et al.* Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology* **138,** 1529–1538 (2018).

263. Han, S. S. *et al.* Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *Journal of Investigative Dermatology* **140,** 1753–1761 (2020).

264. Hancock, J. T. & Bailenson, J. N. The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking* **24,** 149–152 (2021).

265. Hancock, J. T., Naaman, M. & Levy, K. AI-mediated communication: definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication* **25,** 89–100 (2020).

266. Hansen, J. H. L. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech communication* **20,** 151–173 (1996).

267. Hara, K. *et al. A data-driven analysis of workers' earnings on Amazon Mechanical Turk* en. in *Proceedings of the 2018 CHI conference on human factors in computing systems* (Acm, Montreal QC Canada, Apr. 2018), 1–14. ISBN: 978-1-4503-5620-6. https://dl.acm.org/doi/10.1145/3173574.3174023 (2022).

268. Hardy, G. H. *A mathematician's apology* (Cambridge University Press, 1992).

269. Harvey, N. T., Chan, J. & Wood, B. A. Skin biopsy in the diagnosis of inflammatory skin disease. *Australian Family Physician* **46,** 283–288 (2017).

270. Hastie, H. *et al.* I remember you! interaction with memory for an empathic virtual robotic tutor (2016).

271. Haut, K. *et al. Could you become more credible by being White? Assessing Impact of Race on Credibility with Deepfakes* 2021.

272. Hazirbas, C. *et al.* Towards measuring fairness in AI: the Casual Conversations dataset. en, 9.

273. Healey, J., Wang, H. & Chhaya, N. *Challenges in Recognizing Spontaneous and Intentionally Expressed Reactions to Positive and Negative Images* en. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Ieee, Seattle, WA, USA, June 2020), 1622–1630. ISBN: 978-1-72819-360-1. https://ieeexplore.ieee.org/document/9150616/ (2021).

274. Heger, A., Marquis, E. B., Vorvoreanu, M., Wallach, H. & Vaughan, J. W. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *arXiv preprint arXiv:2206.02923* (2022).

275. Hess, U. & Hareli, S. The social signal value of emotions: The role of contextual factors in social inferences drawn from emotion displays. *Oxford series in social cognition and social neuroscience.* 375–393 (2017).

276. Hjortsjö, C.-H. *Man's face and mimic language* (Studentlitteratur, 1969).

277. Hoegen, R., Gratch, J., Parkinson, B. & Shore, D. *Signals of emotion regulation in a social dilemma: Detection from face and context* in (2019), 1–7.

278. Hofman, J. M. *et al.* Integrating explanation and prediction in computational social science. en. *Nature* **595,** 181–188. ISSN: 0028-0836, 1476-4687. http://www.nature.com/articles/s41586-021-03659-0 (2021) (July 2021).

279. Holland, S., Hosny, A., Newman, S., Joseph, J. & Chmielinski, K. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677* (2018).

280. *Home | Hellenic Dermatological Atlas - Over 2700 Dermatology pictures — hellenicdermatlas.com* http://www.hellenicdermatlas.com/en/. [Accessed 17-Feb-2023].

281. Hong, W., Ding, M., Zheng, W., Liu, X. & Tang, J. *CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers* 2022. https://arxiv.org/abs/2205.15868.

282. Hoque, M. E., McDuff, D. & Picard, R. W. Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Transactions on Affective Computing* **3,** 323–334 (3 2012).

283. Hotelling, H. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)* **15,** 193–232 (1953).

284. Huang, D. & De La Torre, F. *Facial action transfer with personalized bilinear regression* in *European Conference on Computer Vision* (2012), 144–158.

285. Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A. & Narayanan, A. The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *arXiv preprint arXiv:2203.06498* (2022).

286. Hutchinson, B., Rostamzadeh, N., Greer, C., Heller, K. & Prabhakaran, V. Evaluation Gaps in Machine Learning Practice. *arXiv preprint arXiv:2205.05256* (2022).

287. Hutto, C. & Gilbert, E. *VADER: A parsimonious rule-based model for sentiment analysis of social media text* in (2014).

288. Ickes, W. Empathic Accuracy. *Journal of Personality* **61,** 4 (1993).

289. Ickes, W. Empathic accuracy. *Journal of personality* **61,** 587–610 (1993).

290. *Iconotheque Numerique de L'Universite Libre de Bruxelles* `https://icono.ulb.ac.be/`. [Accessed 17-Feb-2023].

291. Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G. & Madry, A. *Datamodels: Understanding predictions with data and data with predictions* in *International Conference on Machine Learning* (2022), 9525–9587.

292. Ilyas, A. *et al.* Adversarial Examples are not Bugs, they are Features. en. *Advances in neural information processing systems* **32,** 12 (2019).

293. Irani, L. C. & Silberman, M. S. *Turkopticon: Interrupting worker invisibility in amazon mechanical turk* en. in *Proceedings of the SIGCHI conference on human factors in computing systems* (Acm, Paris France, Apr. 2013), 611–620. ISBN: 978-1-4503-1899-0. `https://dl.acm.org/doi/10.1145/2470654.2470742` (2022).

294. Irvin, J. *et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison* in *Proceedings of the AAAI conference on artificial intelligence* **33** (2019), 590–597.

295. Isen, A., Daubman, K. & Nowicki, G. Positive affect facilitates creative problem solving. *Journal of personality and social psychology* **52,** 1122 (1987).

296. Isen, A. M. Positive affect and decision making. (1993).

297. Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R. & Schyns, P. G. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* **109,** 7241–7244 (19 2012).

298. Jackson-Richards, D. & Pandya, A. G. *Dermatology atlas for skin of color* (Springer, 2014).

299. Jacobs, M. *et al.* How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. en. *Translational psychiatry* **11,** 1–9. ISSN: 2158-3188. `http://www.nature.com/articles/s41398-021-01224-x` (2021) (June 2021).

300. Jain, A. *et al.* Development and Assessment of an Artificial Intelligence–Based Tool for Skin Condition Diagnosis by Primary Care Physicians and Nurse Practitioners in Teledermatology Practices. en. *JAMA Network Open* **4,** e217249. ISSN: 2574-3805. `https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2779250` (2021) (Apr. 2021).

301. Jain, S., Lawrence, H., Moitra, A. & Madry, A. Distilling Model Failures as Directions in Latent Space. *arXiv preprint arXiv:2206.14754* (2022).

302. Jakesch, M., Koren, M., Evtushenko, A. & Naaman, M. The role of source, headline and expressive responding in political news evaluation. *Headline and Expressive Responding in Political News Evaluation (December 5, 2018)* (2018).

303. Jiang, L., Li, R., Wu, W., Qian, C. & Loy, C. C. *DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection* en. arXiv: 2001.03024. Dec. 2020. `http://arxiv.org/abs/2001.03024` (2021).

304. Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Scientific data* **3,** 1–9 (2016).

305. Jospe, K. *et al.* The contribution of linguistic and visual cues to physiological synchrony and empathic accuracy. en. *Cortex* **132,** 296–308. ISSN: 00109452. `https://linkinghub.elsevier.com/retrieve/pii/S0010945220303233` (2021) (Nov. 2020).

306. Jr, R. B. A., Hess, U. & Kleck, R. E. The intersection of gender-related facial appearance and facial displays of emotion. *Emotion Review* **7,** 5–13 (1 2015).

307. Kahneman, D. *Thinking, fast and slow* (Macmillan, 2011).

308. Kahneman, D. & Krueger, A. B. *Developments in the Measurement of Subjective Well-Being* tech. rep. (2006), 3–24.

309. Kailas, A. Taylor and Kelly's dermatology for skin of color. *Journal of the American Academy of Dermatology* **76,** e75 (2017).

310. Kane, K. S., Lio, P. A. & Stratigos, A. Color atlas and synopsis of pediatric dermatology (2009).

311. Kang, S. *Fitzpatrick's Dermatology, 2-Volume Set (Fitzpatricks* (2019).

312. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience* **17,** 4302–4311 (1997).

313. Kapoor, S. & Narayanan, A. Leakage and the Reproducibility Crisis in ML-based Science. *arXiv preprint arXiv:2207.07048* (2022).

314. Karras, T., Laine, S. & Aila, T. *A style-based generator architecture for generative adversarial networks* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 4401–4410.

315. Karras, T. *et al.* *Analyzing and improving the image quality of stylegan* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 8110–8119.

316. Kasos, K. *et al.* Does the electrodermal system "Take Sides" when it comes to emotions? *Applied psychophysiology and biofeedback* **43,** 203–210 (2018).

317. Kasparov, G. *Chess, a Drosophila of reasoning* 2018.

318. Kasra, M., Shen, C. & O'Brien, J. F. Seeing Is Believing: How People Fail to Identify Fake Images on the Web. en, 6 (2018).

319. Katz, D. M., Bommarito, M. J., Gao, S. & Arredondo, P. GPT-4 Passes the Bar Exam. *Available at SSRN 4389233* (2023).

320. Kaur, H., McDuff, D., Williams, A. C., Teevan, J. & Iqbal, S. T. *"I Didn't Know I Looked Angry": Characterizing Observed Emotion and Reported Affect at Work* in *CHI Conference on Human Factors in Computing Systems* (2022), 1–18.

321. Kawahara, J., Daneshvar, S., Argenziano, G. & Hamarneh, G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics* **23,** 538–546 (2018).

322. Keltner, D., Sauter, D., Tracy, J. & Cowen, A. Emotional expression: Advances in basic emotion theory. *Journal of nonverbal behavior,* 1–28 (2019).

323. Kerr, S. On the folly of rewarding A, while hoping for B. *Academy of Management journal* **18,** 769–783 (1975).

324. Kerrigan, D., Hullman, J. & Bertini, E. A survey of domain knowledge elicitation in applied machine learning. *Multimodal Technologies and Interaction* **5,** 73 (2021).

325. Keys, R. T., Taubert, J. & Wardle, S. G. *A visual search advantage for illusory faces in objects* 2021.

326. Khalid, A. T. *et al.* Utility of sun-reactive skin typing and melanin index for discerning vitamin D deficiency. *Pediatric research* **82,** 444–451 (2017).

327. Kiani, A. *et al.* Impact of a deep learning assistant on the histopathologic classification of liver cancer. en. *npj Digital Medicine* **3,** 23. ISSN: 2398-6352. `http://www.nature.com/articles/s41746-020-0232-8` (2022) (Dec. 2020).

328. Kiela, D. *et al.* Dynabench: Rethinking benchmarking in NLP. *arXiv preprint arXiv:2104.14337* (2021).

329. Kilikita, J. *Rosacea is common in dark skin, too. here's what you need to know* `https://www.refinery29.com/en-gb/rosacea-dark-skin`.

330. Kim, A., Moravec, P. L. & Dennis, A. R. Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems* **36,** 931–968 (2019).

331. King, D. *The commissar vanishes: The falsification of photographs and art in Stalin's Russia* (Metropolitan Books New York, 1997).

332. Kinyanjui, N. M. *et al.* Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets. en. *arXiv:1910.13268 [cs, stat].* arXiv: 1910.13268. `http://arxiv.org/abs/1910.13268` (2020) (Oct. 2019).

333. Kirschbaum, C., Pirke, K. M. & Hellhammer, D. H. *The 'Trier social stress test' - A tool for investigating psychobiological stress responses in a laboratory setting* in *Neuropsychobiology* **28** (1993), 76–81.

334. Klein, J., Moon, Y. & Picard, R. W. This computer responds to user frustration: Theory, design, and results. *Interacting with computers* **14,** 119–140 (2002).

335. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. & Mullainathan, S. Human decisions and machine predictions. *The quarterly journal of economics* **133,** 237–293 (2018).

336. Kleinberg, J., Ludwig, J., Mullainathan, S. & Sunstein, C. R. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* **10,** 113–174 (2018).

337. Kleinberg, J., Ludwig, J., Mullainathan, S. & Sunstein, C. R. Algorithms as discrimination detectors. en. *Proceedings of the National Academy of Sciences,* 201912790. ISSN: 0027-8424, 1091-6490. http://www.pnas.org/lookup/doi/10.1073/pnas.1912790117 (2020) (July 2020).

338. Kleinginna, P. R. & Kleinginna, A. M. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion* **5,** 345–379 (1981).

339. Knoop, K. J. *The atlas of emergency medicine* 2010.

340. Köbis, N., Doležalová, B. & Soraperra, I. Fooled Twice–People Cannot Detect Deepfakes But Think They Can. *iScience* **24** (2021).

341. Kocielnik, R., Sidorova, N., Maggi, F. M., Ouwerkerk, M. & Westerink, J. H. D. M. *Smart technologies for long-term stress monitoring at work* in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems* (2013), 53–58.

342. Koenig, J. The dictionary of obscure sorrows. *Acesso em* **12** (2017).

343. Koh, P. W. *et al.* WILDS: A Benchmark of in-the-Wild Distribution Shifts. en. *arXiv:2012.07421 [cs].* arXiv: 2012.07421. http://arxiv.org/abs/2012.07421 (2021) (July 2021).

344. Kolkur, S., Kalbande, D., Shimpi, P., Bapat, C. & Jatakia, J. Human Skin Detection Using RGB, HSV and YCbCr Color Models. en. *Proceedings of the International Conference on Communication and Signal Processing 2016 (ICCASP 2016).* http://dx.doi.org/10.2991/iccasp-16.2017.51 (2021) (2017).

345. Kollareth, D., Fernandez-Dols, J.-M. & Russell, J. A. Shame as a culture-specific emotion concept. *Journal of Cognition and Culture* **18,** 274–292 (2018).

346. Kollareth, D. & Russell, J. A. The English word disgust has no exact translation in Hindi or Malayalam. *Cognition and Emotion* **31,** 1169–1180. ISSN: 14640600 (6 Aug. 2017).

347. Kollias, D. *et al.* Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. en. *International Journal of Computer Vision* **127,** 907–929. ISSN: 0920-5691, 1573-1405. http://link.springer.com/10.1007/s11263-019-01158-4 (2020) (June 2019).

348. Korshunov, P. & Marcel, S. *DeepFakes: a New Threat to Face Recognition? Assessment and Detection* en. arXiv: 1812.08685. Dec. 2018. `http://arxiv.org/abs/1812.08685` (2021).

349. Kosti, R., Alvarez, J. M., Recasens, A. & Lapedriza, A. *Emotion Recognition in Context* en. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Ieee, Honolulu, HI, July 2017), 1960–1968. ISBN: 978-1-5386-0457-1. `http://ieeexplore.ieee.org/document/8099695/` (2020).

350. Kosti, R., Alvarez, J. M., Recasens, A. & Lapedriza, A. Context Based Emotion Recognition using EMOTIC Dataset. en. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* arXiv: 2003.13401, 1–1. ISSN: 0162-8828, 2160-9292, 1939-3539. `http://arxiv.org/abs/2003.13401` (2020) (2019).

351. Kostick-Quenet, K. M. & Gerke, S. AI in the hands of imperfect users. *npj Digital Medicine* **5,** 197 (2022).

352. Krishnapriya, K., Pangelinan, G., King, M. C. & Bowyer, K. W. *Analysis of Manual and Automated Skin Tone Assignments* en. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022), 429–438.

353. Krizhevsky, A., Nair, V. & Hinton, G. The CIFAR-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html* **55** (2014).

354. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012).

355. Krumhuber, E., Manstead, A. S. R. & Kappas, A. Temporal aspects of facial displays in person and expression perception: The effects of smile dynamics, head-tilt, and gender. *Journal of Nonverbal Behavior* **31,** 39–56 (2007).

356. Krumhuber, E. G., Kappas, A. & Manstead, A. S. R. Effects of dynamic aspects of facial expressions: A review. *Emotion Review* **5,** 41–46 (1 2013).

357. Krumhuber, E. G., Küster, D., Namba, S., Shah, D. & Calvo, M. G. Emotion recognition from posed and spontaneous dynamic expressions: Human observers versus machine analysis. *Emotion* (2019).

358. Lahiri, A., Kwatra, V., Frueh, C., Lewis, J. & Bregler, C. *LipSync3D: Data-Efficient Learning of Personalized 3D Talking Faces from Video using Pose and Lighting Normalization* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 2755–2764.

359. Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A. & Tan, C. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

360. Lai, V. & Tan, C. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. en. *Proceedings of the Conference on Fairness, Accountability, and Transparency.* arXiv: 1811.07901, 29–38. `http://arxiv.org/abs/1811.07901` (2021) (Jan. 2019).

361. Laurençon, H. *et al.* The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems* **35,** 31809–31826 (2022).

362. Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S. & David, A. S. Measuring empathy: reliability and validity of the Empathy Quotient. *Psychological medicine* **34,** 911–920 (2004).

363. Lazer, D. Studying human attention on the Internet. *Proceedings of the National Academy of Sciences* **117,** 21–22 (2020).

364. Lazer, D. *et al.* Computational social science. *Science* **323,** 721–723 (2009).

365. Lazer, D. M. *et al.* The science of fake news. en. *Science* **359,** 1094–1096. ISSN: 0036-8075, 1095-9203. `https://www.sciencemag.org/lookup/doi/10.1126/science.aao2998` (2021) (Mar. 2018).

366. Lazer, D. M. *et al.* Computational social science: Obstacles and opportunities. *Science* **369,** 1060–1062 (2020).

367. Lazerus, T., Ingbretsen, Z. A., Stolier, R. M., Freeman, J. B. & Cikara, M. Positivity bias in judging ingroup members' emotional expressions. *Emotion* **16,** 1117–1125. ISSN: 19311516 (8 Dec. 2016).

368. Lebovitz, S., Levina, N. & Lifshitz-Assaf, H. Is AI Ground Truth Really 'True'? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What. *The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What (May 4, 2021). Citation: Lebovitz, S., Levina, N., Lifshitz-Assaf, H,* 1501–1525 (2021).

369. Lebovitz, S., Lifshitz-Assaf, H. & Levina, N. To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization science* **33,** 126–148 (2022).

370. Lee, E.-J. & Shin, S. Y. Mediated Misinformation: Questions Answered, More Questions to Ask. en. *American Behavioral Scientist* **65,** 259–276. ISSN: 0002-7642, 1552-3381. `http://journals.sagepub.com/doi/10.1177/0002764219869403` (2021) (Feb. 2021).

371. Leibowicz, C., McGregor, S. & Ovadya, A. *The Deepfake Detection Dilemma: A Multistakeholder Exploration of Adversarial Dynamics in Synthetic Media* en. arXiv: 2102.06109. Feb. 2021. `http://arxiv.org/abs/2102.06109` (2021).

372. Lenat, D. The Dimensions of Context-Space. en, 78 (1998).

373. Lerner, J. S., Li, Y., Valdesolo, P. & Kassam, K. S. Emotion and decision making. *Annual review of psychology* **66,** 799–823 (2015).

374. Lerner, J. S. & Keltner, D. Beyond valence: Toward a model of emotion-specific influences on judgement and choice. en. *Cognition & Emotion* **14,** 473–493. ISSN: 0269-9931, 1464-0600. `http://www.tandfonline.com/doi/abs/10.1080/026999300402763` (2021) (July 2000).

375. Lerner, J. S., Li, Y., Valdesolo, P. & Kassam, K. S. Emotion and Decision Making. en. *Annual Review of Psychology* **66,** 799–823. ISSN: 0066-4308, 1545-2085. `http://www.annualreviews.org/doi/10.1146/annurev-psych-010213-115043` (2021) (Jan. 2015).

376. Lerner, J. S., Small, D. A. & work(s): G. L. R. Heart Strings and Purse Strings: Carryover Effects of Emotions on Economic Decisions. en. *Psychological Science* **15,** 337–341. http://www.jstor.org/stable/40063984 (2004).

377. Lerner, J. S. & Tiedens, L. Z. Portrait of the angry decision maker: how appraisal tendencies shape anger's influence on cognition. en. *Journal of Behavioral Decision Making* **19,** 115–137. ISSN: 0894-3257, 1099-0771. https://onlinelibrary.wiley.com/doi/10.1002/bdm.515 (2022) (Apr. 2006).

378. Lester, J., Jia, J., Zhang, L., Okoye, G. & Linos, E. Absence of images of skin of colour in publications of COVID-19 skin manifestations. en. *British Journal of Dermatology* **183,** 593–595. ISSN: 0007-0963, 1365-2133. https://onlinelibrary.wiley.com/doi/10.1111/bjd.19258 (2020) (Sept. 2020).

379. Lester, J., Clark Jr, L., Linos, E. & Daneshjou, R. Clinical Photography in Skin of Color: Tips and Best Practices. *The British Journal of Dermatology* (2021).

380. Lester, J. & Shinkai, K. Diversity and Inclusivity Are Essential to the Future of Dermatology. en. *Cutis* **104,** 2 (2019).

381. Li, J., Schmidt, F. & Kolter, Z. *Adversarial camera stickers: A physical camera-based attack on deep learning systems* in *International Conference on Machine Learning* (2019), 3896–3904.

382. Li, S. & Deng, W. Deep Facial Expression Recognition: A Survey. en. *IEEE Transactions on Affective Computing.* arXiv: 1804.08348, 1–1. ISSN: 1949-3045, 2371-9850. https://ieeexplore.ieee.org/document/9039580/ (2020) (2020).

383. Li, Y., Yang, X., Sun, P., Qi, H. & Lyu, S. *Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics* en. arXiv: 1909.12962. Mar. 2020. http://arxiv.org/abs/1909.12962 (2021).

384. Licklider, J. C. Man-computer symbiosis. *IRE transactions on human factors in electronics,* 4–11 (1960).

385. Lipton, Z. C. The mythos of model interpretability. en. *Communications of the ACM* **61,** 36–43. ISSN: 0001-0782, 1557-7317. https://dl.acm.org/doi/10.1145/3233231 (2022) (Sept. 2018).

386. Liu, H., Lai, V. & Tan, C. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. en. *Proceedings of the ACM on Human-Computer Interaction* **5.** arXiv: 2101.05303, 1–45. ISSN: 2573-0142. http://arxiv.org/abs/2101.05303 (2021) (Oct. 2021).

387. Liu, X. *et al.* The medical algorithmic audit. *The Lancet Digital Health* (2022).

388. Liu, Y. *et al.* A deep learning system for differential diagnosis of skin diseases. en. *Nature Medicine* **26,** 900–908. ISSN: 1078-8956, 1546-170x. http://www.nature.com/articles/s41591-020-0842-3 (2020) (June 2020).

389. Logg, J. M., Minson, J. A. & Moore, D. A. Algorithm appreciation: People prefer algorithmic to human judgment. en. *Organizational Behavior and Human Decision Processes* **151,** 90–103. ISSN: 07495978. https://linkinghub.elsevier.com/retrieve/pii/S0749597818303388 (2022) (Mar. 2019).

390. Long, B., Simson, J., Buxó-Lugo, A., Watson, D. G. & Mehr, S. A. How Games Can Make Behavioural Science Better. *Nature* **613,** 433–436. ISSN: 0028-0836, 1476-4687 (Jan. 2023).

391. Louie, P. & Wilkes, R. Representations of race and skin tone in medical textbook imagery. en. *Social Science & Medicine* **202,** 38–42. ISSN: 02779536. `https://linkinghub.elsevier.com/retrieve/pii/S0277953618300790` (2020) (Apr. 2018).

392. Lu, H. *et al. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones* in *Proceedings of the 2012 ACM conference on ubiquitous computing* (2012), 351–360.

393. Lucey, P. *et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression* in *2010 ieee computer society conference on computer vision and pattern recognition-workshops* (2010), 94–101.

394. Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D. & Munafò, M. R. Gamification of Cognitive Assessment and Cognitive Training: A Systematic Review of Applications and Efficacy. *JMIR Serious Games* (2016).

395. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. en. *Nature biomedical engineering* **2,** 749–760. ISSN: 2157-846x. `http://www.nature.com/articles/s41551-018-0304-0` (2021) (Oct. 2018).

396. Luong, H.-T. & Yamagishi, J. Nautilus: a versatile voice cloning system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **28,** 2967–2981 (2020).

397. Lyu, S. DeepFake Detection: Current Challenges and Next Steps. en. *arXiv:2003.09234 [cs].* arXiv: 2003.09234, 1–6. `http://arxiv.org/abs/2003.09234` (2020) (Mar. 2020).

398. Maes, P. in *Readings in human–computer interaction* 811–821 (Elsevier, 1995).

399. Majumder, N. *et al. Dialoguernn: An attentive rnn for emotion detection in conversations* in. **33** (2019), 6818–6825.

400. Makar, M. *et al.* Causally-motivated Shortcut Removal Using Auxiliary Labels. en. *arXiv:2105.06422 [cs].* arXiv: 2105.06422, 739–766. `http://arxiv.org/abs/2105.06422` (2021) (June 2021).

401. Mandel, T. *et al.* Using the crowd to prevent harmful AI behavior. *Proceedings of the ACM on Human-Computer Interaction* **4,** 1–25 (2020).

402. Mao, Y. *et al.* How data scientistswork together with domain experts in scientific collaborations: To find the right answer or to ask the right question? en. *Proceedings of the ACM on Human-Computer Interaction* **3,** 1–23. ISSN: 2573-0142. `https://dl.acm.org/doi/10.1145/3361118` (2022) (Dec. 2019).

403. Marra, F., Gragnaniello, D., Cozzolino, D. & Verdoliva, L. *Detection of GAN-Generated Fake Images over Social Networks* en. in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (Ieee, Miami, FL, Apr. 2018), 384–389. ISBN: 978-1-5386-1857-8. `https://ieeexplore.ieee.org/document/8397040/` (2020).

404. Martel, C., Pennycook, G. & Rand, D. G. Reliance on emotion promotes belief in fake news. en, 50 (2019).

405. Martin, D., Hanrahan, B. V., O'Neill, J. & Gupta, N. *Being a turker* in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (2014), 224–235.

406. Masuda, T. *et al.* Placing the face in context: Cultural differences in the perception of facial emotion. en. *Journal of Personality and Social Psychology* **94,** 365–381. ISSN: 1939-1315, 0022-3514. `http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.94.3.365` (2021) (3 Mar. 2008).

407. Matsumoto, D., Hwang, H. S. & Yamada, H. Cultural Differences in the Relative Contributions of Face and Context to Judgments of Emotions. en. *Journal of Cross-Cultural Psychology* **43,** 198–218. ISSN: 0022-0221, 1552-5422. `http://journals.sagepub.com/doi/10.1177/0022022110387426` (2021) (2 Feb. 2012).

408. Matsumoto, D. & Sung Hwang, H. Judging Faces in Context: Faces in Context. en. *Social and Personality Psychology Compass* **4,** 393–402. ISSN: 17519004. `https://onlinelibrary.wiley.com/doi/10.1111/j.1751-9004.2010.00271.x` (2021) (6 June 2010).

409. Mauss, I. B. & Robinson, M. D. Measures of emotion: A review. *Cognition and emotion* **23,** 209–237 (2 2009).

410. Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P. & Cohn, J. F. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing* **4,** 151–160 (2013).

411. Mayor, A. *Gods and robots: Myths, machines, and ancient dreams of technology* (Princeton University Press, 2019).

412. McCarthy, J. *Chess as the Drosophila of AI* in *Computers, chess, and cognition* (1990), 227–237.

413. McDuff, D., El Kaliouby, R., Kodra, E. & Picard, R. *Measuring Voter's Candidate Preference Based on Affective Responses to Election Debates* en. in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (Ieee, Geneva, Switzerland, Sept. 2013), 369–374. ISBN: 978-0-7695-5048-0. `http://ieeexplore.ieee.org/document/6681458/` (2020).

414. McDuff, D., Gontarek, S. & Picard, R. *Remote measurement of cognitive stress via heart rate variability* in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2014), 2957–2960.

415. McDuff, D., Hernandez, J., Gontarek, S. & Picard, R. W. *Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera* in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 4000–4004.

416. McGrath, T. *et al.* Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences* **119,** e2206625119 (2022).

417. McIlroy-Young, R., Wang, Y., Sen, S., Kleinberg, J. & Anderson, A. Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in Chess. *Advances in Neural Information Processing Systems* (2021).

418. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. en. *Nature* **577,** 89–94. ISSN: 0028-0836, 1476-4687. http://www.nature.com/articles/s41586-019-1799-6 (2020) (Jan. 2020).

419. Mednick, M. T., Mednick, S. A. & Mednick, E. V. Incubation of creative performance and specific associative priming. *The Journal of Abnormal and Social Psychology* **69,** 84 (1964).

420. Mell, J., Beissinger, M. & Gratch, J. An expert-model and machine learning hybrid approach to predicting human-agent negotiation outcomes in varied data. en. *Journal on Multimodal User Interfaces* **15,** 215–227. ISSN: 1783-7677, 1783-8738. https://link.springer.com/10.1007/s12193-021-00368-w (2021) (2 June 2021).

421. Mell, J., Lucas, G. M. & Gratch, J. *Welcome to the real world: How agent strategy increases human willingness to deceive* in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (2018), 1250–1257.

422. Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R. & Rozeira, J. *PH 2-A dermoscopic image database for research and benchmarking* in *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (2013), 5437–5440.

423. Merler, M., Ratha, N., Feris, R. S. & Smith, J. R. *Diversity in Faces* 2019. arXiv: 1901.10436 [cs.CV].

424. Messaris, P. *Visual persuasion: The role of images in advertising* (Sage, 1997).

425. Messaris, P. & Abraham, L. The Role of Images in Framing News Stories. en, 13 (2001).

426. Metzger, M. J., Flanagin, A. J. & Medders, R. B. Social and heuristic approaches to credibility evaluation online. *Journal of communication* **60,** 413–439 (2010).

427. Miceli, M., Posada, J. & Yang, T. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? en. *Proceedings of the ACM on Human-Computer Interaction* **6,** 1–14. ISSN: 2573-0142. https://dl.acm.org/doi/10.1145/3492853 (2022) (Jan. 2022).

428. Miceli, M., Schuessler, M. & Yang, T. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. en. *Proceedings of the ACM on Human-Computer Interaction* **4,** 1–25. ISSN: 2573-0142. https://dl.acm.org/doi/10.1145/3415186 (2022) (Oct. 2020).

429. Miceli, M. *et al. Documenting computer vision datasets: an invitation to reflexive data practices* in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), 161–172.

430. Micheletti, R. G., James, W. D., Elston, D. & McMahon, P. J. *Andrews' Diseases of the Skin Clinical Atlas, E-Book* (Elsevier Health Sciences, 2022).

431. Mills, C., D'Mello, S., Bosch, N. & Olney, A. M. *Mind wandering during learning with an intelligent tutoring system* in *International conference on artificial intelligence in education* (2015), 267–276.

432. Mirsky, Y. & Lee, W. The Creation and Detection of Deepfakes: A Survey. en. *arXiv:2004.11138 [cs, eess].* arXiv: 2004.11138. `http://arxiv.org/abs/2004.11138` (2020) (Sept. 2020).

433. Mitchell, M. *et al. Model cards for model reporting* in *Proceedings of the conference on fairness, accountability, and transparency* (2019), 220–229.

434. Mitchell, M. Why AI is Harder Than We Think. en. *arXiv:2104.12871 [cs].* arXiv: 2104.12871. `http://arxiv.org/abs/2104.12871` (2021) (Apr. 2021).

435. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A. & Manocha, D. *Emotions Don't Lie: A Deepfake Detection Method using Audio-Visual Affective Cues* 2020.

436. Mollahosseini, A., Chan, D. & Mahoor, M. H. *Going deeper in facial expression recognition using deep neural networks* en. in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Ieee, Lake Placid, NY, USA, Mar. 2016), 1–10. ISBN: 978-1-5090-0641-0. `http://ieeexplore.ieee.org/document/7477450/` (2022).

437. Monk Jr, E. P. The cost of color: Skin color, discrimination, and health among African-Americans. *American Journal of Sociology* **121,** 396–444 (2015).

438. Moreno, G., Tran, H., Chia, A. L., Lim, A. & Shumack, S. Prospective study to assess general practitioners' dermatological diagnostic skills in a referral setting. *Australasian journal of dermatology* **48,** 77–82 (2007).

439. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodrıéguez, R., Chawla, N. V. & Herrera, F. A unifying view on dataset shift in classification. en. *Pattern recognition* **45,** 521–530. ISSN: 00313203. `https://linkinghub.elsevier.com/retrieve/pii/S0031320311002901` (2022) (Jan. 2012).

440. Moridis, C. N. & Economides, A. A. Affective learning: Empathetic agents with emotional facial and tone of voice expressions. *IEEE Transactions on Affective Computing* **3,** 260–272 (2012).

441. Mosleh, M., Pennycook, G., Arechar, A. A. & Rand, D. G. Cognitive reflection correlates with behavior on Twitter. *Nature communications* **12,** 1–10 (2021).

442. Mota, S. & Picard, R. W. *Automated posture analysis for detecting learner's interest level* in. **5** (2003), 49.

443. Mullainathan, S. & Obermeyer, Z. Does Machine Learning Automate Moral Hazard and Error? en. **107,** 5 (2017).

444. Mullainathan, S. & Obermeyer, Z. *On the Inequity of Predicting A While Hoping for B* en. in *AEA Papers and Proceedings* **111** (May 2021), 37–42. `https://pubs.aeaweb.org/doi/10.1257/pandp.20211078` (2021).

445. Mullainathan, S. & Obermeyer, Z. Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics* **137,** 679–727 (2022).

446. Muller, M. *et al. How data science workers work with data: Discovery, capture, curation, design, creation* en. in *Proceedings of the 2019 CHI conference on human factors in computing systems* (Acm, Glasgow Scotland Uk, May 2019), 1–15. ISBN: 978-1-4503-5970-2. `https://dl.acm.org/doi/10.1145/3290605.3300356` (2022).

447. Muller, M. *et al. Designing Ground Truth and the Social Life of Labels* en. in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Acm, Yokohama Japan, May 2021), 1–16. ISBN: 978-1-4503-8096-6. `https://dl.acm.org/doi/10.1145/3411764.3445402` (2022).

448. Mumenthaler, C., Sander, D. & Manstead, A. Emotion recognition in simulated social interactions. *IEEE Transactions on Affective computing* (2018).

449. Murali, P., Trinh, H., Ring, L. & Bickmore, T. *A Friendly Face in the Crowd: Reducing Public Speaking Anxiety with an Emotional Support Agent in the Audience* in (2021), 156–163.

450. Muse, E. D. *et al.* From second to hundredth opinion in medicine: A global consultation platform for physicians. *NPJ Digital Medicine* **1,** 55 (2018).

451. Nadarevic, L., Reber, R., Helmecke, A. J. & Köse, D. Perceived truth of statements and simulated social media postings: an experimental investigation of source credibility, repeated exposure, and presentation format. *Cognitive Research: Principles and Implications* **5,** 1–16 (2020).

452. Nader, L. Up the anthropologist: Perspectives gained from studying up. (1972).

453. Narain, J. *et al.* Personalized Modeling of Real-World Vocalizations fromNonverbal Individuals. *Proceedings of the 2020 International Conference on Multimodal Interaction* (2020).

454. Narayanan, V., Manoghar, B. M., Dorbala, V. S., Manocha, D. & Bera, A. ProxEmo: Gait-based Emotion Learning and Multi-view Proxemic Fusion for Socially-Aware Robot Navigation. *arXiv preprint arXiv:2003.01062* (2020).

455. Newell, A., Shaw, J. C. & Simon, H. A. Elements of a theory of human problem solving. *Psychological review* **65,** 151 (1958).

456. Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J. & Lindsay, D. S. Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review* **19,** 969–974 (2012).

457. Newman, E. J., Jalbert, M. C., Schwarz, N. & Ly, D. P. Truthiness, the illusory truth effect, and the role of need for cognition. en. *Consciousness and Cognition* **78,** 102866. ISSN: 10538100. `https://linkinghub.elsevier.com/retrieve/pii/S1053810019301977` (2022) (Feb. 2020).

458. Nichol, A. *et al.* Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

459. Nightingale, S. J. & Farid, H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. en. *Proceedings of the National Academy of Sciences* **119,** e2120481119. ISSN: 0027-8424, 1091-6490. `https://pnas.org/doi/full/10.1073/pnas.2120481119` (2022) (Feb. 2022).

460. Nightingale, S. J., Wade, K. A., Farid, H. & Watson, D. G. Can people detect errors in shadows and reflections? en. *Attention, Perception, & Psychophysics* **81,** 2917–2943. ISSN: 1943-3921, 1943-393x. `http://link.springer.com/10.3758/s13414-019-01773-w` (2020) (Nov. 2019).

461. Nightingale, S. J., Wade, K. A. & Watson, D. G. Can people identify original and manipulated photos of real-world scenes? en. *Cognitive Research: Principles and Implications* **2,** 30. ISSN: 2365-7464. `http://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-017-0067-2` (2020) (Dec. 2017).

462. Nirkin, Y., Keller, Y. & Hassner, T. *Fsgan: Subject agnostic face swapping and reenactment* in *Proceedings of the IEEE/CVF international conference on computer vision* (2019), 7184–7193.

463. Norgeot, B. *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature medicine* **26,** 1320–1324 (2020).

464. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint arXiv:2303.13375* (2023).

465. Northcutt, C. G., Athalye, A. & Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. en. *arXiv preprint arXiv:2103.14749.* arXiv: 2103.14749. `http://arxiv.org/abs/2103.14749` (2021) (Apr. 2021).

466. Nouri, K., Ballard, C., Patel, A. & Brasie, R. Basal cell carcinoma. *Nouri K, et al. Skin Cancer. McGraw Hill Medical, China,* 61–81 (2008).

467. O'Brien, S., Groh, M. & Dubey, A. Evaluating Generative Adversarial Networks on Explicitly Parameterized Distributions. en. *arXiv:1812.10782 [cs, stat].* arXiv: 1812.10782. `http://arxiv.org/abs/1812.10782` (2021) (Dec. 2018).

468. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Re, C. *Hidden stratification causes clinically meaningful failures in machine learning for medical imaging* en. in *Proceedings of the ACM Conference on Health, Inference, and Learning* (Acm, Toronto Ontario Canada, Apr. 2020), 151–159. ISBN: 978-1-4503-7046-2. `https://dl.acm.org/doi/10.1145/3368555.3384468` (2020).

469. Oakley, A. & NZ, N. *Dermatology Made Easy* 2017. `https://books.google.com/books?id=ETaAEAAAQBAJ`.

470. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. en. *Science* **366,** 447–453. ISSN: 0036-8075, 1095-9203. `https://www.sciencemag.org/lookup/doi/10.1126/science.aax2342` (2020) (Oct. 2019).

471. Oh, Y., Markova, A., Noor, S. & Rotemberg, V. Standardized clinical photography considerations in patients across skin tones. en. *British Journal of Dermatology,* bjd.20766. ISSN: 0007-0963, 1365-2133. `https://onlinelibrary.wiley.com/doi/10.1111/bjd.20766` (2022) (Nov. 2021).

472. Okoji, U., Taylor, S. & Lipoff, J. Equity in skin typing: why it is time to replace the Fitzpatrick scale. en. *British Journal of Dermatology* **185,** 198–199. ISSN: 0007-0963, 1365-2133. `https://onlinelibrary.wiley.com/doi/10.1111/bjd.19932` (2021) (July 2021).

473. Oliva, A. & Torralba, A. The role of context in object recognition. en. *Trends in Cognitive Sciences* **11,** 520–527. ISSN: 13646613. `https://linkinghub.elsevier.com/retrieve/pii/S1364661307002550` (2020) (Dec. 2007).

474. Ong, D. C., Soh, H., Zaki, J. & Goodman, N. D. Applying Probabilistic Programming to Affective Computing. en. *IEEE Transactions on Affective Computing.* arXiv: 1903.06445, 1–1. ISSN: 1949-3045, 2371-9850. `http://arxiv.org/abs/1903.06445` (2020) (2019).

475. Ong, D. C., Zaki, J. & Goodman, N. D. Affective cognition: Exploring lay theories of emotion. en. *Cognition* **143,** 141–162. ISSN: 00100277. `https://linkinghub.elsevier.com/retrieve/pii/S0010027715300196` (2021) (Oct. 2015).

476. Ørting, S. *et al.* A survey of crowdsourcing in medical image analysis. *arXiv preprint arXiv:1902.09159* (2019).

477. Osto, M., Hamzavi, I. H., Lim, H. W. & Kohli, I. Individual Typology Angle and Fitzpatrick Skin Phototypes are Not Equivalent in Photodermatology. *Photochemistry and photobiology* **98,** 127–129 (2022).

478. Pacheco, A. G. *et al.* PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief* **32,** 106221 (2020).

479. Paiva, A., Leite, I., Boukricha, H. & Wachsmuth, I. Empathy in Virtual Agents and Robots: A Survey. en. *ACM Transactions on Interactive Intelligent Systems* **7,** 1–40. ISSN: 2160-6455, 2160-6463. `https://dl.acm.org/doi/10.1145/2912150` (2022) (Oct. 2017).

480. Palan, S. & Schitter, C. Prolific.ac—A subject pool for online experiments. en. *Journal of Behavioral and Experimental Finance* **17,** 22–27. ISSN: 22146350. `https://linkinghub.elsevier.com/retrieve/pii/S2214635017300989` (2021) (Mar. 2018).

481. Palmieri, J. R. Missed Diagnosis and the Development of Acute and Late Lyme Disease in Dark Skinned Populations of Appalachia. en. *Biomedical Journal of Scientific & Technical Research* **21.** ISSN: 25741241. `https://biomedres.us/fulltexts/BJSTR.MS.ID.003583.php` (2020) (Sept. 2019).

482. Pantic, M., Valstar, M., Rademaker, R. & Maat, L. *Web-based database for facial expression analysis* in *2005 IEEE international conference on multimedia and Expo* (2005), 5–pp.

483. Paolacci, G., Chandler, J. & Ipeirotis, P. G. Running experiments on amazon mechanical turk. *Judgment and Decision making* **5,** 411–419 (2010).

484. Paris, B. & Donovan, J. *Deepfakes and cheapfakes* en. 2019.

485. Parmar, D., Olafsson, S., Utami, D., Murali, P. & Bickmore, T. Designing empathic virtual agents: manipulating animation, voice, rendering, and empathy to create persuasive agents. *Autonomous Agents and Multi-Agent Systems* **36,** 1–24 (2022).

486. Pataranutaporn, P. *et al.* AI-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence* **3,** 1013–1022 (2021).

487. Patel, B. N. *et al.* Human–machine partnership with artificial intelligence for chest radiograph diagnosis. en. *NPJ digital medicine* **2,** 1–10. ISSN: 2398-6352. `http://www.nature.com/articles/s41746-019-0189-7` (2021) (Dec. 2019).

488. Paullada, A., Raji, I. D., Bender, E. M., Denton, E. & Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* **2,** 100336 (2021).

489. Peirce, C. S. *Peirce on signs: Writings on semiotic* (UNC Press Books, 1991).

490. Pelz, M. C., Allen, K. R., Tenenbaum, J. B. & Schulz, L. E. Foundations of intuitive power analyses in children and adults. *Nature Human Behaviour,* 1–12 (2022).

491. Peña, A., Fierrez, J., Lapedriza, A. & Morales, A. Learning Emotional-Blinded Face Representations. *arXiv preprint arXiv:2009.08704* (2020).

492. Pennycook, G., Cheyne, J. A., Koehler, D. J. & Fugelsang, J. A. Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods* **48,** 341–348 (2016).

493. Pennycook, G. & Rand, D. G. Fighting misinformation on social media using crowd-sourced judgments of news source quality. en. *Proceedings of the National Academy of Sciences* **116,** 2521–2526. ISSN: 0027-8424, 1091-6490. `http://www.pnas.org/lookup/doi/10.1073/pnas.1806781116` (2021) (Feb. 2019).

494. Pennycook, G. & Rand, D. G. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality* **88,** 185–200 (2020).

495. Pennycook, G. & Rand, D. G. *The psychology of fake news* 2021.

496. Pennycook, G. & Rand, D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. en. *Cognition* **188,** 39–50. ISSN: 00100277. `https://linkinghub.elsevier.com/retrieve/pii/S001002771830163X` (2021) (July 2019).

497. Pennycook, G. *et al. Shifting attention to accuracy can reduce misinformation online* en. Apr. 2021. `http://www.nature.com/articles/s41586-021-03344-2` (2021).

498. Phillips, M. *et al.* Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA network open* **2,** e1913436–e1913436 (2019).

499. Phillips, P. J. *et al.* Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. en. *Proceedings of the National Academy of Sciences* **115,** 6171–6176. ISSN: 0027-8424, 1091-6490. `http://www.pnas.org/lookup/doi/10.1073/pnas.1721355115` (2021) (June 2018).

500. Picard, R. W. *Affective computing* (MIT press, 1995).

501. Picard, R. W., Vyzas, E. & Healey, J. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence* **23,** 1175–1191 (10 2001).

502. Picard, R. W., Fedor, S. & Ayzenberg, Y. Multiple Arousal Theory and Daily-Life Electrodermal Activity Asymmetry. en. *Emotion Review* **8,** 62–75. ISSN: 1754-0739, 1754-0747. `http://journals.sagepub.com/doi/10.1177/1754073914565517` (2020) (Jan. 2016).

503. Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S. & Obermeyer, Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. en. *Nature Medicine* **27,** 136–140. ISSN: 1078-8956, 1546-170x. http://www.nature.com/articles/s41591-020-01192-7 (2021) (1 Jan. 2021).

504. Poh, M.-Z., McDuff, D. & Picard, R. W. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering* **58,** 7–11 (2010).

505. Pollak, S. D. & Kistler, D. J. Early experience is associated with the development of categorical representations for facial expressions of emotion. en. *Proceedings of the National Academy of Sciences* **99,** 9072–9076. ISSN: 0027-8424, 1091-6490. http://www.pnas.org/cgi/doi/10.1073/pnas.142165999 (2021) (June 2002).

506. Pollak, S. D., Messner, M., Kistler, D. J. & Cohn, J. F. Development of perceptual expertise in emotion recognition. *Cognition* **110,** 242–247 (2 2009).

507. Poria, S., Cambria, E., Bajpai, R. & Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. en. *Information Fusion* **37,** 98–125. ISSN: 15662535. https://linkinghub.elsevier.com/retrieve/pii/S1566253517300738 (2020) (Sept. 2017).

508. Poria, S. *et al. Context-dependent sentiment analysis in user-generated videos* in. **37** (Elsevier, 2017), 873–883.

509. Powell, T. E., Boomgaarden, H. G., De Swert, K. & de Vreese, C. H. Video killed the news article? Comparing multimodal framing effects in news videos and articles. *Journal of broadcasting & electronic media* **62,** 578–596 (2018).

510. Prajwal, K., Mukhopadhyay, R., Namboodiri, V. P. & Jawahar, C. *A lip sync expert is all you need for speech to lip generation in the wild* en. arXiv: 2008.10010. Seattle, WA, USA, Oct. 2020. http://arxiv.org/abs/2008.10010 (2021).

511. Prelec, D., Seung, H. S. & McCoy, J. A solution to the single-question crowd wisdom problem. en. *Nature* **541,** 532–535. ISSN: 0028-0836, 1476-4687. http://www.nature.com/articles/nature21054 (2021) (Jan. 2017).

512. Preston, C. C. & Colman, A. M. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica* **104,** 1–15 (2000).

513. Qi, H. *et al.* DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. en. *arXiv:2006.07634 [cs].* arXiv: 2006.07634. http://arxiv.org/abs/2006.07634 (2020) (Aug. 2020).

514. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. D. *Dataset shift in machine learning* (2008).

515. Raghu, M. *et al.* Direct Uncertainty Prediction for Medical Second Opinions. en, 10.

516. Rahwan, I. *et al.* Machine behaviour. en. *Nature* **568,** 477–486. ISSN: 0028-0836, 1476-4687. http://www.nature.com/articles/s41586-019-1138-y (2020) (Apr. 2019).

517. Raji, I. D., Bender, E. M., Paullada, A., Denton, E. & Hanna, A. AI and the Everything in the Whole Wide World Benchmark. en. *arXiv:2111.15366 [cs].* arXiv: 2111.15366. http://arxiv.org/abs/2111.15366 (2021) (Nov. 2021).

518. Raji, I. D. *et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing* in (2020), 33–44.

519. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nature medicine* **28,** 31–38 (2022).

520. Rajpurkar, P. *et al.* Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).

521. Ramezani, M., Karimian, A. & Moallem, P. Automatic detection of malignant melanoma using macroscopic images. *Journal of medical signals and sensors* **4,** 281 (2014).

522. Rand, D. G. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of theoretical biology* **299,** 172–179 (2012).

523. Recht, B., Roelofs, R., Schmidt, L. & Shankar, V. *Do imagenet classifiers generalize to imagenet?* in *International Conference on Machine Learning* (2019), 5389–5400.

524. Reeves, B., Yeykelis, L. & Cummings, J. J. The Use of Media in Media Psychology. en. *Media Psychology* **19,** 49–71. ISSN: 1521-3269, 1532-785x. http://www.tandfonline.com/doi/full/10.1080/15213269.2015.1030083 (2021) (Jan. 2016).

525. *Regionalderm.com* https://www.regionalderm.com/contact_info.html. [Accessed 17-Feb-2023].

526. Reid, V. M. *et al.* The human fetus preferentially engages with face-like visual stimuli. *Current Biology* **27,** 1825–1828 (2017).

527. Reschke, P. J., Knothe, J. M., Lopez, L. D. & Walle, E. A. Putting "context" in context: The effects of body posture and emotion scene on adult categorizations of disgust facial expressions. en. *Emotion* **18,** 153–158. ISSN: 1931-1516, 1528-3542. http://doi.apa.org/getdoi.cfm?doi=10.1037/emo0000350 (2020) (Feb. 2018).

528. Reschke, P. J., Walle, E. A., Knothe, J. M. & Lopez, L. D. The influence of context on distinct facial expressions of disgust. en. *Emotion* **19,** 365–370. ISSN: 1931-1516, 1528-3542. http://doi.apa.org/getdoi.cfm?doi=10.1037/emo0000445 (2020) (Mar. 2019).

529. Rhodes, G., Brake, S. & Atkinson, A. P. What's lost in inverted faces? *Cognition* **47,** 25–57 (1993).

530. Ribers, M. A. & Ullrich, H. Machine Predictions and Human Decisions with Variation in Payoffs and Skills. en. *SSRN Electronic Journal.* ISSN: 1556-5068. https://www.ssrn.com/abstract=3726018 (2021) (2020).

531. Richler, J. J., Cheung, O. S. & Gauthier, I. Holistic processing predicts face recognition. *Psychological Science* **22,** 464–471 (2011).

532. Richler, J. J. & Gauthier, I. A Meta-Analysis and Review of Holistic Face Processing. en. *Psychological Bulletin* **140,** 1281–1302. ISSN: 1939-1455, 0033-2909. http://doi.apa.org/getdoi.cfm?doi=10.1037/a0037004 (2021) (2014).

533. Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S. & Lewandowsky, S. Psychological inoculation improves resilience against misinformation on social media. *Science Advances* **8,** eabo6254 (2022).

534. Rossler, A. *et al. Faceforensics++: Learning to detect manipulated facial images* 2019.

535. Rössler, A. *et al. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces* en. arXiv: 1803.09179. Mar. 2018. `http://arxiv.org/abs/1803.09179` (2021).

536. Rotemberg, V. *et al.* A patient-centric dataset of images and metadata for identifying melanomas using clinical context. en. *Scientific data* **8,** 1–8. ISSN: 2052-4463. `http://www.nature.com/articles/s41597-021-00815-z` (2022) (Dec. 2021).

537. Roy, A. G. *et al.* Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. en. *Medical Image Analysis* **75.** arXiv: 2104.03829, 102274. `http://arxiv.org/abs/2104.03829` (2021) (Apr. 2022).

538. Ruba, A. L. & Pollak, S. D. The Development of Emotion Reasoning in Infancy and Early Childhood. en. *Annual Review of Developmental Psychology* **2,** 503–531. ISSN: 2640-7922, 2640-7922. `https://www.annualreviews.org/doi/10.1146/annurev-devpsych-060320-102556` (2021) (Dec. 2020).

539. Rudovic, O., Lee, J., Dai, M., Schuller, B. & Picard, R. W. Personalized machine learning for robot perception of affect and engagement in autism therapy. en. *Science Robotics* **3,** eaao6760. ISSN: 2470-9476. `https://robotics.sciencemag.org/lookup/doi/10.1126/scirobotics.aao6760` (2020) (June 2018).

540. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *International journal of computer vision* **115,** 211–252 (2015).

541. Russell, J. A. A circumplex model of affect. *Journal of personality and social psychology* (1980).

542. Russell, J. A. & Barrett, L. F. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology* **76,** 805 (1999).

543. Russell, J. A. Core affect and the psychological construction of emotion. en. *Psychological Review* **110,** 145–172. ISSN: 1939-1471, 0033-295x. `http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.110.1.145` (2021) (2003).

544. Ryan, R. M., Rigby, C. S. & Przybylski, A. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion* (2006).

545. Sachdeva, M., Price, K. N., Hsiao, J. L. & Shi, V. Y. Gender and rank salary trends among academic dermatologists. *International Journal of Women's Dermatology* **6,** 324 (2020).

546. Sagawa, S., Koh, P. W., Hashimoto, T. B. & Liang, P. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. en. *arXiv:1911.08731 [cs, stat].* arXiv: 1911.08731. `http://arxiv.org/abs/1911.08731` (2021) (Apr. 2020).

547. Sagers, L. W. *et al. Improving dermatology classifiers across populations using images generated by large diffusion models* in *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research* ().

548. Salzman, H. The Color Atlas and Synopsis of Family Medicine. *Family Medicine* **52,** 226–227 (2020).

549. Sambasivan, N. *et al. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI* en. in *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Acm, Yokohama Japan, May 2021), 1–15. ISBN: 978-1-4503-8096-6. `https://dl.acm.org/doi/10.1145/3411764.3445518` (2022).

550. Sankaranarayanan, A., Groh, M., Picard, R. & Lippman, A. The Presidential Deepfakes Dataset. en, 16 (2021).

551. Santurkar, S., Tsipras, D. & Madry, A. *BREEDS: Benchmarks for Subpopulation Shift* en. Number: arXiv:2008.04859 arXiv:2008.04859 [cs, stat]. Aug. 2020. `http://arxiv.org/abs/2008.04859` (2022).

552. Scarantino, A. How to do things with emotional expressions: The theory of affective pragmatics. *Psychological Inquiry* **28,** 165–185 (2017).

553. Scheuerman, M. K., Hanna, A. & Denton, E. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. en. *Proceedings of the ACM on Human-Computer Interaction* **5,** 1–37. ISSN: 2573-0142. `https://dl.acm.org/doi/10.1145/3476058` (2021) (Oct. 2021).

554. Scheuerman, M. K., Wade, K., Lustig, C. & Brubaker, J. R. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. en. *Proceedings of the ACM on Human-Computer Interaction* **4,** 1–35. ISSN: 2573-0142. `https://dl.acm.org/doi/10.1145/3392866` (2021) (May 2020).

555. Schneider, K. & Josephs, I. The expressive and communicative functions of preschool children's smiles in an achievement-situation. en. *Journal of Nonverbal Behavior* **15,** 185–198. ISSN: 0191-5886, 1573-3653. `http://link.springer.com/10.1007/BF01672220` (2021) (Sept. 1991).

556. Schoth, D. E. & Liossi, C. A systematic review of experimental paradigms for exploring biased interpretation of ambiguous information with emotional and neutral associations. *Frontiers in psychology* **8,** 171 (2017).

557. Schrittwieser, J. *et al.* Mastering Atari, Go, chess and shogi by planning with a learned model. en. *Nature* **588,** 604–609. ISSN: 0028-0836, 1476-4687. `http://www.nature.com/articles/s41586-020-03051-4` (2021) (Dec. 2020).

558. Schroeder, J. & Epley, N. Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General* **145,** 1427 (2016).

559. Schwarz, N. Feelings-as-information theory. *Handbook of theories of social psychology* **1,** 289–308 (2011).

560. Seferbekov, S. *Deepfake Detection Challenge Submission* `https://github.com/selimsef/dfdc_deepfake_challenge`. 2020.

561. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. en. *Nature Medicine* **27,** 2176–2182. ISSN: 1078-8956, 1546-170x. `https://www.nature.com/articles/s41591-021-01595-0` (2022) (Dec. 2021).

562. Shankar, S., Garcia, R., Hellerstein, J. M. & Parameswaran, A. G. *Operationalizing Machine Learning: An Interview Study* 2022. `https://arxiv.org/abs/2209.09125`.

563. Shen, C. *et al.* Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New media & society* **21,** 438–463 (2019).

564. Shen, H. *et al.* Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. en. *Proceedings of the ACM on Human-Computer Interaction* **4,** 1–22. ISSN: 2573-0142. `https://dl.acm.org/doi/10.1145/3415224` (2022) (Oct. 2020).

565. Shen, J. H., Lapedriza, A. & Picard, R. W. *Unintentional affective priming during labeling may bias labels* in (2019), 587–593.

566. Shin, M., Kim, J., van Opheusden, B. & Griffiths, T. L. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences* **120,** e2214840120 (2023).

567. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. en. *Nature* **529,** 484–489. ISSN: 0028-0836, 1476-4687. `http://www.nature.com/articles/nature16961` (2020) (Jan. 2016).

568. Silver, D. *et al.* Mastering the game of Go without human knowledge. en. *Nature* **550,** 354–359. ISSN: 0028-0836, 1476-4687. `http://www.nature.com/articles/nature24270` (2020) (Oct. 2017).

569. Silver, D. *et al.* A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. en. *Science* **362,** 1140–1144. ISSN: 0036-8075, 1095-9203. `https://www.sciencemag.org/lookup/doi/10.1126/science.aar6404` (2021) (Dec. 2018).

570. Simon, H. & Chase, W. Skill in chess. *Computer chess compendium,* 175–188 (1973).

571. Simon, H. A. & Newell, A. Human problem solving: The state of the theory in 1970. en. *American Psychologist* **26,** 145–159. ISSN: 0003-066x. `http://content.apa.org/journals/amp/26/2/145` (2020) (1971).

572. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

573. Singh, M. & Venkataramani, A. *Capacity Strain and Racial Disparities in Hospital Mortality* tech. rep. (National Bureau of Economic Research, 2022).

574. Sinha, P., Balas, B., Ostrovsky, Y. & Russell, R. Face recognition by humans: Nineteen results all computer vision researchers should know about. en. *Proceedings of the IEEE* **94,** 1948–1962. ISSN: 0018-9219, 1558-2256. `http://ieeexplore.ieee.org/document/4052483/` (2020) (Nov. 2006).

575. Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M. & Perer, A. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. *arXiv preprint arXiv:2302.00096* (2023).

576. Small, D. A. & Lerner, J. S. Emotional policy: Personal sadness and anger shape judgments about a welfare case. en. *Political Psychology* **29,** 149–168. ISSN: 0162895x, 14679221. `http://doi.wiley.com/10.1111/j.1467-9221.2008.00621.x` (2021) (Apr. 2008).

577. Smith, J. J., Amershi, S., Barocas, S., Wallach, H. & Wortman Vaughan, J. *REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research* en. in *2022 ACM Conference on Fairness, Accountability, and Transparency* (Acm, Seoul Republic of Korea, June 2022), 587–597. ISBN: 978-1-4503-9352-2. `https://dl.acm.org/doi/10.1145/3531146.3533122` (2022).

578. Smith, P. B. Acquiescent Response Bias as an Aspect of Cultural Communication Style. en. *Journal of Cross-Cultural Psychology* **35,** 50–61. ISSN: 0022-0221, 1552-5422. `http://journals.sagepub.com/doi/10.1177/0022022103260380` (2021) (1 Jan. 2004).

579. Snoswell, C., Finnane, A., Janda, M., Soyer, H. P. & Whitty, J. A. Cost-effectiveness of store-and-forward teledermatology: a systematic review. *JAMA dermatology* **152,** 702–708 (2016).

580. Soenksen, L. R. *et al.* Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Science Translational Medicine* **13** (2021).

581. Spreng*, R. N., McKinnon*, M. C., Mar, R. A. & Levine, B. The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of personality assessment* (2009).

582. Srinivasan, R. & Chander, A. Biases in AI systems. *Communications of the ACM* **64,** 44–49 (2021).

583. Stagnaro, M., Pennycook, G. & Rand, D. G. Performance on the Cognitive Reflection Test is stable across time. *Stagnaro, MN, Pennycook, G., & Rand, DG (2018) Performance on the Cognitive Reflection Test is stable across time. Judgment and Decision Making* **13,** 260–267 (2018).

584. Stratou, G., Ghosh, A., Debevec, P. & Morency, L.-P. *Effect of illumination on automatic expression recognition: a novel 3D relightable facial database* in *Face and Gesture 2011* (2011), 611–618.

585. Sun, X., Yang, J., Sun, M. & Wang, K. *A benchmark for automatic visual classification of clinical skin disease images* in *European Conference on Computer Vision* (2016), 206–222.

586. Sun, Y. *et al.* Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137* (2021).

587. Sundar, S. S. The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital Media, Youth, and Credibility* (eds Metzger, M. J. & Flanagin, A. J.) (2008).

588. Sundar, S. S., Molina, M. D. & Cho, E. Seeing Is Believing: Is Video Modality More Powerful in Spreading Fake News via Online Messaging Apps? en. *Journal of Computer-Mediated Communication* **26,** 301–319. ISSN: 1083-6101. `https://academic.oup.com/jcmc/article/26/6/301/6336055` (2021) (Nov. 2021).

589. Suresh, H. & Guttag, J. V. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. en. *arXiv:1901.10002 [cs, stat].* arXiv: 1901.10002, 1–9. `http://arxiv.org/abs/1901.10002` (2021) (June 2021).

590. Surowiecki, J. *The wisdom of crowds. Anchor* 2005.

591. Suvorov, R. *et al. Resolution-robust Large Mask Inpainting with Fourier Convolutions* in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022), 2149–2159.

592. Suwajanakorn, S., Seitz, S. M. & Kemelmacher-Shlizerman, I. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)* **36,** 1–13 (2017).

593. Swinney, C. C., Han, D. P. & Karth, P. A. Incontinentia pigmenti: a comprehensive review and update. *Ophthalmic Surgery, Lasers and Imaging Retina* **46,** 650–657 (2015).

594. Syed, O. The arimaa challenge: From inception to completion. *ICGA Journal* **38,** 3–11 (2015).

595. Tahir, R. *et al. Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos* en. in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Acm, Yokohama Japan, May 2021), 1–16. ISBN: 978-1-4503-8096-6. `https://dl.acm.org/doi/10.1145/3411764.3445699` (2021).

596. Taleb, N. N. *The black swan: The impact of the highly improbable* (Random house, 2007).

597. Tan, M. & Le, Q. *EfficientNet: Rethinking model scaling for convolutional neural networks* in *International Conference on Machine Learning* (2019), 6105–6114.

598. Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J. & Schiebinger, L. Sex and gender analysis improves science and engineering. *Nature* (2019).

599. Taylor, S., Jaques, N., Nosakhare, E., Sano, A. & Picard, R. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing* (2017).

600. Taylor, S. *et al.* Automatic identification of artifacts in electrodermal activity data. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* 1934–1937 (2015).

601. Ternovski, J., Kalla, J. & Aronow, P. M. *Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments* en. preprint. Jan. 2021. `https://osf.io/dta97` (2021).

602. Thomas, R. L. & Uminsky, D. Reliance on metrics is a fundamental challenge for AI. en. *Patterns* **3,** 100476. ISSN: 26663899. `https://linkinghub.elsevier.com/retrieve/pii/S2666389922000563` (2022) (May 2022).

603. Thornton, M. A. & Tamir, D. I. Mental models accurately predict emotion transitions. en. *Proceedings of the National Academy of Sciences* **114,** 5982–5987. ISSN: 0027-8424, 1091-6490. `http://www.pnas.org/lookup/doi/10.1073/pnas.1616056114` (2021) (23 June 2017).

604. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. & Ortega-Garcia, J. Deep-Fakes and Beyond: A Survey of Face Manipulation and Fake Detection. en. *arXiv:2001.00179 [cs].* arXiv: 2001.00179. `http://arxiv.org/abs/2001.00179` (2020) (June 2020).

605. Torralba, A. & Efros, A. A. *Unbiased look at dataset bias* in *Cvpr 2011* (2011), 1521–1528.

606. Tran, H., Chen, K., Lim, A. C., Jabbour, J. & Shumack, S. Assessing diagnostic skill in dermatology: a comparison between general practitioners and dermatologists. *Australasian journal of dermatology* **46,** 230–234 (2005).

607. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* **5,** 1–9 (2018).

608. Tschandl, P. *et al.* Human–computer collaboration for skin cancer recognition. en. *Nature Medicine* **26,** 1229–1234. ISSN: 1078-8956, 1546-170x. `http://www.nature.com/articles/s41591-020-0942-0` (2020) (Aug. 2020).

609. Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A. & Madry, A. *From imagenet to image classification: Contextualizing progress on benchmarks* in *International Conference on Machine Learning* (2020), 9625–9635.

610. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A. & Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152* (2018).

611. Tucker, J. D. *et al.* Ethical concerns of and risk mitigation strategies for crowdsourcing contests and innovation challenges: scoping review. *Journal of medical Internet research* **20,** e75 (2018).

612. Tyson, A., Pasquini, G., Spencer, A. & Funk, C. 60% of Americans Would Be Uncomfortable With Provider Relying on AI in Their Own Health Care (2023).

613. Usatine, R. P., Smith, M. A., Mayeaux, E. & Chumley, H. S. *The color atlas of family medicine* 2009.

614. Vaccari, C. & Chadwick, A. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. en. *Social Media + Society* **6,** 205630512090340. ISSN: 2056-3051, 2056-3051. `http://journals.sagepub.com/doi/10.1177/2056305120903408` (2021) (Jan. 2020).

615. Vaccaro, M. & Waldo, J. The effects of mixing machine learning and human judgment. *Communications of the ACM* **62,** 104–110 (2019).

616. Valstar, M. F., Jiang, B., Mehu, M., Pantic, M. & Scherer, K. *The first facial expression recognition and analysis challenge* in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)* (2011), 921–926.

617. Van Such, M., Lohr, R., Beckman, T. & Naessens, J. M. Extent of diagnostic agreement among medical referrals. *Journal of evaluation in clinical practice* **23,** 870–874 (2017).

618. Van Zyl, L., Du Plessis, J. & Viljoen, J. Cutaneous tuberculosis overview and current treatment regimens. *Tuberculosis* **95,** 629–638 (2015).

619. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine* **5,** 48 (2022).

620. Verdoliva, L. Media Forensics and DeepFakes: an overview. en. *arXiv:2001.06564 [cs].* arXiv: 2001.06564. `http://arxiv.org/abs/2001.06564` (2020) (Jan. 2020).

621. Vereschak, O., Bailly, G. & Caramiaux, B. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. en. *Proceedings of the ACM on Human-Computer Interaction* **5,** 1–39. ISSN: 2573-0142. `https://dl.acm.org/doi/10.1145/3476068` (2021) (Oct. 2021).

622. Vinyals, O. *et al.* Grandmaster level in StarCraft II using multi-agent reinforcement learning. en. *Nature* **575,** 350–354. ISSN: 0028-0836, 1476-4687. `http://www.nature.com/articles/s41586-019-1724-z` (2021) (Nov. 2019).

623. Vlachostergiou, A., Caridakis, G. & Kollias, S. *Investigating context awareness of Affective Computing systems: A critical approach* in *Procedia Computer Science* **39** (Elsevier B.V., 2014), 91–98.

624. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. en. *Science* **359,** 1146–1151. ISSN: 0036-8075, 1095-9203. `https://www.sciencemag.org/lookup/doi/10.1126/science.aap9559` (2020) (Mar. 2018).

625. Wallace, E. *et al. Automated Crossword Solving* in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2022), 3073–3085.

626. Wang, A. *et al.* GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).

627. Wang, A. *et al.* Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems* **32** (2019).

628. Wang, A., Ramaswamy, V. V. & Russakovsky, O. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. *arXiv preprint arXiv:2205.04610* (2022).

629. Wang, P. *et al. TAL EmotioNet Challenge 2020 Rethinking the Model Chosen Problem in Multi-Task Learning* (2020).

630. Wang, T. T. *et al.* Adversarial Policies Beat Professional-Level Go AIs. *arXiv preprint arXiv:2211.00241* (2022).

631. Wang, Z. *et al. Towards fairness in visual recognition: Effective strategies for bias mitigation* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), 8919–8928.

632. Ware, O. R., Dawson, J. E., Shinohara, M. M. & Taylor, S. C. Racial limitations of Fitzpatrick skin type. en. *Cutis.* **105,** 77–80 (2020).

633. Watts, D. J. Should social science be more solution-oriented? *Nature Human Behaviour* **1,** 1–5 (2017).

634. Watts, D. J., Rothschild, D. M. & Mobius, M. Measuring the news and its impact on democracy. *Proceedings of the National Academy of Sciences* **118** (2021).

635. Wen, D. *et al.* Characteristics of publicly available skin cancer image datasets: a systematic review. en. *Lancet Digital Health,* S2589750021002521. ISSN: 25897500. `https://linkinghub.elsevier.com/retrieve/pii/S2589750021002521` (2021) (Nov. 2021).

636. Werner, P., Saxen, F. & Al-Hamadi, A. *Facial Action Unit Recognition in the Wild with Multi-Task CNN Self-Training for the EmotioNet Challenge* en. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Ieee, Seattle, WA, USA, June 2020), 1649–1652. ISBN: 978-1-72819-360-1. `https://ieeexplore.ieee.org/document/9151065/` (2021).

637. Widmer, G. & Kubat, M. Learning in the presence of concept drift and hidden contexts. *Machine learning* **23,** 69–101 (1996).

638. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine* **25,** 1337–1340 (2019).

639. Wieser, M. J. & Brosch, T. Faces in context: a review and systematization of contextual influences on affective face processing. *Frontiers in psychology* **3,** 471 (2012).

640. Wilkes, M., Wright, C. Y., du Plessis, J. L. & Reeder, A. Fitzpatrick Skin Type, Individual Typology Angle, and Melanin Index in an African Population: Steps Toward Universally Applicable Skin Photosensitivity Assessments. en. *JAMA Dermatology* **151,** 902–903. ISSN: 2168-6068. eprint: `https://jamanetwork.com/journals/jamadermatology/articlepdf/2280387/dld150006.pdf`. `https://doi.org/10.1001/jamadermatol.2015.0351` (Aug. 2015).

641. Williams, D. R. & Wyatt, R. Racial bias in health care and health: challenges and opportunities. *Jama* **314,** 555–556 (2015).

642. Wilson-Mendenhall, C. D., Barrett, L. F., Simmons, W. K. & Barsalou, L. W. Grounding emotion in situated conceptualization. en. *Neuropsychologia* **49,** 1105–1127. ISSN: 00283932. `https://linkinghub.elsevier.com/retrieve/pii/S0028393210005658` (2020) (Apr. 2011).

643. Winkler, J. K. *et al.* Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. en. *JAMA Dermatology* **155,** 1135. ISSN: 2168-6068. `https://jamanetwork.com/journals/jamadermatology/fullarticle/2740808` (2020) (Oct. 2019).

644. Witkower, Z. & Tracy, J. L. A facial-action imposter: How head tilt influences perceptions of dominance from a neutral face. *Psychological science* **30,** 893–906 (2019).

645. Wittenberg, C., Tappin, B. M., Berinsky, A. J. & Rand, D. G. The (minimal) persuasive advantage of political video over text. en. *Proceedings of the National Academy of Sciences* **118,** e2114388118. ISSN: 0027-8424, 1091-6490. `http://www.pnas.org/lookup/doi/10.1073/pnas.2114388118` (2021) (Nov. 2021).

646. Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E. & Wells, W. Estimating the reliability of eyewitness identifications from police lineups. en. *Proceedings of the National Academy of Sciences* **113,** 304–309. ISSN: 0027-8424, 1091-6490. `http://www.pnas.org/lookup/doi/10.1073/pnas.1516814112` (2021) (Jan. 2016).

647. Wohler, L., Castillo, S., Zembaty, M. & Magnor, M. *Towards Understanding Perceptual Diferences between Genuine and Face-Swapped Videos* 2021.

648. Wolff, K. *et al. Fitzpatricks Textbook of Dermatology in General Medicine* 2008.

649. Wondra, J. D. & Ellsworth, P. C. An appraisal theory of empathy and other vicarious emotional experiences. *Psychological Review* **122,** 411–428. ISSN: 0033295X (3 July 2015).

650. Woodard, K., Plate, R. C., Morningstar, M., Wood, A. & Pollak, S. D. Categorization of Vocal Emotion Cues Depends on Distributions of Input. en. *Affective Science.* ISSN: 2662-2041, 2662-205x. `https://link.springer.com/10.1007/s42761-021-00038-w` (2021) (Apr. 2021).

651. Wu, E. *et al.* How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. en. *Nature Medicine* **27,** 582–584. ISSN: 1078-8956, 1546-170x. `http://www.nature.com/articles/s41591-021-01312-x` (2022) (Apr. 2021).

652. Xiao, K., Engstrom, L., Ilyas, A. & Madry, A. Noise or Signal: The Role of Image Backgrounds in Object Recognition. `http://arxiv.org/abs/2006.09994` (June 2020).

653. Yadav, A. *et al.* If a picture is worth a thousand words is video worth a million? Differences in affective and cognitive processing of video and text cases. *Journal of Computing in Higher Education* **23,** 15–37 (2011).

654. Yalcin, O. N. *Evaluating Empathy in Artificial Agents* en. in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (Ieee, Cambridge, United Kingdom, Sept. 2019), 1–7. ISBN: 978-1-72813-888-6. `https://ieeexplore.ieee.org/document/8925498/` (2022).

655. Yalçın, O. N. & DiPaola, S. *Evaluating levels of emotional contagion with an embodied conversational agent* in (2019).

656. Yang, X., Li, Y. & Lyu, S. *Exposing Deep Fakes Using Inconsistent Head Poses* en. arXiv: 1811.00661. Nov. 2018. `http://arxiv.org/abs/1811.00661` (2021).

657. Yannakakis, G. N., Cowie, R. & Busso, C. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing* (2018).

658. Yannakakis, G. N. & Paiva, A. Emotion in games. *Handbook on affective computing* **2014,** 459–471.

659. Yin, R. K. Looking at upside-down faces. *Journal of Experimental Psychology* **81,** 141–145 (1969).

660. Young, A. W. & Burton, A. M. Are we face experts? *Trends in cognitive sciences* **22,** 100–110 (2018).

661. Zakharov, E., Shysheya, A., Burkov, E. & Lempitsky, V. *Few-shot adversarial learning of realistic neural talking head models* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), 9459–9468.

662. Zhang, B., Essl, G. & Provost, E. M. *Predicting the distribution of emotion perception: capturing inter-rater variability* in (2017), 51–59.

663. Zhang, K., Zhang, Z., Li, Z. & Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* **23,** 1499–1503 (2016).

664. Zhang, Y., Weninger, F., Bjorn, S. & Picard, R. Holistic Affect Recognition Using PaNDA: Paralinguistic Non-metric Dimensional Analysis. en. *IEEE Transactions on Affective Computing,* 1–1. ISSN: 1949-3045, 2371-9850. `https://ieeexplore.ieee.org/document/8941312/` (2020) (2020).

665. Zhang, Y. *et al. Identify experts through Revealed Confidence: application to Wisdom of Crowds* PhD thesis (Massachusetts Institute of Technology, 2020).

666. Zhao, M., Adib, F. & Katabi, D. *Emotion recognition using wireless signals* in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking* (2016), 95–108.

667. Zhou, Z., Nartker, M. & Firestone, C. When will AI misclassify? Human intuition for machine (mis) perception. *Journal of Vision* **20,** 1325–1325 (2020).