Human and Artificial Intelligence in Decision Systems
for Social Development

by

Alejandro Noriega Campero

B.Sc., Mexican Autonomous Institute of Technology (2012)
S.M., Technology and Policy, Institute for Data, Systems, and Society,
Massachusetts Institute of Technology (2015)

Submitted to the Program in Media Arts and Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

© Massachusetts Institute of Technology 2019. All rights reserved.
Human and Artificial Intelligence in Decision Systems for Social Development

by

Alejandro Noriega Campero

Submitted to the Program in Media Arts and Sciences on August 09, 2019, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Media Arts and Sciences

Abstract

Today there is widespread expectation about how ubiquitous data and intelligent systems may revolutionize society towards shared prosperity; or conversely, deepen social inequalities, bring the end of human agency, and forgo the right to privacy. In this two-part thesis, we investigate the societal value and perils of hybrid decision systems—which integrate elements of human and artificial intelligence.

Part I of this thesis focuses on their potential for promoting social development goals, with applications in poverty alleviation and public health. Towards public health, in the context of early detection of diabetic blindness, we show that human-AI hybrid systems can be more accurate than either human or algorithm in isolation, and that both opinions benefit from mutual exposure. Towards improved poverty action, we argue that poverty-targeting rules are among the most relevant algorithms operating in the world today. We demonstrate that a shift towards the use of AI methods in poverty-based targeting can substantially increase accuracy, extending the coverage of social policies by nearly a million people in the case of two Latin American countries, without increasing budgets. However, it is also shown that both the status quo and AI-systems induce disparities across population subgroups. Hence, we close by proposing a decision support tool that empowers diverse social institutions to design fair targeting rules under a distributed governance framework.

Part II addresses cross-cutting challenges that arise as one applies these technologies in real-world domains towards social development. In particular, the work presented provides academic and practical contributions on: 1) achieving fairness in algorithmic decision systems by means of adaptive information collection, a novel paradigm we call active fairness; and 2) preserving privacy and mapping its tradeoff against utility in development contexts.

Thesis Supervisor: Alex "Sandy" Pentland
Title: Toshiba Professor of Media Arts and Sciences
Human and Artificial Intelligence in Decision Systems
for Social Development
by
Alejandro Noriega Campero

This doctoral thesis has been examined by a Committee as follows:

Thesis Advisor........................................................
Alex "Sandy" Pentland
Toshiba Professor of Media Arts and Sciences
Massachusetts Institute of Technology
Human and Artificial Intelligence in Decision Systems
for Social Development
by
Alejandro Noriega Campero

This doctoral thesis has been examined by a Committee as follows:

Thesis reader ..........................................................
Iyad Rahwan
Director
Center for Human and Machines
Max Planck Institute for Human Development
Human and Artificial Intelligence in Decision Systems
for Social Development
by
Alejandro Noriega Campero

This doctoral thesis has been examined by a Committee as follows:

Thesis reader..............................................................................

Esteban Moro
Associate Professor
Universidad Carlos III de Madrid
Human and Artificial Intelligence in Decision Systems for Social Development
by
Alejandro Noriega Campero

This doctoral thesis has been examined by a Committee as follows:

Thesis reader ..........................................................
Rayid Ghani
Director
Center for Data Science & Public Policy
University of Chicago
Acknowledgments

The saying goes: “It takes a village to raise a child.”
I say it takes a village to raise a PhD.

My joy and any success in this doctoral work is owed to my direct collaborators—Abdullah, Michiel, Yves, Peter, Bernardo, Mehdi, Eaman, Yan, Abdulrahman, Dhaval, Oren, Luis, Fernando, Dalia and others—, wise mentors—Sandy, Esteban, Iyad, Rayid, Dean, Sinan, and others—, and dear friends and peers—Martin Saveski, Nico Ricci, Julius, Scott, Jacob, Justin, Henry, Sami, Alex, Mike, Tiziana, Cecilia, the dear Mexican community, and many many others with whom I’ve shared this journey.

In particular, I’d like to thank Sandy, whose unconditional support often relates more to that of a parent rather than strictly an advisor. Life was extremely generous in placing me for a PhD at the MIT Media Lab, and in the Human Dynamics Lab under your leadership. Thanks for all the support, mentorship, and freedom throughout the years.

Apart from that, life and fortune were only more generous when having me born in my parents’ house—Pamen and Carlos. Together with Elena and Rafa, they formed me into loving life, embracing people, and striving for intellectual adventure and social change. To them I owe all I am.

The deepest personal experiences along my PhD journey have come from the families I’ve had the pleasure to choose and live with. Starting with Pierre, Juan, and Rami, and following with Andres, Bjarke, Flo, and Alvaro. Thank you. You are brothers to me, and have woven into my soul. Lastly, Julie, with whom we’ve ridden a comet since back in Ilhia Grande. Thanks for your love and unconditional support throughout these years, and for illuminating our years ahead.

Above all, thanks to God, for without his ..

\footnote{1 just kidding..}
Contents

1 Introduction 29

2 Architecting Multi-Agent Decision Systems for Collective Intelligence 33
   2.1 Background 35
   2.2 Results 38
   2.3 Discussion 44
   2.4 Appendix: Experiment Details 45

3 Human + Artificial Intelligence for Early Diagnosis of Diabetic Blindness 47
   3.1 Relevance: Massifying the Early Diagnosis of Diabetic Blindness 48
   3.2 Human + AI Experiment 51
      3.2.1 Research Questions 51
      3.2.2 Experimental Design 54
      3.2.3 AI Screening Algorithm 57
   3.3 Results 61
   3.4 Conclusions 63

4 Human + Artificial Intelligence for Fair and Efficient Targeting of Social Policies 65
   4.1 Worldwide Relevance of Targeted Social Policies —and their Underlying Algorithms 67
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2</td>
<td>The Accuracy Edge: Large-Scale Impact of AI-Based Targeting</td>
<td>69</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Algorithmic Targeting of Social Policies</td>
<td>69</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Empirical Evaluation</td>
<td>70</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Results</td>
<td>73</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Inter-Temporal Robustness and Sensitivity Analysis</td>
<td>77</td>
</tr>
<tr>
<td>4.3</td>
<td>Disparities in Predictive Performance across Population Subgroups</td>
<td>80</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Methodology</td>
<td>80</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Results</td>
<td>81</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Section Conclusions</td>
<td>83</td>
</tr>
<tr>
<td>4.4</td>
<td>Fair Targeting and Distributed Governance via Interactive Decision Support</td>
<td>87</td>
</tr>
<tr>
<td>4.4.1</td>
<td>The Architecture of Real-World Social Targeting</td>
<td>87</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Heterogeneous Preferences and Disagreement Over Prioritization and Fairness Criteria</td>
<td>88</td>
</tr>
<tr>
<td>4.4.3</td>
<td>An Interactive Decision Support Platform for Fair Targeting and Distributed Governance</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>Active Fairness in Algorithmic Decision-Making</td>
<td>99</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>100</td>
</tr>
<tr>
<td>5.2</td>
<td>Related Work</td>
<td>101</td>
</tr>
<tr>
<td>5.3</td>
<td>Active Fairness</td>
<td>106</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Group-Level Information Budgets</td>
<td>106</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Individual-Level Information Budgets</td>
<td>107</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Random Forest Implementation</td>
<td>108</td>
</tr>
<tr>
<td>5.4</td>
<td>Datasets</td>
<td>109</td>
</tr>
<tr>
<td>5.5</td>
<td>Achieving Equal Opportunity &amp; Calibration</td>
<td>110</td>
</tr>
<tr>
<td>5.6</td>
<td>Achieving Equal Odds</td>
<td>111</td>
</tr>
<tr>
<td>5.7</td>
<td>Conclusions</td>
<td>113</td>
</tr>
<tr>
<td>6</td>
<td>Privacy-Utility Tradeoffs in the Use of High-Dimensional Data for Development</td>
<td>115</td>
</tr>
</tbody>
</table>
6.1 Introduction ................................................. 116
  6.1.1 Mobile Phone Data for Development ...................... 116
  6.1.2 Privacy Risk in Mobile Phone Metadata .................... 117
6.2 Methods .................................................... 118
  6.2.1 Assessing Privacy ...................................... 118
  6.2.2 Assessing Utility ....................................... 120
6.3 Results ..................................................... 122
  6.3.1 Reidentification Results ................................. 122
  6.3.2 Utility Results .......................................... 123
  6.3.3 Privacy-Utility Tradeoff .................................. 123
6.4 Conclusions ............................................... 125

7 Conclusions and Future Work ................................. 127
List of Figures

2-1 Experimental design. Panel (A) shows an illustration of the experiment design. Panel (B) illustrates example of the scatter plots used in the experiment. For any given round, all participants saw plots that shared an identical true correlation, but difficulty levels could differ among them. Participants were not informed about the difficulty level they or other participants were facing. ............................................. 37

2-2 Individual and collective outcomes. Groups connected by dynamic influence networks incur substantially lower individual errors ($P < 10^{-5}$) as shown in Panel (A) and collective errors ($P < 0.001$) in Panel (B). The reduction is notably larger and more significant in periods where networks had adapted to the information environment (i.e., rounds [6,10] and [16,20]). Errors are normalized with respect to average errors in the solo condition. Statistics on individual-level outcomes are based on White cluster-robust standard errors. Error bars indicate 95% confidence intervals. ............................................. 39

2-3 Mechanisms promoting collective intelligence in dynamic networks. Panel (A) shows that the network becomes more centralized with time (Freeman global centralization—i.e., how far the network is from a star network). Panel (B) depicts the relation between performance (i.e., average error) and popularity (i.e., number of followers). Panel (c) shows the relationship between accuracy of initial estimate and confidence (i.e., resistance to social influence). Error bars indicate 95% confidence intervals. ............................................. 41
2-4 **An illustrative example of the network dynamics.** The circle color represents performance. The size of each circle represents the number of followers (i.e., popularity). The dashed orange line is the distribution of estimates prior to social influence, the blue solid line is the distribution of post-social influence estimates, while the dashed vertical line is the true correlation.

2-5 **Mean-variance trade-off.** Mean and standard deviation of absolute errors incurred by top-k estimates during the adapted periods (rounds $\in [6,10] \cup [16,20]$). Top-12 estimates correspond to the full-group mean, and top-1 to the group’s best individual. Within each condition, top-k trade-off curves first gain in both objectives, then trade off lower average error for higher variability, and finally regress in both objectives as $k \to 1$. Across conditions, for any $k \in [1,12]$, groups in the dynamic condition outperformed groups in the static and solo conditions. Moreover, the full-group mean of dynamic networks averaged 28% lower error and 48% less variability than the best individual playing solo (dynamic top-12 vs. solo top-1; $P < 10^{-2}$); and the best individual in the dynamic condition averaged 32% lower error and 38% less variability than her analogue in solo (dynamic top-1 vs. solo top-1; $P < 10^{-2}$). Bars indicate 95% confidence intervals.

3-1 **Hybrid decision system for massifying early diagnosis of diabetic blindness.**

3-2 **Solo condition.** The participant responds to the task without exposure to the DR-screening algorithm.

3-3 **Confidence condition.** The participant can observe the answer and confidence of the algorithm (slider on the right).

3-4 **Attention condition.** The participant can observe the answer and confidence of the algorithm (slider on the right), as well as the attention map.
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-5</td>
<td>Comparison between VGG16 architecture and the architecture used in the current work.</td>
</tr>
<tr>
<td>3-6</td>
<td>Receiver operating characteristic (ROC) curves in EyePACS test and APEC datasets.</td>
</tr>
<tr>
<td>3-7</td>
<td>Fondus image with and without attention map overlaid.</td>
</tr>
<tr>
<td>3-8</td>
<td>Performance results.</td>
</tr>
<tr>
<td>3-9</td>
<td>Social influence magnitudes across conditions.</td>
</tr>
<tr>
<td>4-1</td>
<td>Presence of nation-wide targeted cash transfer programs worldwide.</td>
</tr>
<tr>
<td>4-2</td>
<td>Exclusion vs. Inclusion errors curve. Grey bars around the red points indicate 95% confidence intervals, computed non-parametrically via bootstrap resampling.</td>
</tr>
<tr>
<td>4-3</td>
<td>Exclusion vs. Inclusion errors curve. Grey bars around the red points indicate 95% confidence intervals, computed non-parametrically via bootstrap resampling.</td>
</tr>
<tr>
<td>4-4</td>
<td>Inter-temporal exclusion vs. inclusion errors curves.</td>
</tr>
<tr>
<td>4-5</td>
<td>Sensitivity map. Screenshot of the interactive visualization. The area of each rectangle represents the average magnitude of the effect that perturbations of a particular variable have on the poverty score. The visualization interactively displays information of variables and categories as the user hovers over them.</td>
</tr>
<tr>
<td>4-6</td>
<td>Subgroup-specific performance improvements. Improvements achieved by the AI-based over the status quo, for three selected examples of country-attribute pairs. All subgroups substantially benefited from switching to the AI-based method. Error bars denote 95% confidence intervals computed non-parametrically via bootstrap sampling.</td>
</tr>
</tbody>
</table>
Group-specific comparative results for Colombia. Third rows report % improvements of AI-based vs. status quo methods. First columns show unbalance measures $\sigma_A = \sqrt{\sum_{g \in A} (e_g - \bar{e})^2}$ for each attribute-method pair. All results are statistically significant at 95% except noted by $^{ns}$. 

Interactive decision-support platform. Adjusting thresholds across subgroups towards achieving different fairness criteria. Threshold parity is instantiated by default after segmenting a node (shown in Figure 4-11(b)). Subfigures here exemplify the cases of coverage parity (a) and positive discrimination (b).
Achieving calibration and single error rate parity: classifiers with group-level information budgets vs. naive randomization. Rows correspond to analysis on two different datasets: the Mexican poverty and adult income datasets. Green and yellow colors correspond to error rates for two population subgroups (e.g., white and non-white individuals).

Solid black lines represent the space of calibrated classifiers. Panels in the first column (A and D) show the generalized false positive and false negative rates (GFPR and GFNR) of classifiers that randomize an increasing proportion of individuals (0% to 100%), as in [106]. In line with [106], naive randomization is able to achieve calibration and any single error parity constraint. Panels in the second column (B and E) show the same analysis for classifiers with group-level budgets. These classifiers are effective in achieving parity on either false positive or false negative rates, or equal cost, while maintaining calibration; yet without resorting to naive randomization. Finally, panels in the third column (C and F) compare the efficiency of both methods, by showing the best classifiers that achieve equal opportunity and calibration, under an information budget restriction $b = \frac{1}{n} \sum_{i} b_i < b_{\text{max}}$. Classifiers with group-level budgets Pareto-dominate randomized classifiers by a wide margin, i.e. for the same information budget, both population subgroups are better off, incurring substantially lower false positive and false negative errors.
5-2 Achieving parity in false positives and false negatives (*equal odds*) via group-level information budgets. Results correspond to the *Mexican poverty dataset*. Achievable regions of classifiers for each population subgroup are plotted in blue and yellow. The outer and inner FPR-FNR curves of each achievable region correspond to classifiers using maximum and minimum information budgets. Points along the curve correspond to different values of the decision threshold. It is observed that active classifiers with group-level budgets achieve parity in both FNR and FPR (*equal odds*). Moreover, they provide *equal odds* solutions anywhere on the overlap of the achievable regions for both subgroups and thus along the entire FNR-FPR trade-off.

5-3 Achievable space of *equal odds* classifiers. Comparison of active classifiers with group-level (green) and individual-level (yellow) information budgets, and naive randomization (blue, following [65]). Classifiers are constrained by an information budget $b \leq b_{\text{max}}$. Figure A corresponds to *Mexican poverty*, B to *German credit* (center) and C to *Heart health* under different budget constraints. It is observed that active classifiers yield *equal odds* solutions along the FPR-FNR trade-off, without resorting to randomization. Moreover, in a budget-constraint setting, the active classifiers are substantially more information-efficient. The individual-level budget Pareto-dominates the group-level classifiers which in turn dominate randomized classifiers in budget-constrained environments.

6-1 **Reidentification results.** Reidentification results for datasets at varying levels of spatial and temporal granularity: (a) *information cost* and (b) *information ratio*. All 95% confidence intervals are non-overlapping (except for pairs $(D1, Z6)$ and $(D24, M6)$, as shown in Additional File 1).
6-2 Utility results. Usefulness of mobile phone metadata per generalization profile. Grey bars denote bootstrapped 95% confidence intervals. Spatial granularity levels are Z = ZIP code, D = district, and M = municipality.

6-3 Privacy-Utility trade-off in mobile phone data. Utility vs. re-identification risk in mobile phone data for development, across spatial and temporal granularities \{ZIP, District, Municipality\} × \{1h, 6h, 12h, 24h\}.

The more useful the dataset, the less auxiliary information is needed to reidentify its individuals. Conversely, while data generalization increasingly hinders reidentification, it strongly diminishes datasets’ value.

7-1 Integrated system for early diagnosis of chronic disease. Targeting population at risk, coupled with H+AI screening, towards massifying early diagnostics in developing countries.
## List of Tables

3.1 EyePACS and APEC distributions ........................................ 57

4.1 Household survey data statistics ........................................ 71

4.2 Comparative Accuracy Results and their Impact on Coverage of the Poor .................................................. 76

6.1 Example of CDR records .................................................. 117

6.2 Experts data ............................................................. 121
Chapter 1

Introduction
Consider a scenario in which you are appointed health minister of (fictional) country Uqbar, where, as in many other developing countries, diabetes is a leading cause of disability and death. You seek to leverage artificial intelligence (AI) and related technologies for improving prevention, diagnosis, and treatment of diabetes. Yet, despite their potential societal value, you soon realize numerous challenges. In particular, how should you best couple human and machine intelligence to yield more accurate and cost-efficient decision systems and interventions? Even in light of large potential benefits, how should you build intelligent systems to be fair by design? What are the underlying fairness-utility, and privacy-utility trade-offs, and how can you manage and decide over them? Given only coarse national statistics, and a limited budget for individual-level information collection, how can you target interventions over a large population?

Today’s age of data and artificial intelligence holds high potential to enhance the way we pursue and monitor development goals, such as poverty alleviation, health, and education. Yet, numerous challenges still hinder the path from the lab to positive social impact. This thesis aims to yield academic and practical contributions addressing three key elements in this space:

1. Architecting human and artificial intelligence for improving development-related decision systems.

2. Demonstrating the substantial positive impact that these can have on advancing Sustainable Development Goals, such as no poverty (SDG 1) and good health and well-being (SDG 3).

3. Understanding and overcoming transversal issues related to fairness and privacy.

In line with MIT’s motto Mens et Manus, as well as the Media Lab’s emphasis on demonstration and real-world deployment, the ultimate goal of this doctoral work is to follow a progression, from academic work in controlled environments, to development and deployment of human+AI systems that materialize positive social impact in the field.
Outline

The thesis starts by investigating questions regarding the architecture of multi-agent and hybrid (human + AI) decision systems, in Chapters 2 and 3. Chapter 2 studies the effects that different communication structures among agents have on the accuracy of individual and group decision-making. In particular, it elicits notable advantages from endowing multi-agent decision systems with adaptive rather than static communication structures.

Chapter 3 leverages the insights and framework from Chapter 2 to study multi-agent decision-making in a domain with two key characteristics. 1) Real-world relevance towards sustainable development goals, focusing on systems that could help massify early diagnosis of diabetic blindness in developing countries. And 2) Hybrid human and AI decision systems, studying architectural designs that could optimize the accuracy and cost of early diagnosis.

Chapter 4 presents an in-depth study focused on a relevant application domain: the targeting of social development policies. It offers a comprehensive view, from the generation of accurate AI models, to the development of a decision-support platform that enhances distributed governance, fairness, and transparency.

On the one hand, the chapter demonstrates that a shift towards the use of AI methods in poverty-based targeting can substantially increase accuracy, increasing the coverage of social policies by nearly a million people in two Latin American countries, without increasing budgets. However, it is also shown that both the status quo and AI-systems suffer from performance disparities across population subgroups. The chapter closes by proposing and describing an interactive decision support platform that empowers social institutions towards designing fair and accurate targeting rules contextualized to their needs.

The latter part of the thesis deals with cut-crossing issues that naturally arise when deploying intelligent decision systems, particularly in domains akin to sustainable development goals: fairness and privacy. In particular, Chapter 5 presents a novel approach to leverage adaptive information collection, towards jointly optimiz-
ing decision systems’ accuracy and fairness across population subgroups. Chapter 6 contributes relevant insights on the fundamental tradeoff between, on the one hand, the use of novel sources of data for—high-dimensional data such as mobile phone records, credit card records, and satellite imaging—development purposes; and the risk to individuals’ privacy on the other.

Finally, Chapter 7 wraps-up by listing several exciting avenues of future work, towards further synergizing the methods and frameworks developed, and materializing the societal value of the present doctoral work.
Chapter 2

Architecting Multi-Agent Decision Systems for Collective Intelligence
Human social networks evolve over time. The structure of face-to-face interaction networks, advice networks, communication networks, and online social networks, tends to change continuously as people create new ties and break existing ones. It is widely noted that our social embedding exerts a strong influence on what information we receive, and on how we form beliefs and ultimately make decisions. However, most studies on collective intelligence overlook the dynamic nature of social influence networks, and its potential role in fostering adaptive collective intelligence (e.g., [11]). It remains unknown (1) how network structures adapt to the skills and performances of individuals, and (2) whether this adaptation promotes or impairs the accuracy of individual and collective decisions. In other words, are adaptive social networks wise?

Here, we answer these questions through a series of behavioral experiments involving 719 participants and complementary numerical simulations. Our results reveal that groups of people embedded in dynamic social networks can adapt to their peers’ performances, even in biased and non-stationary information environments. As a result, individual and collective accuracy is substantially improved over static networks and unconnected groups. Moreover, we show that groups in dynamic networks far outperform their best-performing member and that even the best member’s judgment substantially benefits from group engagement. Thereby, our findings demonstrate the role of dynamic social networks as adaptive mechanisms for refining individual judgments and harnessing groups’ collective intelligence.

Significance Statement

Social networks have a strong influence on how people form judgments and make decisions. We address the question whether the structure of such networks can adapt to leverage the skills of individuals and promote collective intelligence. In our experiment, groups of participants were embedded in social networks and asked to solve a series of estimation tasks. We show, for the first time to our knowledge, that groups in dynamic networks—where network structure can change by forming and breaking ties in response to peers’ performance—improve individual and collective performance substantially (compared to static networks and unconnected groups), and far outper-
form even their best-performing member in isolation. Hence, we reveal adaptive social networks as key mechanisms for harnessing groups’ collective intelligence.

2.1 Background

Intelligent systems, both natural and artificial, rely on feedback, empirical learning, and adaptation [133]. Such systems are widespread, and can often be viewed as networks of interacting entities that dynamically evolve over time. Cell reproduction, for example, relies on protein networks to combine sensory inputs into gene expression choices adapted to environmental conditions [47]. Neurons in the brain dynamically rewire for human learning [62]. Animal swarms modify their connectivity to enhance their collective intelligence [73]. Dynamic interaction networks have been shown to promote human cooperation [109], and culture transmission networks over generations enabled human groups to develop technologies above any individual’s capabilities [66].

In the artificial realm, prominent machine learning algorithms rely on similar logics, where dynamically updated networks integrate input signals into useful output [18, 96]. Across the board, networks’ dynamic properties embody key mechanisms that enable systems to adapt to environmental changes [68, 81].

In our view, the information processing capabilities of interacting human groups are no exception. People’s behavior, opinion formation, and decision-making are deeply rooted in cumulative bodies of social information [7], accessed through social networks formed by choices of whom we friend [4, 130], follow [121], call [45, 104], imitate [143, 115], trust [29, 126], and cooperate with [131, 109, 51]. Moreover, peer choices are frequently revised, most often based on notions of peer success and reliability, or approximations such as reputation, popularity, and socio-demographics [136, 67, 84, 53, 93].

It is widely noted, however, that social influence strongly correlates individuals’ judgment in estimation tasks [95, 85, 11, 58], compromising the first of two assumptions underlying common statistical accounts of ‘wisdom-of-crowds’ phenomena [120]: namely, that (i) individual estimates are uncorrelated, or negatively correlated, and
(ii) individuals are correct in mean expectation \([54, 58]\). In recent years, numerous studies have offered conflicting findings, showing that social interaction can either significantly benefit group and individual estimates \([6, 11, 99]\), or, conversely, lead them astray by inducing social bias, herding, and group-think \([95, 58, 85]\). Notably, both theoretical and experimental work has been limited mainly to frameworks where agents are randomly placed in static social structures—dyads \([6, 78]\), groups \([85, 138, 99]\), or networks \([58, 11]\)—and has found that these divergent effects are a function of whether well-informed individuals are placed in prominent positions in the network structure \([58, 11]\), and how self-confident they are \([6, 78, 90]\).

However, unlike what is assumed in most existing work, the social networks we live in are not random or imposed by external forces, but emerge shaped by endogenous social processes and gradual evolution. The present study builds on the observation that agent characteristics, such as skill and information access, are not randomly located in network structures. Instead, their distribution is often the outcome of social heuristics that form and break ties influenced by social cues \([136, 20, 67]\). Intuitively, groups can benefit from awarding centrality to and amplifying the influence of well-informed individuals. Hence we hypothesize that dynamic social influence networks may be central to human collective intelligence, acting as core mechanisms by which crowds, which may not initially be wise, evolve into wisdom, adapting to biased and potentially non-stationary information environments.

To test these hypotheses, we developed a web-based experiment that allows us to identify the role of dynamic networks in fostering an adaptive ‘wisdom of crowds.’ Participants \((n = 719)\) from Amazon Mechanical Turk (population details in SI) engaged in a sequence of 20 estimation tasks. Each task consisted of estimating the correlation of a scatter plot, and monetary prizes were awarded in proportion to performance. Participants were randomly allocated to groups of 12 \((n = 60 \text{ groups})\), and each group was randomized to one of three treatment conditions: a solo condition, where each individual solved the sequence of tasks in isolation; a static condition, in which participants were randomly placed in static communication networks; and a dynamic condition, in which participants at each round were allowed to select up
Figure 2-1: **Experimental design.** Panel (A) shows an illustration of the experiment design. Panel (B) illustrates example of the scatter plots used in the experiment. For any given round, all participants saw plots that shared an identical true correlation, but difficulty levels could differ among them. Participants were not informed about the difficulty level they or other participants were facing.

to three neighbors to communicate with. Fig. 2-1A illustrates the experiment flow according to each condition. At each round, participants were initially asked to submit an independent guess. Then those in static and dynamic conditions entered a social exposure stage, where they could observe the answers of their network peers, update their own, and see peers’ updates in real time. After submitting a final guess, participants in all conditions were shown feedback on their score and the correct answer. Lastly, those in the dynamic condition were shown scores of all participants and allowed to revise which peers to follow in subsequent rounds.
The estimation task was designed so that it allowed manipulation of the quality of information provided to participants. Scatter plots with three levels of difficulty were used (varying the number of points, and adding outliers or non-linearities; see Figure 2-1B). At every round, all plots seen by participants shared an identical true correlation, but difficulty levels could differ among them. The design allowed the simulation of a shock to the distribution of information among participants. Specifically, each participant experienced a constant difficulty level across the first ten rounds; then, at round eleven, we introduced shocks by reshuffling difficulties to new levels that remained constant thereafter. Participants were not informed about the difficulty levels they or their peers faced.

2.2 Results

Individual and collective outcomes

We first compared individual- and group-level errors across conditions. Evolutionary reasoning suggests that people’s propensity to imitate follows from its direct benefits to the individual, but it may, nonetheless, induce benefits to the population as a whole. Fig. 2-2 shows the individual and group error rates of the static and dynamic conditions—using the arithmetic mean as group estimate—normalized with respect to baseline errors in the solo condition. Overall, we find that participants in dynamic networks achieved the lowest error rates, averaging 33% lower individual error ($P < 10^{-5}$), and 34% lower group error, compared to participants in static networks ($P < 10^{-4}$). In particular, the performance edge of groups in dynamic networks was larger in periods where networks had adapted to their information environment (rounds $[6, 10] \cup [16, 20]$, the ‘adapted periods’), reducing individual error by 45% ($P < 10^{-10}$) and group error by 47% ($P < 10^{-10}$) compared to groups connected by static networks. Hence, these results support our primary hypothesis that dynamic networks’ adaptiveness can benefit both individual and collective judgment.
Figure 2-2: **Individual and collective outcomes.** Groups connected by dynamic influence networks incur substantially lower individual errors ($P < 10^{-5}$) as shown in Panel (A) and collective errors ($P < 0.001$) in Panel (B). The reduction is notably larger and more significant in periods where networks had adapted to the information environment (i.e., rounds [6, 10] and [16, 20]). Errors are normalized with respect to average errors in the *solo* condition. Statistics on individual-level outcomes are based on White cluster-robust standard errors. Error bars indicate 95% confidence intervals.
Adaptive Mechanisms

Two social mechanisms underlie the favorable performance of networked groups. First, dynamic networks adaptively centralized over high-performing individuals. This behavior was predicted by abundant evidence from cognitive science and evolutionary anthropology, which indicate that people naturally engage in selective social learning—i.e., the use of social cues related to peer competence and reliability to choose whom we pay attention to and learn from selectively. Figs. 2-3A and 2-3B show that participants in dynamic networks consistently used peers’ past performance information as success cues to guide their peer choices. As rounds elapsed, performance information accrued, and social networks evolved from fully distributed into networks that amplified the influence of well-informed individuals. Upon receiving an information shock, the networks slightly decentralized, entering a transient exploration stage before finding a configuration adapted to the new distribution of information among participants (see Fig. 2-4).

A centralization mechanism alone could suggest that group members may merely follow and imitate the best individual among them, hence bounding collective performance by that of the group’s top performer. However, research on the two-heads-better-than-one effect indicates that, in the simpler case of dyads, even the best individual can benefit from social interaction; and that the critical mechanism enabling this effect is a positive relationship between individuals’ accuracy and their confidence. Fig. 2-4C shows that participants in dynamic networks had, overall, a positive correlation between the accuracy of their initial estimates and their self-confidence (measured in terms of resistance to social influence). Participants were likely to rely on private judgments whenever these were accurate and likely to rely on social information otherwise. Fig. 2-4C also shows that, as rounds elapsed, participants used task feedback to calibrate their accuracy-confidence relation gradually, and were able to readapt gradually upon the shock. Consistent with prior literature, a positive correlation of confidence and accuracy was found in static networks too (see S18 in SI), explaining their favorable performance compared to
Figure 2-3: Mechanisms promoting collective intelligence in dynamic networks. Panel (A) shows that the network becomes more centralized with time (Freeman global centralization—i.e., how far the network is from a star network). Panel (B) depicts the relation between performance (i.e., average error) and popularity (i.e., number of followers). Panel (C) shows the relationship between accuracy of initial estimate and confidence (i.e., resistance to social influence). Error bars indicate 95% confidence intervals.
unconnected groups.

Figure 2-4: **An illustrative example of the network dynamics.** The circle color represents performance. The size of each circle represents the number of followers (i.e., popularity). The dashed orange line is the distribution of estimates prior to social influence, the blue solid line is the distribution of post-social influence estimates, while the dashed vertical line is the true correlation.

**Mean-Variance Trade-off**

The joint effect of centralization and confidence mechanisms explains the adaptive advantage of dynamic networks. Moreover, it suggests that their collective performance may not be bounded by that of the best individual, and that even the best individual may benefit from network interaction. To test these implications, we generalize the use of group means as collective estimates, common in ‘wisdom of crowds’ studies, and analyze the performance of *top-k* estimates—that is, collective estimates where only the guesses of the *k* best-performing group members are averaged. Top-*k* subsets within each group were computed based on post-hoc individual performances across all rounds. In particular, *top-12* estimates correspond to the group mean, and *top-1* to estimates of the group’s best-performing individual. Fig. 2-5 reports the mean and standard deviation of estimation errors incurred by *top-* *k* estimates during the adapted periods. Ideal estimates would minimize both mean error and variability, approaching the lower left end of the trade-off space.
Figure 2-5: **Mean-variance trade-off.** Mean and standard deviation of absolute errors incurred by top-$k$ estimates during the adapted periods (rounds $\in [6, 10] \cup [16, 20]$). Top-$12$ estimates correspond to the full-group mean, and top-$1$ to the group’s best individual. Within each condition, top-$k$ trade-off curves first gain in both objectives, then trade off lower average error for higher variability, and finally regress in both objectives as $k \to 1$. Across conditions, for any $k \in [1, 12]$, groups in the dynamic condition outperformed groups in the static and solo conditions. Moreover, the full-group mean of dynamic networks averaged 28% lower error and 48% less variability than the best individual playing solo (dynamic top-$12$ vs. solo top-$1$; $P < 10^{-2}$); and the best individual in the dynamic condition averaged 32% lower error and 38% less variability than her analogue in solo (dynamic top-$1$ vs. solo top-$1$; $P < 10^{-2}$). Bars indicate 95% confidence intervals.

The shape of top-$k$ curves reveals that, as we remove low-performing individuals (from $k = 12$ to $k = 1$), estimates initially improve in both mean and standard deviation. Then, as we further curate the crowd beyond $k = 6$, top-$k$ estimates trade off between decreasing mean error and increasing variability, and finally regress in both objectives as $k \to 1$. Comparison across conditions shows that, for any $k \in [1, 12]$, dynamic influence networks improved estimation errors in terms of both mean and standard deviation. In particular, Fig. 2-5 shows that the full-group average in dynamic networks got 28% lower error and 48% less variability than the best individual in the solo groups (dynamic top-$12$ vs. solo top-$1$; $P < 10^{-2}$). Moreover,
even the best individual derived substantial benefits from social interaction, averaging 32% lower error and 38% less variability when forming and revising social connections rather than working in isolation (dynamic top-1 vs. solo top-1; $P < 10^{-2}$).

We implemented numerical simulations to further explore the effects of dynamic influence networks in collective estimation. We simulated interacting agents that update beliefs according to a DeGroot model [41], and rewire social connections according to a performance-based preferential attachment process (see Section S7). Results of these simulations corroborate our experimental results. Dynamic networks adapted to shocks by shifting influence weight to agents with better information, substantially decreasing individual and group error. Results also indicate a relationship between the network learning rate—the network’s sensitivity to changes in agents’ performance—and the arrival rate of environmental shocks (Fig. S21). Networks with faster learning rates could adapt to environments with frequent information shocks. Conversely, networks with slower learning rates could leverage longer learning periods, achieving lower error rates when in environments characterized by infrequent shocks. This short-term versus long-term accuracy trade-off implies that optimal network learning rates depend on the pace at which the information environment changes, analogous to notions of regularization and learning rates in artificial intelligence algorithms [18, 96].

2.3 Discussion

Existing work on collective intelligence and social influence has not considered the rewiring of influence networks, a phenomenon that is widespread in natural situations. We have shown that dynamic influence networks can adapt to biased and non-stationary environments, inducing individual and collective beliefs even far more accurate than the independent beliefs of the best-performing individual. We expect this adaptive systems view on collective intelligence to further sprout connections with fields such as evolutionary biology and artificial intelligence. Our findings also emphasize the relevance of high-quality and accessible social information, and the
relation between the stability of social influence ties and how fast the information environment changes. The insights here provided suggest design guidelines germane to real-world collective intelligence mechanisms, in contexts such as commodity markets, social trading platforms, crowdfunding, crowd work, prediction markets, and online education (e.g., MOOCs).

2.4 Appendix: Experiment Details

We recruited 720 participants using Amazon Mechanical Turk (AMT; see Supplementary Information). Following a between-subjects design, each subject participated only once, was randomly assigned to one of the three conditions, and did not know that other conditions existed. Participants interacted anonymously over the Internet using web-application playable in a browser window. Consistent with standard payment rates on AMT, participants received a show-up fee of $2.0, and then had the opportunity to earn an additional bonus of up to $8 based on their performance in the game—average hourly pay of $15. We excluded one participant from our analysis who was assigned to the solo (control) condition and dropped from the game without completing any of the tasks. Consequently, the analysis of this study is based on 719 valid participants. No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during analysis.

Participants were randomly allocated into groups of size 12, and each group was randomized to one of three treatment conditions: a solo condition, where each individual solved the sequence of tasks in isolation; a static condition, in which participants were randomly placed in static communication networks; and a dynamic condition, in which participants at each round were allowed to select up to three neighbors to communicate with. The initial communication network for the static and dynamic conditions were the same (i.e., each participant had three random incoming edges and three random outgoing ones). At each round, participants were initially asked to submit an independent guess. Then those in static and dynamic conditions entered a social exposure stage, where they could observe the answers of their network peers,
update their own and see peers’ updates in real time. After submitting a final guess, participants in all conditions were shown feedback on their score and the correct answer. Lastly, those in the dynamic condition were shown scores of all participants and allowed to revise which peers to follow in subsequent rounds.

Our experimental task consisted of estimating the correlation from a scatter plot. Participants engaged in 20 correlation estimation tasks. The correlation estimation task was designed so that it allowed manipulation of the task difficulty. We manipulated the difficulty of a scatter plot by varying the number of points, adding outliers, or non-linear relationship between the two variables (see Supplementary Information for visual examples). At every round, all plots seen by participants shared an identical true correlation (i.e., the correct answer for everyone is the same), but difficulty levels could differ among them. The design allowed the simulation of a shock to the participants. Specifically, each participant experienced a constant difficulty level across the first ten rounds; then, at round eleven, we introduced shocks by reshuffling difficulties to new levels that remained constant after that. Participants were not informed about the difficulty levels they or their peers faced.

To evaluate the difference in performance between the different treatment conditions, we used a treatment dummy to estimate a linear regression for the average normalized error (i.e., the average across the 20 rounds). We examined the effects at two levels of analyses: individual outcomes (i.e., one data point per user; n = 719); and 2) group-level outcomes (i.e., one data point per-group). Statistics on individual-level outcomes are based on White cluster-robust standard errors to account for correlated error terms within each group. Statistics on group-level outcomes are based on two-sample t-test. We also performed various statistical robustness checks (see Supplementary Information).

The study was reviewed and approved by the Committee on the Use of Humans as Experimental participants (COUHES) at MIT. All participants provided an explicit consent to participants in this study and COUHES approved the consent procedure.
Chapter 3

Human + Artificial Intelligence for Early Diagnosis of Diabetic Blindness
This third chapter moves forward from Chapter 2 in two directions: 1) towards multi-agent systems composed of human and AI elements, and 2) towards application domains of high societal relevance, and aligned with the Sustainable Development Goals.

Here we focus on studying the interplay of human and artificial decision makers as they strive to solve a socially relevant task: screening diabetic patients according to the presence of early signs of diabetic blindness. In what follows, Section 3.1 argues for the relevance of this problem domain, Section 3.2 describes the human-AI experiment conducted, Section 3.3 presents results, and Section 3.4 concludes.

### 3.1 Relevance: Massifying the Early Diagnosis of Diabetic Blindness

**Diabetes** is among the top public health challenges in the world today. According to the Institute for Health Metrics and Evaluation (IHME) [50], diabetes is one of the top-ten leading causes of death of most countries in Latin America and the Caribbean (LAC), including Mexico, Brazil, Colombia, Venezuela, Peru, Chile, Argentina, Puerto Rico, Jamaica, Belize, Trinidad and Tobago, and others. Furthermore, the trend is upwards: diabetes has climbed up the ranking in the last decade, relegating other causes such as lower respiratory infections.

Particularly relevant to the present work, diabetes is not only a cause of death, but also a major cause of disability. For example, it is the first leading cause of disability in countries like Mexico, Puerto Rico, Jamaica, Trinidad and Tobago, Guyana, and Suriname; and top three in Guatemala, Belize, El Salvador, Dominican Republic, Venezuela, Ecuador, and Uruguay [50]. Among the physical consequences of diabetes, the most prevalent incapacitating complication is visual impairment [1].

**Diabetic retinopathy.** In Mexico for example, 54% of diabetic patients experience vision loss [1]. In diabetic patients, high blood sugar levels often cause damage to blood vessels in the retina—swelling, leaking, obstructing blood flow, or creating abnormal growth. This disease is called *diabetic retinopathy* (DR), which deteriorates
sight, and if not treated adequately, provokes blindness.

Towards Timely Diagnosis in the Developing World

Treatments exist to stop the progress of DR, including the control of blood sugar levels, insulin injections, and laser photocoagulation surgery. However, the eye damage caused by DR is irreversible, and without proper treatment leads to blindness. In countries like Mexico, DR is the leading cause of irreversible blindness, particularly in the working-age population [1, 9].

Hence, early detection and proper treatment are the key to preventing latter DR stages. It is therefore recommended that diabetic patients get screened by an ophthalmic specialist once every 3-12 months [9]. Yet, in most developing countries, due to lack of resources and organization, diabetic patients only reach out to an ophthalmic specialist at advanced DR stages, once their vision is rather deteriorated. In Mexico, although 54% of diabetic patients experience vision loss, only 13% have visited an ophthalmologist [1]. Similarly, a study in Chile showed that only 9% of diagnosed diabetic patients had been examined by an ophthalmic specialist in the previous 12 months [9].

The above context presents a paramount opportunity for AI-assisted systems that enable the massification of early DR screening in low-income populations. In particular, it was recently shown that AI algorithms can classify referable vs. non-referable cases based on images of the retina (called fondus images), with an accuracy comparable to human experts [61, 16, 139]. This technology could enable public health systems to conduct regular DR screenings, by deploying fondus cameras in neighborhood and rural clinics which are already frequented by diabetic patients (as opposed to ophthalmic specialty hospitals).

We envision a distributed network of fondus cameras, coupled with a mix of DR-screening algorithms and ophthalmic specialists in a telemedicine framework. If successful, such system could help society abate substantial public health and economic costs, by massifying the timely detection of DR. Figure 3-1 shows a high-level diagram of such potential H+AI system, where diabetic patients are screened by the AI,
or the AI-assisted human, and referred to ophthalmic specialty hospitals if referable DR levels are found.

Figure 3-1: Hybrid decision system for massifying early diagnosis of diabetic blindness.
3.2 Human + AI Experiment

3.2.1 Research Questions

Motivated by the opportunity described above, we are interested in studying how to best architect human and artificial intelligence towards effective and efficient screening systems for massifying the early detection of DR. The present work investigates the following research questions:

1) **Individual accuracy.** What is the accuracy of an AI screening algorithm compared to the status quo screening performed by human ophthalmologists? In particular, how do these compare in developing countries like Mexico, considering local ophthalmologists, local disease and genetic characteristics, and local processes and equipment for collecting fondus images.

2) **Social learning effects:** Are humans more accurate post exposure to the AI? Conversely, is the AI more accurate post exposure to the human?

3) **Two-heads-better-than-one (THBTO) effect:** Are the human plus the AI together more accurate than either of them in isolation?

4) **Communication depth:** Finally, how do these effects depend on the depth of communication exchanged? From exchanging more shallow confidence-only information, to exchanging deeper information such as attention maps.

Related Literature

There is abundant evidence in the social and computational social sciences, including our own work in Chapter 2 demonstrating commonly-found advantages of multi-agent decision systems over individual decision makers [6, 78, 136, 66, 11, 101].

In human-to-human systems, several mechanisms conducive to collective intelligence have been identified. First, judgments from multiple agents often have uncorrelated errors, hence their aggregation can cancel out inaccuracies and distill improved estimates [120, 85]. This mechanism underlies the traditional notion of the wisdom

---

Note that in this first human experiment, we don’t base results on ophthalmic experts, but on non-expert online participants. We do however work with Mexican patients and image collection processes (as described in Section 3.2.2).
Second, beyond naive aggregation, agents that interact can exchange information about their degree of confidence on each estimation task, allowing adaptive strategies that modulate the influence of their partners’ opinion on their own. Similarly, exchanging deeper information about their decision-making process, such as identifying the pieces of task-related information that are most influential to their judgment, allow to assess the quality of social information, and modulate its influence adaptively. These adaptive social mechanisms are key to collective intelligence, and, most relevantly, have been shown to enable accuracies even beyond that of the most accurate individual [6, 78, 136, 66, 101].

In the vein of hybrid decision systems, academic work has more recently studied the interplay of human and AI agents in estimation tasks. In particular, several studies have focused on studying the degree of trust that humans place on predictive algorithms, and the effects that exposure to algorithmic opinions has on the accuracy of human judgments [140, 107, 141, 74, 80].

Trust. Overall, it is found that humans tend to trust algorithms, even beyond their trust in human peers and human experts [141, 80, 74]. Yet, results from previous work show that the level of trust depends widely on multiple factors, such as: the nature of the task [141, 74, 140], the humans’ education levels and overall attitude towards automation [74], the domain expertise and relative accuracy between the human and the algorithm [141], and the amount and type of information about the algorithm’s opinion that is communicated to the human [80, 107].

Accuracy. In terms of accuracy, not all studies report the performance of the hybrid human-AI duo. The ones that do have found a strict tradeoff between accuracy of the hybrid decision system and the degree of agency left to the human. In other words, the more the algorithm’s opinion prevails over the pre-exposure opinion of the human, the more accurate the decisions are [80, 141]. However, these results most probably depend on the task chosen, and the relative accuracies between the humans and the AI.

Closest in spirit to the present work is that of Lai et al. 2019 [80]. There,
authors evaluate the accuracy of human judgments after being exposed to an algorithm, and how it varies under communication schemes that induce increasing degrees of AI influence (from human autonomy to full AI automation). Notably, the task used—discerning truthful from deceptive hotel reviews—is one where the algorithm is correct 90% of the times, but where the average human is not better than random guessing [80]. Hence, the authors find a stark tradeoff between “human agency” and human accuracy post-AI-exposure.

In summary, human-AI opinion dynamics highly depend on factors related to the human, the AI, the communication interface, and the task. Hence, the potential collective intelligence of hybrid human-AI decision systems must be empirically evaluated. In what follows of this Chapter 3 we report on human experiments that test for such collective intelligence in the context of DR screening.
3.2.2 Experimental Design

We developed a web-based experiment that allows to test the research questions laid out in Section 3.2.1. Participants \( n = 50 \) from Amazon Mechanical Turk (AMT) engaged in a sequence of 30 estimation tasks. Each task consisted of classifying a retinal fondus image in one of two classes: “non-referable” or “referable DR”. Participants were asked to provide both their answer and confidence level, and were rewarded in proportion to their classification accuracy.

The experiment followed a within-subjects design, where each participant was subject to three treatment conditions ordered at random (10 task images under each treatment). In the solo condition, participants responded the task in isolation, without any exposure to the DR-screening algorithm. Conversely, in the confidence condition, participants first were asked to submit their independent estimate, then were exposed to the answer and confidence of the algorithm for that particular task, and finally were asked to submit their revised, post AI exposure answer and confidence level. Finally, the attention condition was identical to the confidence condition, with the exception that, additional to the algorithm’s answer and confidence, participants in this condition could see an attention map—a heatmap highlighting the areas of the image that the algorithm thinks are most indicative of a referable DR. Figures 3-2, 3-3, and 3-4 show screenshots of the experimental platform as seen by participants in the solo, confidence, and attention conditions respectively.

Expertise level. Participants in this first experiment were not ophthalmic experts, but online workers recruited from AMT. Participants were 100% based in the United States, 59% were females, average age was 36, and in terms of education, the highest degree obtained by 41% of participants was high school, 50% had a bachelors degree, and 9% had a masters degree. Because participants are non-experts, we provided them a visual guide for screening the fondus images, based on the international clinical diabetic retinopathy disease severity scales [134]. In all conditions participants were informed about the accuracy of the algorithm in this task (“correct 75% of the times”).
Figure 3-2: **Solo condition.** The participant responds to the task without exposure to the DR-screening algorithm.

Figure 3-3: **Confidence condition.** The participant can observe the answer and confidence of the algorithm (slider on the right).
(a) Not showing the attention map.

(b) Showing the attention map overlaid.

Figure 3-4: **Attention condition.** The participant can observe the answer and confidence of the algorithm (slider on the right), as well as the attention map.
3.2.3 AI Screening Algorithm

We trained a neural network to classify fondus images as referable or non-referable diabetic retinopathy. To train the network we used the EyePACS dataset \[15\]. The dataset contains 88,702 images, from those 19.33% have referable DR. The dataset we used to perform the experiment was a small anonymized sample \((n = 215)\) of fondus images provided by the Association to Prevent Blindness in Mexico (Asociacion Para Evitar la Ceguera en Mexico, APEC), where 66% have referable DR. Table 3.1 shows the overall amount of instances in each DR category.

<table>
<thead>
<tr>
<th>EyePACS</th>
<th>APEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>No apparent retinopathy</td>
<td>65,343</td>
</tr>
<tr>
<td>Mild nonproliferative DR</td>
<td>6,205</td>
</tr>
<tr>
<td>Moderate nonproliferative DR</td>
<td>13,153</td>
</tr>
<tr>
<td>Severe nonproliferative DR</td>
<td>2,087</td>
</tr>
<tr>
<td>Proliferative DR</td>
<td>1,914</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>88702</td>
</tr>
</tbody>
</table>

Table 3.1: EyePACS and APEC distributions

The EyePACS dataset was divided into training, validation, and test sets as done by Voets et al. \[128\]. The training and validation sets together had a total of 57,146 images and the test set had 8,790 (80% and 20% respectively).

The architecture we used in this experiment was based on the VGG16 network \[117\]. The VGG16 architecture has 16 trainable layers and receives images of size 224 \(\times\) 224 in RGB. All convolutional layers have kernels of size 3 \(\times\) 3, and maxpooling layers use 2 \(\times\) 2 windows with strides of size 2. The architecture we used is identical to the VGG16 up to the last max pooling layer. From there we added a dropout layer with drop out probability of 0.5, a dense layer with ReLU activation, another dropout layer with drop out probability of 0.5, and finally a dense layer with sigmoid activation. Figure 3-5 shows the architecture used compared to the VGG16 architecture.

To train the network we initialized it with pretrained weights from the ImageNet dataset \[42\] for the convolutional layers, and random weights for the two dense layers. The learning rate used was \(1e - 5\) and a decay of 0. The code was implemented using the Keras library with the Tensorflow backend \[27\] in Python 3.5 \[127\].
Figure 3-5: Comparison between VGG16 architecture and the architecture used in the current work.
Images were preprocessed by centering the retina, cropping the background, and re-sizing them to $224 \times 224$ as in [128].

The performance of the network in the EyePACS test set has an area under the curve (AUC) of the receiver operating curves (ROC) of 0.95, while in the APEC dataset the AUC is 0.93. The corresponding ROCs are shown in Figure 3-6.

**AI Attention Maps**

In order to make the algorithm’s decision explainable to the user, we created heatmaps that showed how much the different regions of the eye contribute to a referable DR prediction. To construct the heatmaps we used an *occlusion approach*. We defined a sliding window of size $64 \times 64$ and a stride of 4. The images were padded with 63 black pixels in each direction. Predictions were obtained by setting black all pixels that did not correspond to the sliding window. Each pixel was given an importance score corresponding to the average of all the predicted scores of windows that contained that pixel.
The above process yielded a pixel-wise importance map. For presentation purposes, the attention matrix was smoothed by applying a convolution with kernels of size $8 \times 8$, resulting in the final attention heatmaps (shown in Figure 3-7).

Figure 3-7: Fondus image with and without attention map overlaid.
3.3 Results

We computed human performance for each of the three experimental conditions—solo, confidence, and attention—, and divided each by estimates pre-AI-exposure and estimates post-AI-exposure. We measure performance as the simple classification accuracy (percentage of instances correctly classified). Figures 3-8 and 3-9 show results for the focal questions laid out in Section 3.2.1.

**Individual accuracy.** The AI was more accurate than human participants. Participants were, on average, accurate on 70% of the images evaluated (condition solo in Figure 3-8 A); while the algorithm was accurate in 74% of them. Note that the AI's accuracy is lower than reported in Figure 3-6, as we intentionally selected a sub-sample that is challenging for the algorithm.

**Social learning effects.** We find that the AI’s opinion exerted strong influence on the opinion of human participants. On average, participants changed their estimate by .15 after being exposed to the AI’s (Figure 3-9). Most importantly, Figure 3-8 A shows that participants strongly benefited from updating their estimates, on average improving their accuracy from 70% to 75% (grey vs. purple bars). Similarly, Figure 3-8 B shows that the AI can significantly benefit from interacting with a human.
partner, and using a simple social learning strategy—averaging its own estimate and the estimate of its human peer (bars above the green dashed line).

**Two-heads-better-than-one (THBTO) effects.** Figures 3-8 A and B show that both human participants and the AI, after interaction with one another, achieved accuracies above what either could have achieved in isolation (purple bars above the green dashed line), confirming a THBTO effect. This result is salient, as most prior literature has worked in contexts where a strict tradeoff between accuracy and human agency has been found [80, 141].

**Communication depth:** Finally, we did not find significant differences in the magnitude of influence, nor the post-interaction accuracies, induced by different depths in the information communicated by the human-AI dyad. Figure 3-9 shows similar influence levels across confidence and attention treatments, and Figure 3-8 shows similar post-exposure accuracies (blue bars).

![Figure 3-9: Social influence magnitudes across conditions.](image-url)
3.4 Conclusions

This Chapter 3 takes steps forward from Chapter 2 in two directions: 1) towards multi-agent systems composed of human and AI elements, and 2) towards application domains of high societal relevance, and aligned with the Sustainable Development Goals.

We confirmed the existence of significant synergies derived from the interaction of the human and AI dyads. In particular, human accuracy improved after interaction with the AI, and likewise AI’s accuracy improved after interaction with its human peers. Most relevantly, post-interaction accuracies outperformed the accuracy that either human or AI agents could achieve in isolation, confirming a *two-heads-better-than-one* effect between the human and the machine.

However, these results are only preliminary in terms of the social impact goals motivated in Section 3.1. In particular, the study here reported relied on non-expert human participants. It is expected that participation of expert ophthalmologists will induce different human-AI dynamics compared to non-experts, as seen in [141] in the context of geopolitical forecasting. On the one hand, experts are expected to be more accurate and confident about their estimates; yet on the other, their confidence might be better calibrated, allowing them to harness further human-AI synergies. Looking ahead, we are already collaborating with researchers from the APEC hospital in Mexico, towards replicating this experimental design with ophthalmic specialist participants.

Finally, towards real-world impact, we should couple these fondus image-based screening systems with an *a priori* targeting system—as proposed in [92]—in order to determine, for example, where a large-scale DR screening system should deploy its fondus cameras, and which high-risk individuals in those locations to test. The following chapter delves into the development of large-scale targeting systems, and Section 7 elaborates on potential future work coupling targeting tools with image-based screening, towards the massification of early diagnosis in developing countries.
Chapter 4

Human + Artificial Intelligence for Fair and Efficient Targeting of Social Policies
This chapter presents an in-depth study focused on an impactful application domain: the targeting of social development policies. It offers a comprehensive view: from the generation of accurate AI models, to the development of a decision-support platform for distributed governance, enabling a diversity of institutions to customize the use of AI-based insights into their targeting decisions.

Section 4.1 argues that, due to the scale, diversity, and global spread of targeted social policies, the algorithms that determine or influence decisions regarding who benefits from them, and who doesn’t, are among the top algorithms of paramount relevance operating in the world today.

Section 4.2 shows that AI-based approaches offer superior accuracy than status quo poverty targeting systems. The error reductions of 20-30% in two countries mean that, shifting to an AI-based approach, nearly a million people in poverty currently misclassified could be included in the countries’ set of targeted social policies, without increasing their budget.

Section 4.3 goes beyond global accuracy and studies predictive performance disaggregated by population subgroups. There we show that: 1) the AI-based approach not only improves global accuracy, it reduces the subgroup-specific errors for all subgroups defined by four segmentation attributes, in both countries. 2) The AI-based approach has a strong effect in narrowing disparities in predictive performance across subgroups. And 3) although narrower than the status quo, significant disparities were still found for the AI-based approach in terms of predictive performance across subgroups, highlighting the need for algorithmic fairness considerations.

Lastly, Section 4.4 describes the multi-stakeholder architecture of real-world social targeting decisions, and the tensions that arise from heterogeneous preference over prioritization criteria and notions of fairness. Hence, we conclude by proposing an approach to design fair targeting rules under a distributed governance paradigm, by means of an interactive decision-support platform designed for non-technical experts.
4.1 Worldwide Relevance of Targeted Social Policies—and their Underlying Algorithms

Algorithmic decision-making systems (ADS) have become increasingly ubiquitous—e.g., in criminal justice [76], medical diagnosis and treatment [75], human resource management [26], social work [56], credit [69], and insurance. Although there is widespread excitement for the potential societal benefit that this type of technology can bring, there is also commensurate concern about how it can deepen social inequalities and systematize discrimination [103, 100]. Consequently, substantial work has surged in recent years, on conducting fairness audits of deployed systems, as well as on defining and optimizing for algorithmic fairness. Notably, this rising field of research has majorly been motivated by and focused on developed-world and US-centric contexts: online domains such as targeted advertising [122], search engines [57], and face recognition algorithms [21]; and offline domains such as the criminal justice system [32, 13], child maltreatment [33], and predictive policing [113].

Algorithmic targeting of social policies. The present work focuses on targeted social development policies, i.e.: policies that promote social development outcomes, and target only a subset of the population—most commonly, populations in poverty or extreme poverty. Since two decades ago, algorithmic rules underlie the targeting decisions of a large fraction of social policies in the developing world, e.g., poverty prediction algorithms that precondition eligibility to cash transfer programs [70, 63, 48]. Here, we argue that the algorithms that influence or determine critical decisions regarding who benefits from a targeted social policy, and who doesn’t, at a large scale and across the global south, are without a doubt among the top ADS of paramount importance operating in the world today.

Diversity. Targeted social policies constitute a major vehicle for fighting poverty and redistributing wealth across the developing world [63, 8]. In most countries, targeted programs run by governments and NGOs proliferate, touching every corner of social development: from cash transfer programs, to scholarship programs, subsidized health care systems, targeted housing and energy subsidies, targeted childcare, food
security, retirement pension programs, targeted microloans, and others. Moreover, internationally, aid agencies such as USAID require all their funded institutions to guide the selection of beneficiaries based on poverty assessment tools [25], in an effort to promote interventions’ effectiveness and combat corruption.

Scale and Spread. To gauge the scale and global spread of targeted policies, consider the case of cash transfer programs (CTs). CTs provide a financial stipend to families in poverty, and are often conditional, i.e., requiring beneficiaries to comply with “co-responsibilities”, such as maintaining children in school, and attending regular medical appointments [36, 48]. More than 110 countries worldwide have implemented national CTs (see a map in Figure 4-1 [8]), playing a central role in the countries’ poverty alleviation and wealth redistribution strategies. Only in Mexico and Brazil, for example, national CTs reach more than 80 million people (roughly 25% of the population), distributing +8 billion USD yearly (0.3% of GDP) [70].

The present work is motivated by the scale, diversity, and widespread relevance of targeted social policies like CTs. Ultimately, we aim at contributing to the development of better targeting systems, and realizing the substantial positive impact these may bring to society in terms of increased fairness, accuracy, and transparency.
4.2 The Accuracy Edge: Large-Scale Impact of AI-Based Targeting

4.2.1 Algorithmic Targeting of Social Policies

Most social policies, like CTs, are targeted to the poor. However, in the developing world, reliable income data is typically not available and costly to procure, because households in the target population participate mainly in informal economic markets. Hence, targeting most commonly relies on poverty prediction algorithms that estimate households’ eligibility based on observable and less costly proxy data, such as education levels, demographics, and the assets and services of households. In most countries, these algorithms are trained to estimate poverty based on large, periodic, statistically representative household surveys, which collect both the proxy features and the income ground truth. In their daily operations, institutions can’t collect ground truth income data directly from potential beneficiaries, due to the cost of eliciting trustable income data at such massive scale, as well as candidates’ strong incentives for underreporting.

In practice, predictive algorithms are imperfect, leading to targeting errors. In Latin America and other regions, it is estimated that targeting systems incur more than 25% of both exclusion errors and inclusion errors (false positives and false negatives). However, the methods used during the past two decades for estimating poverty have relied on econometric approaches which are not optimized for out-of-sample prediction, and cannot leverage the predictive value of high-dimensional data. Hence, we hypothesize that substantial accuracy gains may exist from the use of modern computational and statistical methods borrowed from the field of artificial intelligence. In what follows of this section, we investigate the comparative performance of AI-based targeting systems against the status quo.

Related work. Closest in spirit to the work presented in this Section is the work of Mcbride et al. 2016. In it, it’s shown that out-of-sample estimation can reduce exclusion errors (undercoverage), but increase inclusion errors (leak-
age), compared to in-sample linear methods traditionally used in poverty prediction tools. However, they found no consistent advantage from machine learning methods (forests) over the cross-validated linear methods. Our results in this section partially contradict results in [91], showing that AI-based predictive approaches substantially outperform the status quo methods, in both exclusion and inclusion errors. This partial discrepancy is most likely due to the fact that [91] works on small samples \( n \in [1800, 11280] \), doesn’t perform feature engineering, doesn’t conduct metaparameter search to control overfitting, and doesn’t compute the entire exclusion-inclusion error curves when comparing targeting approaches.

### 4.2.2 Empirical Evaluation

#### Data

The present study uses publicly-available data from two countries: Costa Rica, and Colombia. We chose these countries as they represent populations of medium (47M) and smaller (5M) sizes, characteristic of Latin American countries.

**Household surveys data.** In particular, the present study is based on data from household surveys that are collected annually by the countries’ national statistical offices; the same data on which status quo targeting systems are trained and evaluated [70, 3]. These surveys constitute one of the most important information instruments in the countries, based on which the main poverty, prices, labor, and other socioeconomic indices are computed. The surveys are carefully designed to be statistically representative of the population, and collect ground truth income data, as well as most relevant household information, such as socio-demographics, living conditions, education, assets, and services (information & communication services, utilities, sanitary, etc.). The ground truth income data in these surveys is widely considered the best income data available in the countries—as compared to income from census data and other surveys—due to the robustness of the sampling methodology, extensiveness of the questionnaire applied to each household, professionality of the field social workers, and the lack of incentives to under-report.
<table>
<thead>
<tr>
<th></th>
<th>Poverty ratio</th>
<th>Years</th>
<th>Total Sample Size</th>
<th>Population Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Costa Rica</td>
<td>22%</td>
<td>2015-18</td>
<td>22 k</td>
<td>4.9 M</td>
</tr>
<tr>
<td>Colombia</td>
<td>28%</td>
<td>2016-17</td>
<td>462 k</td>
<td>47.6 M</td>
</tr>
</tbody>
</table>

Table 4.1: Household survey data statistics.

Table 4.1 presents basic characteristics of the datasets for each country.

**Algorithms**

We aim at assessing potential advantages of modern statistical and computational methods, borrowed from the field of artificial intelligence, compared to the status quo methods used to prioritize households in terms of income poverty.

**Status quo methods.** The status quo predictors used for income-based targeting of social policies—which have been used over the past two decades, and still prevail today—stem mainly from econometric methods \[25, 70\] more suitable for causal inference than predictive inference. Linear regressions are by far the most common method of choice, particularly quantile linear regressions \[77\], as it’s often found empirically that these provide a slight accuracy edge over the former (due to robustness to income outliers). In what follows of this work, we’ll refer to quantile linear regressions as the status quo methodology.

**AI-based methods.** We implement a number of algorithms based on the AI paradigm of machine learning. In particular, the methodological elements expected to yield increased predictive performance compared to the status quo are: 1) feature engineering, 2) regularization, and 3) better approximator models.

Discussing the performance of different AI-based predictors is off the scope of this work. Here we describe only the best-performing methodology, which consists of:

- **Feature engineering.** We preprocessed the survey data to generate three types of features. First, *expert features*, generated by human experts, such as the ratio of people over rooms in the households, and the age of the head of the households. Second, *statistical features*, including means, modes and entropies for all individual-level variables of household members, such as age, gender, and
education. Lastly, deep features, generated by a recursive neural network that condenses information of the individual-level features into a one-dimensional encoding—a technique akin to the AI subfield of multiple instance learning (MIL) [23]

- Predictive algorithm. The predictive algorithm that provided best performance was a gradient boosting classifier, trained on the three feature sets concatenated. We used k-fold cross-validation to set the models’ meta-parameters.

In what follows of this work, we’ll refer to the above combination of feature engineering and predictive algorithm as the AI-based methodology.

Accuracy metrics

The two key metrics used by institutions in the social sector to assess the quality of their targeting are exclusion and inclusion errors [70, 63]. Exclusion errors measure the percentage of poor households incorrectly classified as non-poor, denoted by $\epsilon^e = \frac{TP}{TP+FN}$; while inclusion errors measure the percentage of non-poor households incorrectly classified as poor, denoted by $\epsilon^i = \frac{FP}{TP+FP}$. TP, FP, TN, FN correspond respectively to true positives, false positives, true negatives, and false negatives. In what follows of this work, we’ll compare the accuracy performance of alternative targeting systems based on these two measures.

Exclusion-Inclusion Curve. Targeting rules are composed of two elements: a poverty-based score provided by the poverty predictors, and an acceptance threshold above/below which candidates are accepted. Exclusion and inclusion errors are dependent on the acceptance threshold applied. Hence, for comparison of poverty prediction methodologies, we compute the entire exclusion-inclusion curve (EIC), which maps the entire space of targeting rules that a predictive methodology enables: from the universal program, to all sizes of targeted programs, and down to the non-existent program (similar to ROC curves prevalent in the health sciences and machine learning literature). Thereby, EICs map the fundamental tradeoff between exclusion and inclusion errors. Figures 4-2 and 4-3 present instances of EICs.
Lastly, all accuracy measures presented in this work are computed out-of-sample, and 95% confidence intervals are computed non-parametrically by means of bootstrapped resampling.

4.2.3 Results

Figures 4-2 and 4-3 compare the accuracy performance of the AI-based methodology vs. the status quo. For each method, the exclusion-inclusion curve (EIC) is plotted. The EICs map the entire set of targeting rules that each method enables, and the fundamental tradeoff between exclusion and inclusion errors. The upper left corner corresponds to programs with no beneficiaries (100% exclusion and zero inclusion error). The lower right corresponds to universal programs (zero exclusion and max inclusion error). In between, solutions correspond to all targeted programs ranging in sizes from zero to the population size.

In particular, point $a$ on the status quo curve corresponds to the targeting rule with an acceptance threshold $t_a$ that accepts a number of beneficiaries equal to the country’s poverty rate. It is also the acceptance threshold that balances exclusion and inclusion errors (i.e., $e_{exc}^{sq}(t_a) = e_{inc}^{sq}(t_a)$). Similarly, point $b$ on the AI-based curve corresponds to the threshold $t_b$ that accepts a number of beneficiaries equal to the country’s poverty rate, and which balances the errors $e_{exc}^{ai}(t_b) = e_{inc}^{ai}(t_b)$. Let $s(t)$ denote the size of the accepted population of a targeting rule with threshold $t$. Because $s^{sq}(t_a) = s^{ai}(t_b)$, the difference between points $a$ and $b$ provides a rather meaningful assessment, as it compares the performance of the two methodologies when both are constrained by a constant budget equal to the amount of poor in the population.

Points $c$ and $d$ on the AI-based curve correspond to solutions with either equal exclusion errors ($e_{exc}^{ai}(t_c) = e_{exc}^{sq}(t_a)$), or equal inclusion errors ($e_{inc}^{ai}(t_d) = e_{inc}^{sq}(t_a)$), than point $a$ on the status quo curve. Finally, if we want to compare performances irrespective of any particular threshold level, but averaged across all, we compare the areas under the inclusion-exclusion curves (AUEICs).\footnote{This metric is exactly equivalent to the area under the precision-recall curve, also known as}
Figure 4-2: Exclusion vs. Inclusion errors curve.
Grey bars around the red points indicate 95% confidence intervals, computed non-parametrically via bootstrap resampling.

Figure 4-2 and Table 4.2 present the accuracy comparison for Colombia. It is shown that the AI-based method dominates the status quo by a wide margin, and along the entire exclusion-inclusion error curve. This means that for any given budget, a social program will incur less exclusion error and less inclusion error if it targets based on the AI method. In particular, comparing points \(a\) and \(b\), a social program or policy accepting an amount of candidates equal to the country’s poor population, would reduce both, its exclusion errors by 19.7%, and its inclusion errors by 19.8% average precision.
Figure 4-3: **Exclusion vs. Inclusion errors curve.** Grey bars around the red points indicate 95% confidence intervals, computed non-parametrically via bootstrap resampling.

(Table 4.2), if it switches to the AI-based method. Most importantly, this reduction in errors means that 828,940 people in poverty, previously misclassified, would now be correctly included in the country’s set of social protection policies.

Figure 4-3 and Table 4.2 present analogous results for Costa Rica, where a social program or policy accepting an amount of candidates equal to the country’s poor population, would reduce both its exclusion and inclusion errors by 31.6% (Table 4.2), if it switches to the AI-based method. The reduction in errors would mean that 144,306 people in poverty, previously misclassified, would be now correctly included.
in the country’s set of social protection policies.

\[
\begin{align*}
\text{Reduction in Exclusion Error} & \quad \text{Reduction in Inclusion Error} & \quad \text{Increase in Poor Population Covered} \\
@\text{constant budget} & \quad @\text{constant budget} & \quad \text{@constant budget} \\
\frac{\epsilon_{\text{exc}}^sq(t_a) - \epsilon_{\text{exc}}^ai(t_b)}{\epsilon_{\text{exc}}^sq(t_a)} & \quad \frac{\epsilon_{\text{inc}}^sq(t_a) - \epsilon_{\text{inc}}^ai(t_b)}{\epsilon_{\text{inc}}^sq(t_a)} & \quad \\
\text{Colombia} & \quad 19.7\% & \quad 19.8\% & \quad 828,940 \\
\text{Costa Rica} & \quad 31.6\% & \quad 31.6\% & \quad 144,306 \\
\end{align*}
\]

Table 4.2: Comparative Accuracy Results and their Impact on Coverage of the Poor.

76
4.2.4 Inter-Temporal Robustness and Sensitivity Analysis

We conducted two types of complimentary analyses: inter-temporal robustness and sensitivity analysis.

**Inter-temporal robustness.** The accuracy analysis presented so far has been based on cross-validated out of sample evaluation. However, it assumes that the \((X, Y)\) distribution remains constant from training time to deployment time. In reality, poverty estimation models are trained with data from a couple of years past, and deployed on the subsequent year. Moreover, the training/deployment cycle is at least one year long.

Hence we report on an additional inter-temporal accuracy study. In particular, Figure 4-4 compares results for Colombia, for: a) an inter-temporal model trained on years 2015-2016, and tested on 2017 data; vs. b) the regular setup, trained and tested on years 2016-2017 with cross-validated out of sample evaluation. Figure 4-4 shows that there is virtually no difference between the exclusion-inclusion error curves of both setups, indicating that there is inter-year stability of the \((X, Y)\) distribution across a one year leap.

**Sensitivity analysis.** Moreover, it is relevant to know how the model depends on its input variables. This analysis is particularly relevant for two reasons. First, it is known that household surveys applied to elicit the proxy information \((X)\) variables) have some degree of measurement error. Hence, a model that depends only on a few variables, or a model that depends heavily on variables that tend to have high measurement error, will yield noisy outputs. Second, it is desired for all alternative models evaluated to have a common layer of explainability, so that diverse stakeholders can familiarize with the high-level rational of the input-output mappings.

Hence, we conducted a sensitivity analysis where the input variables were systematically perturbed, and the corresponding changes in the output poverty score were measured. Critically, we conducted this analysis taking together the entire analytic pipeline, including the pre-processing, feature engineering, and prediction model described in Section 4.2.2 The perturbation distribution implemented was uniform on
the support of each variable, and independent across variables. Notably, this type of sensitivity analysis is independent from model types, therefore can be applied to all proposed model alternatives.

We constructed an interactive visualization for stakeholders to explore models’ sensitivity results, called the sensitivity map. Figure 4-5 shows the results for the AI algorithm in Figure 4-3. The area of each rectangle represents the average magnitude of the effect that perturbations of a particular variable have on the poverty score, normalized to sum 1. It also shows a hierarchy of categories, which add a higher-level of interpretability. Categories are user-defined, and can be for example: socio-demographic variables (like age, education and gender of household members), house

Figure 4-4: Inter-temporal exclusion vs. inclusion errors curves.
construction variables (type of walls, number of rooms, etc.), house services variables (electricity, water, internet, etc.), etc.

From Figure 4-5 we observe that the AI algorithm is fairly well distributed, with a low Herfindahl-Hirshman concentration index (HHI = .05) [110]. We also see that the variable with the maximum sensitivity accounts for only 10.7% of sensitivity, and that categories 1 and 4 have the most impact on the score.

![Sensitivity Analysis, concentration (HHI) = 0.05](image)

Figure 4-5: **Sensitivity map.** Screenshot of the interactive visualization. The area of each rectangle represents the average magnitude of the effect that perturbations of a particular variable have on the poverty score. The visualization interactively displays information of variables and categories as the user hovers over them.
4.3 Disparities in Predictive Performance across Population Subgroups

Section 4.2 shows that AI-based targeting systems for social policies can be substantially more accurate, enabling government institutions and NGOs to increase the coverage of their target population while maintaining budget constant, or reduce budget while maintaining coverage constant.

However, these results are aggregates across the population as a whole, which could potentially hide systematic disadvantages to minorities or other population subgroups. In this Section 4.3 we study and compare the performance of AI-based vs. status quo targeting systems, computing error rates not at the population level, but disaggregated into subgroups defined by relevant household characteristics, such as urban/rural, geographic region, gender of the head of the family, and family type.

Relevance. To the best of our knowledge, this is the first time that algorithmic rules of this sort are audited for potential exclusion disparities across population subgroups. As argued in Section 4.1, algorithms that determine or influence such critical decisions as inclusion into social protection programs, at such a massive scale, and global spread, should without a doubt be top in the agendas of algorithmic fairness advocates.

Contributions. In summary, 1) we show that the AI-based targeting reduces the errors local to every single population subgroup, in both countries, rendering the paradigm switch uncontroversial; 2) we find substantial disparities in exclusion errors across population subgroups, in both methods, and for most segmentation attributes in both countries; and 3) this error disparities across subgroups are consistently and substantially narrowed by the shift from the status quo method to the AI-based.

4.3.1 Methodology

We extend the accuracy analysis in Section 4.2 to the subgroup level. For each subgroup, and each method (i.e., status quo and AI-based), we computed the exclusion
errors by setting acceptance thresholds corresponding to the poverty rate (base rate) of each subgroup. This choice of thresholds is analogous to the analysis in Section 4.2 and emulates the most prevalent practice, where social institutions set acceptance thresholds that admit a number of beneficiaries equal to the poverty rate. For example, figures 4-6(a) 4-6(d) and show results for the exclusion error rates of each urban/rural subgroup in both countries.

We then compare the performance across methodologies, by computing the percentage of improvement of the AI-based method over the status quo. Analogous to improvement rates in Section 4.2 if the status quo and AI-based systems yield exclusion rates of 20% and 25% respectively for a given subgroup, then we report that the AI-based achieves a 25% reduction in exclusion errors compared to the status quo for that subgroup. For all charts, we report 95% confidence intervals on each estimate by repeating the procedure \( n \) times with bootstrapped re-sampling \( (n = 500) \).

Beyond assessing the subgroup-level benefits of transitioning to the AI system, we also want to assess the disparities among subgroups that exist within each method. We define the subgroup disparity metric as the standard deviation of the exclusion errors across subgroups, i.e., 
\[
\sigma_A = \sqrt{\sum_{g \in A} (e_g - \bar{e})^2},
\]
where \( A \) is a segmentation attribute such as family size, and \( g \in A \) are the subgroups that it defines.

4.3.2 Results

Subgroup-specific improvements.

The right column in Figure 4-6 shows the subgroup-specific improvements in exclusion errors achieved by the AI-based over the status quo, for three selected examples of country-attribute pairs. It is observed that the AI-based reduced errors in all cases, with the minimum reduction being rural households in Colombia, with 10% improvement, and the maximum improvement being unipersonal households in Costa Rica, with improvements above 50%.

Tables in 4-7 and 4-8 summarize results for all combinations of the two countries and four segmentation attributes: urban/rural, gender of the head, family size, and
The third row in each table reports percentual reductions in the group-specific errors for each subgroup. Overall, we find that the AI-based reduced the errors of each and every subgroup, across the four segmentation attributes and two countries. The minimum improvement was 4.8%, for region 27 in Colombia. All subgroup improvement results were significant with 95% confidence.

Disparities across subgroups.

We are interested in knowing: a) whether or not there are significant disparities in predictive performance across population subgroups; and b) whether or not these disparities are more salient for one method or another.

The left and center columns of Figure 4-6 show that substantial imbalances do exist in predictive performance across subgroups. For example, figures 4-6(a-b) show that, in Colombia poor households in urban areas are more likely to be misclassified by either algorithm. Similarly, figures 4-6(g-h) show that, in Costa Rica, poor unipersonal households are more significantly more likely to be misclassified than households with three members or more, by either method.

However, we also observe the AI-based has a strong effect not only in reducing errors overall, but also on balancing error disparities across subgroups. It is observed that the subgroups where the AI-based achieves most improvement, are exactly the ones where the status quo performs worst. In other words, the overall accuracy gains observed in Section 4.2 tend to be driven by the subgroups were the status quo is least accurate.

Tables in 4-7 and 4-8 summarize results for all combinations of the two countries and four segmentation attributes. The unbalance column in each table reports the group unbalance measure ($\sigma_A$) for group-specific errors for country, attribute $A$, and method. Moreover, third rows report the percentual reduction in disparities achieved by the AI-based over the status quo. Overall, we find that disparities do exist for most country-attribute pairs, and for both methods. However, we consistently find that the AI-based has a substantial positive effect, narrowing such disparities in all cases (the only non-positive improvements are in the three cases where the errors are
already balanced, and hence the percentual change is not statistically significant).

4.3.3 Section Conclusions

The above results provide three key results. First, AI-based targeting systems reduced predictive errors for each and every population subgroup, according to the four segmentation attributes, and in both countries. Because all subgroups result benefited, the AI-based Pareto-dominates the status quo at the subgroup level. This result is rather positive, as it renders uncontroversial a decision to transition from the status quo to the AI-based targeting system.

Second, substantial disparities were found across population subgroups, for both methods. These disparities were dependent on the particularities of each country and sensitive attributes, hence supporting the case for requiring empirical evaluations and fairness audits to understand subgroup-level performance prior to deploying algorithmic targeting systems. Lastly, although both methods entailed disparities in prediction errors across population subgroups, the AI-based had a strong effect in balancing errors and narrowing disparities compared to the status quo, supporting again the case for a transition to AI-based methods.
Figure 4-6: **Subgroup-specific performance improvements.** Improvements achieved by the AI-based over the status quo, for three selected examples of country-attribute pairs. All subgroups substantially benefited from switching to the AI-based method. Error bars denote 95% confidence intervals computed non-parametrically via bootstrap sampling.
Figure 4-7: Group-specific comparative results for Colombia. Third rows report % improvements of AI-based vs. status quo methods. First columns show unbalance measures $\sigma_A = \sqrt{\sum_{g \in A} (e_g - \bar{e})^2}$ for each attribute-method pair. All results are statistically significant at 95% except noted by $^{ns}$. 
Figure 4-8: **Group-specific comparative results for Colombia.** Third rows report % improvements of AI-based vs. status quo methods. First columns show unbalance measures $\sigma_A = \sqrt{\sum_{g \in A} (e_g - \bar{e})^2}$ for each attribute-method pair. All results are statistically significant at 95% except noted by ns.
4.4 Fair Targeting and Distributed Governance via Interactive Decision Support

Sections 4.2 and 4.3 demonstrate that the AI-based targeting system provides superior accuracy, both at the global and subgroups levels. Moreover, it is shown that the AI-based compares favorably against the status quo in terms of reducing disparities in prediction performance across population subgroups. Unfortunately, these results do not entail that the AI targeting system is fair in absolute terms. On the contrary, Figure 4-6 and tables in 4-8 and 4-7 show that substantial performance disparities exist, demanding careful reflection regarding algorithmic fairness, and potential fairness mechanisms to be implemented, before the operational deployment of any of these systems (note, of course, that the status quo has been deployed in multiple countries for more than two decades [25][70]).

The present Section 4.4 describes the multi-stakeholder architecture of real-world social targeting decisions (Subsection 4.4.1), the existence of heterogeneous preferences and disagreement over prioritization and fairness criteria (Subsection 4.4.2), and a proposed framework for fair decision-making and distributed governance (Subsection 4.4.3), currently being piloted at the national level in two countries.

4.4.1 The Architecture of Real-World Social Targeting

In many countries, like Costa Rica, Colombia, and Panama, a unified poverty index is constructed by the central government, establishing a common methodology for assessing the poverty status of households in the country [70][3]. Similarly, agencies like USAID establish a per-country poverty assessment tool [25]. This central methodologies are then consumed in a distributed manner by a wide diversity of social programs and NGOs in the country, serving as core and unified criterion for prioritizing and selecting beneficiaries. The use of the central methodology is most often mandated by law, in an effort to focus public funds on the population segments most in need, as well as avoiding manipulation of public funds for political or private interests [30][25].
However, the centralization of prioritization and selection criteria creates strong tensions across the diversity of institutions mandated to consume the central index. These institutions include everything from cash transfer programs, to subsidized healthcare systems, public child care services, scholarship programs, pensions to non-formal workers, etc. They also include a range of geographic scopes, encompassing the national, regional, and local levels of government.

Most often, these institutions demand more complex prioritization and selection criteria, beyond a monolithic income poverty score. Many struggle to accommodate additional criteria that respond to their policy goals, and to the particular characteristics of their social programs and target population. For example, while cash transfer programs may want to focus mainly on household income poverty, scholarship programs need to incorporate distance to education facilities as key selection criterion. Similarly, subsidized healthcare systems need to incorporate the quality of alternative healthcare services accessible to potential beneficiaries. Moreover, social programs often attempt to incorporate contextualized preferences over population subgroups, such as indigenous population, population with disabilities, and victims of violence and forced migration.

Hence, targeting decisions are the compound product of a central government or agency issuing the guiding criterion for prioritization; and a diversity of social programs consuming that index, applying acceptance thresholds based on it, and most often complimenting it with additional selection criteria like multidimensional poverty, and positive/negative preferences across population subgroups.

4.4.2 Heterogeneous Preferences and Disagreement Over Prioritization and Fairness Criteria

It is in this context of centralized vs. distributed preferences and decision making, that we must reflect on what fair decision-making is, and how we can design systems that support it.
Tensions Over Prioritization Criteria

So far, attempts to resolve the tension between centralization and decentralization have been dissatisfying. On the one hand, one type of solution has been to propose, and in some cases implement, multidimensional poverty indices as the centrally-prescribed instruments for prioritization and selection of beneficiaries [111, 5, 30]. However, centrally-defined multidimensional approaches have failed to resolve the centralization-decentralization tensions. In most cases, it has been concluded that it is more convenient for individual institutions to complement a unified, well-defined income poverty index, than to force all institutions to adopt an index that bundles a number of socioeconomic deprivations that are not relevant to all.

On the other hand, additional to centrally-defined multidimensional poverty indices, many institutions attempt to circumvent the rule of a central income poverty criterion by means of diverse ad hoc mechanisms that introduce additional prioritization criteria, often carrying risk of incurring administrative penalties for deviating from a mandated central rule.

Tensions Over Fairness Criteria

Even if agreement existed on what the appropriate criterion for prioritizing households is, there is no consensus on what fair decision-making based on that criterion is.

There are at least three salient fairness criteria discussed in the academic literature applicable to the context of selecting beneficiaries of social programs:

- **Demographic parity**, which requires the selected group to be composed by equal amounts of individuals from each subgroup [65]. For example, a 50/50 gender composition in a president’s cabinet would be an instance of demographic parity.

- **Threshold parity**, which defines fair classification as that which simply applies the same acceptance threshold to all individuals in the population [31].

- **Error rate parity**, which defines fair classification as that which guarantees equal classification error rates across relevant population subgroups. Here in
particular, we consider the special case of equal opportunity, which requires only parity in exclusion error—or equivalently, parity in coverage of the positive class—, because the context of targeting social benefits is a canonical example of positive interventions [65].

On the one hand, it has been shown that fairness criteria like demographic parity, threshold parity, and error rate parity are most often incompatible with one another [76, 106]. On the other hand, there is no consensus in the academic literature about the appropriateness of one of these over the rest. Moreover, we have found from our interactions with multiple institutions operating social policies, that there is also no consensus across institutions about which fairness criterion should overrule the rest.

Positive discrimination. Even more fundamental, there is no consensus about parity per se, regardless of the underlying fairness criterion, being the a priori desirable fairness property for social targeting systems. On the contrary, positive discrimination prevails. Institutions targeting social programs commonly prioritize population subgroups that are considered socially disadvantaged, guided by a paradigm of positive discrimination and affirmative action, rather than parity. Examples of these include victims of violence and forced migration in Colombia, and households with a member with disabilities in Costa Rica.

Hence, it isn’t clear that one should, for example, attempt to re-calibrate a centrally-defined poverty score to balance exclusion errors a priori, unless there is certainty that such balance conforms to the policy goals and ethical frameworks of all institutions that will build targeting rules based on it.

4.4.3 An Interactive Decision Support Platform for Fair Targeting and Distributed Governance

We developed an interactive decision support tool with the following design goals, addressing the opportunities described in sections 4.1-4.2, and challenges described in sections 4.3-4.4:

1. **Accuracy**: Empower social institutions to leverage highly accurate AI-based
predictors on which to build their targeting rules, without requiring technical expertise in statistics and machine learning.

2. **Distributed Governance**: Allow institutions to resolve tensions among different prioritization criteria and among fairness criteria, in a decentralized manner; according to their own policy goals, ethical frameworks, and budgetary constraints.

3. **Awareness**: Educate stakeholders in understanding the space of possible targeting rules available to them, and particularly on the fundamental tradeoffs among different prioritization and fairness criteria.

4. **Inclusiveness**: Enable wider multi-stakeholder discussions over the space of options.

5. **Transparency**: Increase transparency and auditability of deployed targeting rules and their rationales.

The typical usage of the platform and its features transcurrs as follows. We’ll use the working example of a nation-wide cash transfer program in a Latin American country. With the start of the fiscal year, the institution’s prioritization committee has to select which of the 100,000 applicants will take one of the 65,000 available spots this year. The committee session starts, the decision support tool is projected on the meeting room’s screen, and an officer/user logs into her institutional account. First, the platform asks her to specify their:

- **Prioritization criteria**: the core prioritization criteria to use, e.g., a centrally-defined, AI-based income poverty score as constructed in Section 4.2.

- **Segmentation attributes**: the categorical attributes by which they might choose to segment the population, e.g., urban/rural, family type, and membership to a vulnerable group (e.g., disability, or victims of violence and forced migration).
• **Budget:** their institutional budget for this year, i.e., 65,000 available spots; and

• **Hard filters:** any hard exclusion criteria to apply, according to the target population of the institution and its program. E.g., a program focusing on promoting children’s health and education outcomes, will exclude households without children.

After defining the above key elements, the user moves to the main interactive tool for designing targeting rules. The interface is composed of three main elements: 1) the **population tree**, which displays—potentially for each population segment—the acceptance thresholds set, and the percentage of coverage that the thresholds entail (see Figures 4-9 and 4-10), 2) the **thresholds adjustment window** (see Figure 4-10), and 3) the **stacked bar**, which displays the segmented costs vs. the institution’s budget (Figures 4-9-4-12).

The key functions supported by the platform are:

**a) Thresholds adjustment.** Adjusting acceptance thresholds is the basic unit of decision support that the platform provides. To adjust the acceptance threshold of the population, or any population segment, the user clicks on the corresponding tree node, which pops-up the thresholds adjustment window (Figure 4-10). In it, the user can see the distribution of the population or segment in terms of the prioritization variable, and how the threshold divides the distribution in accepted/non-accepted population. The user can then adjust the threshold, interactively observing the impact that threshold shifts have on: 1) the % of coverage, which corresponds to $1 - \text{exclusion error}$ (pie chart), and 2) the costs for that segment, stacked with the rest, and compared to the institution’s budget (stacked bar).

**b) Group segmentation.** The second main function the platform supports, is the segmentation of the population in subgroups. The user can click the “split node" button on any leaf node in the population tree, which pops-up a menu asking her to select the attribute to use as basis of segmentation (e.g., family type, or vulnerability, Figure 4-11). The platform segments the node, initially assigning an equal threshold
(a) Initial state: total population without segmentation.

(b) Segmentation by urban/rural.

Figure 4-9: Interactive decision-support platform.
to all subgroups. Then the user can proceed to adjust thresholds for each individual group, according to the institution’s policy goals and ethical frameworks. At all times, the platform interactively computes the effects of the threshold shifts, in terms of the % of coverage for each segment, and the amount/costs compared to the institution’s budget.

Hence, the platform forces users to reflect on the fundamental tradeoffs between coverage and cost, and among population subgroups; yet providing flexibility for defining relevant segments, and priorities among them. For example, in that institution’s country the law may require to give strong preference to households in any of the following groups: indigenous households, with disabilities, displaced by armed conflict, or victims of domestic violence. In that case, the user may explore shifting thresholds to reach 100% coverage of vulnerable groups, as well to reduce coverage of the non-vulnerable group as unfortunately necessary to keep the program in budget.

c) Balance for parity. Chiefly, the platform enables users to easily explore different parity criteria across subgroups, and reflect on their impact and tradeoffs. Clicking the “balance button” on any parent node will trigger a menu asking the user to specify one of three parity criteria to implement: coverage parity (which implies exclusion error parity), thresholds parity, or demographic parity. The platform then assists the user in finding the combination of thresholds that achieves the specified parity (while maintaining constant the cost incurred by the parent node). The user can then proceed to further adjust the thresholds manually, explore different parity criteria, etc. Figures \ref{fig:threshold-parity}, \ref{fig:coverage-parity}, and \ref{fig:positive-discrimination} respectively exemplify the cases of threshold parity, coverage parity, and positive discrimination for the vulnerability attribute.

a) Dual prioritization criteria. Institutions often require to complement income poverty scores with additional indices as core prioritization criteria. Relevant use cases include: national scholarship programs requiring criteria based on both economic need and academic performance; social programs requiring criteria based on both income poverty and the amount of time (months) applicants have waited in the admissions queue; loan programs for entrepreneurship, education, or housing,
requiring criteria based on both income poverty and repayment risk; and social programs requiring to complement income poverty with additional dimensions of poverty (e.g., health or education deprivations), along the lines of multidimensional poverty [111, 30].

Hence, users of the platform can choose to specify a second prioritization criterion according to their goals, such as academic performance or a multidimensional poverty index. In this case, the thresholds adjustment window maps out the joint distribution of the population (or segment) in terms of both prioritization criteria, and enables the user to explore combinations of thresholds, and their impact on the coverage and budget (see Figure 4-10) (note that different combinations can entail same coverage and costs). Thereby, institutions can reflect on emphases over one or the other criteria, in accordance with their policy goals, and with proper awareness of its consequences.
(a) one-dimensional prioritization criteria.

(b) two-dimensional prioritization criteria.

Figure 4-10: Interactive decision-support platform.
(a) Segmentation by *family type*.

(b) Segmentation by *vulnerability*.

Figure 4-11: Interactive decision-support platform.
Figure 4-12: Interactive decision-support platform. Adjusting thresholds across subgroups towards achieving different fairness criteria. Threshold parity is instantiated by default after segmenting a node (shown in Figure 4-11(b)). Subfigures here exemplify the cases of coverage parity (a) and positive discrimination (b).
Chapter 5

Active Fairness in Algorithmic Decision-Making
5.1 Introduction

As automated decision-making systems (ADS) have become increasingly ubiquitous—e.g., in criminal justice [76], medical diagnosis and treatment [75], human resource management [26], social work [56], credit [69], and insurance [116]—there is widespread concern about how these can deepen social inequalities and systematize discrimination. Consequently, substantial work on defining and optimizing for algorithmic fairness has surged in the last few years.

Inspired by domains such as race biases in criminal risk predictions [49], a substantial body of literature has focused on the problem of balancing classification errors across protected population subgroups, towards achieving equal false positive rates, false negative rates, or both (equal odds). To that end, recent research has proposed “optimal” post-processing methods that randomize decisions of a fraction of individuals to attain group fairness [65, 106]. Yet, strong limitations of randomized approaches have been noted, such as information wastefulness, Pareto sub-optimality, and intra-group unfairness [65, 106, 31].

Our work aims at overcoming such limitations. We propose a complementary approach, active fairness, where, in deployment, an ADS adaptively collects information (features) about decision subjects; gathering more information about groups or individuals harder to classify, towards achieving equity in predictive performance. Thereby, the approach leverages a natural affordance of many real-world decision systems—adaptive information collection—and allocates an ADS’s information budget according to group- or individual-level needs.

Summary of contributions. We propose two methods for achieving fairness, based on group-level and individual-level budgets. We show that, without resorting to randomization, these methods are able to achieve: a) calibration and a single-error parity constraint, and b) parity in both false positive and false negative rates (i.e., equal odds). We show in four real-world datasets that, with constrained information budgets, active approaches can substantially outperform randomized approaches previously considered optimal (lower false positive and false negative rates). Finally, we
show that classifiers using individual-level budgets in combination with active inquiry tend to dominate classifiers that use group-level budget constraints.

**Intuition and motivating contexts.** Consider a patient entering a hospital seeking diagnosis, typically undergoing a progressive inquiry—measuring vitals, procuring lab tests, specialists’ opinions, etc. At each step, absent sufficient certainty, the inquiry continues. Intuitively, a fair health system allocates resources to provide all patients similar-quality diagnoses. Likewise, active inquiry under cost constraints underlies contexts like disaster response, poverty mapping, homeland security, recruitment, telemedicine, refugee status determination, credit and insurance pricing, etc.

**Problem formulation.** Let $X$ be an $n \times d$ feature matrix. Let $X^{(q)} \subset X$ denote a query on a subset of features in $X$, with $q \subset \{0, ..., d\}$, and $x_i^{(q)}$ the partial feature vector of individual $i$; and let $f(X^{(q)})$ be a predictor of class probability $P\{Y = 1|X^{(q)}\}$. We study the classification context where a decision-maker can choose what information to collect about each decision subject, and seeks to maximize accuracy and fairness under an information budget constraint $\bar{b} = \frac{1}{n} \sum_{i} b_i < b_{max}$, where $b_i = |q_i| \in [0, d]$ is the amount of information collected for individual $i$.

Although this setting is natural to many real-world decision systems, its affordances and implications to algorithmic decision systems—at the intersection of accuracy, fairness, and cost-efficiency—have not been thoroughly studied. Here, we focus on contexts with constant costs across features. Yet we note that the active fairness framework allows generalizations to contexts with varied costs across features, as well as richer and context-specific utility functions with potential costs to decision-subjects, such as monetary, opportunity, or privacy costs.

## 5.2 Related Work

**Active feature acquisition (AFA).** Several probabilistic and non-probabilistic methods exist for sequential feature querying under budget constraints [55, 84], dating at least back to [88], and applied in domains such as medical diagnosis [59], customer targeting [72], and image classification [55]. To the best of our knowledge, this work
Figure 5-1: Achieving calibration and single error rate parity: classifiers with group-level information budgets vs. naive randomization. Rows correspond to analysis on two different datasets: the Mexican poverty and adult income datasets. Green and yellow colors correspond to error rates for two population subgroups (e.g., white and non-white individuals). Solid black lines represent the space of calibrated classifiers. Panels in the first column (A and D) show the generalized false positive and false negative rates (GFPR and GFNR) of classifiers that randomize an increasing proportion of individuals (0% to 100%), as in [106]. In line with [106], naive randomization is able to achieve calibration and any single error parity constraint. Panels in the second column (B and E) show the same analysis for classifiers with group-level budgets. These classifiers are effective in achieving parity on either false positive or false negative rates, or equal cost, while maintaining calibration; yet without resorting to naive randomization. Finally, panels in the third column (C and F) compare the efficiency of both methods, by showing the best classifiers that achieve equal opportunity and calibration, under an information budget restriction $\bar{b} = \frac{1}{n} \sum b_i < b_{\text{max}}$. Classifiers with group-level budgets Pareto-dominate randomized classifiers by a wide margin, i.e. for the same information budget, both population subgroups are better off, incurring substantially lower false positive and false negative errors.

is the first to study the implications AFA has to the algorithmic fairness literature and policy debate. Here, we use an approach based on probabilistic random forests, but more sophisticated methods can be used, for example, for dealing with domains
with very high-dimensional input like medical images [123].

**Active learning.** Similar to the general AFA setting, this paper assumes that a fixed set of training data is used and that incremental features of a test sample can be queried. This differs from the active learning setting in which the system actively queries training examples that optimize learning e.g. by balancing exploration and exploitation or maximizing the expected model change [114]. We foresee future work studying synergies in systems that attain fairness by actively choosing training samples using active learning while also applying AFA at test time.

**Notions of fairness.** Several notions of fairness and their corresponding formalizations have been proposed, most of which require that statistical properties hold across two or more population subgroups. *Demographic or statistical parity* requires that decision rates are independent from group membership [22, 142, 86], such that $P\{\hat{Y} = 1|A = 0\} = P\{\hat{Y} = 1|A = 1\}$, for the case of binary classification and a sensitive attribute $A \in \{0, 1\}$. Most recent work focuses on meritocratic notions of fairness, or *error rate matching* [65, 10], such as requiring population subgroups to have equal false positive rates (FPR), equal false negative rates (FNR), or both, i.e., $P\{\hat{Y} = 1|A = 0, Y = y\} = P\{\hat{Y} = 1|A = 1, Y = y\}$, $y \in 0, 1$. In this work we focus on the latter set of fairness notions, although generalizations to others, such as statistical parity, are possible. Refer to [144] for a survey on computational measures of fairness.

**Achieving Equal Opportunity and Equal Odds**

Hardt et al. 2016 proposes parity in FNRS and/or parity in FPRs as a measure of unfair discrimination across population subgroups [65]. Parity in both types of error is referred to as *equal odds*, and its relaxation, equality in only FPRs, is conceptualized as *equal opportunity*, as in contexts of positive classification it means that subjects within the positive class have an equal probability of being correctly classified positive, regardless of group membership.

*Equal opportunity* can be achieved simply by shifting up or down the decision threshold $t_A$—where $\hat{Y} = 1_{[t_A,1]}(\hat{P}(Y = 1|X,A))$—for group $A$ or $A$. Yet, doing
so also directly affects FNRs, impeding achievement of equal odds. In this context, Hardt et al. 2016 propose a classifier that balances both FPRs and FNRs, based on naive randomization of a fraction of individuals in the advantaged group; and prove conditions under which the classifier is optimal with respect to accuracy [65].

Although effective in achieving equal odds, these randomization-based results have been considered discouraging for reasons outlined below, and, as shown in Section 5.6, are outperformed by active approaches.

**Achieving Calibration and Error Parity**

In many real-world uses of algorithms for risk estimation, it is common practice to require that predictions are calibrated—e.g., in recidivism [49, 32], child maltreatment hotlines [56, 28], and credit risk assessments [69]. A calibrated estimator is one where, if we look at the subset of people who receive any given probability estimate $p \in [0, 1]$, we find indeed a $p$ fraction of them to be positive instances of the classification problem. In the context of credit assignment, for example, we would expect a $p$ fraction of credit applicants with estimated default risk of $p$ to default. Moreover, in the context of algorithmic fairness across population groups, it is desired that calibration holds for each group [49].

Calibration is not necessary nor sufficient to achieve parity in classification errors [31]. However, it is particularly desirable in cases where the output of an algorithm is not directly a decision but used as input to the subsequent judgment of a human decision-maker. In such contexts, risk estimates of an uncalibrated algorithm would carry a different meaning for different groups (e.g., African-American and white defendants), and hence their use in informing human judges’ decisions would likely entail disparate impact.

Recently, Kleinberg et al. 2016 demonstrated that a tension exists between minimizing error disparity across different population groups and maintaining calibrated probability estimates [76]. In particular, it showed that calibration is compatible only with a single error constraint (i.e. equal FNR or equal FPR). On the same vein, Pleiss et al. 2017 showed that the results hold for even a strong relaxation of equal odds,
named *equal cost*, where FPRs and FNRs are allowed to compensate one another according to a cost function \[106\]. Finally, they propose a method that, using naive randomization, is able to achieve parity on either error rate or *equal cost*. We compare our methods to these benchmarks in Section 5.5.

**Objections to naive randomization**

The above results on achieving *equal odds*, as well as on jointly achieving calibration and a single error parity measure, rely on naive randomization as means to fairness. Hence, they have been interpreted as unintuitive, discouraging, and unsettling \[65, 106, 31\]. Several objections have been put forth against the use of naive randomization to achieve classification parity. Among them:

**Inefficiency.** As pointed out by \[65, 106, 31\], it is inefficient and appears unintuitive to withhold information that is already in hand, by naively randomizing the classification of a subset of individuals.

**Individual unfairness.** Classifiers based on naive randomization, such as those in \[65, 106, 31\], entail intra-group unfairness. Individuals who are randomized are not necessarily those with higher uncertainty but simply the ones who were unlucky, hence breaking ordinality between the probability of classification error and the underlying uncertainty.

**Pareto sub-optimality and undesirability.** Consider an unconstrained and unfair classifier \(\hat{Y}_U\), which incurs higher errors on group \(A\) than group \(B\); and consider an alternative "fair" classifier \(\hat{Y}_F\), where a percentage of individuals of group \(B\) are randomized to achieve parity in errors. Considering groups \(A\) and \(B\) as the system’s stakeholders, we note that the original unfair classifier \(\hat{Y}_U\) Pareto dominates the fair alternative \(\hat{Y}_F\), i.e.: the disadvantaged group \(A\) will be indifferent, as its classification remained unchanged, while group \(B\) will strongly prefer \(\hat{Y}_U\), the original classifier before accuracy was degraded by randomization. No group would prefer \(\hat{Y}_F\).
5.3 Active Fairness

The present work explores active feature acquisition approaches for achieving fairness, where a decision-maker adaptively acquires information according to the needs of different groups or individuals, in order to balance disparities in classification performance. This section defines two such strategies, one that allocates group-level information budgets—constant for all members of a group—and one that allocates individual-level information budgets, which are computed dynamically at test time. Sections 5.5 and 5.6 demonstrate their use and advantages in attaining fairness.

Preliminaries. We denote data of each decision subject as a pair \((x, y)\), where \(x\) is a feature vector of dimensionality \(d\), and \(y\) is an outcome of interest. Let \(S = (x^i, y^i)_{i=1}^n\) denote a labeled dataset, and \(A \subset S\) represent a population subgroup. Let \(\hat{Y}(X)\) be a binary classifier. We denote by \(FPR_A(\hat{Y})\) and \(FNR_A(\hat{Y})\) the false positive and false negative rates of \(\{(x, y) \in A\}\), and define disparity measures with respect to \(A\) in terms of the following FPR and FNR differences:

\[
D^A_{FPR} = \left| FPR_A(\hat{Y}) - FPR_{A(\hat{Y})} \right|
\]
\[
D^A_{FNR} = \left| FNR_A(\hat{Y}) - FNR_{A(\hat{Y})} \right|
\]

*Equal opportunity*—or FNR parity—with respect to \(A\) requires that \(D^A_{FNR} = 0\), while *equal odds* requires that both \(D^A_{FNR} = D^A_{FPR} = 0\) [65].

5.3.1 Group-Level Information Budgets

Let \(b_A, b_B\) be the information budgets for population sub-groups \(A, B\). We define predictor \(h^g\) with group-level information budgets \(b_A, b_B\) by:

\[
h^g(x_i) = \begin{cases} 
  f(x_i^{(qA)}) & \text{if } i \in A \\
  f(x_i^{(qB)}) & \text{if } i \in B 
\end{cases}
\]
where $q_A, q_B$ are feature sets that satisfy $|q_A| = b_A$ and $|q_B| = b_B$. Sections 5.5 and 5.6 show how decision-makers can achieve calibration and group-level equity by allocating budgets $b_A, b_B$.

### 5.3.2 Individual-Level Information Budgets

Beyond group-level budgets, an ADS may adaptively collect information of each decision subject until a confidence threshold is met, upon which a classification decision is made. Thereby, individual-level information budgets are set dynamically according to the needs of each decision subject, towards attaining equity.

In particular, Algorithm 1 specifies active inquiry at the individual level as the decision-making process that, given lower and upper probability thresholds $\alpha_l, \alpha_u \in (0, 1)$, and for each decision subject $i$, progressively expands the information set $x_i^{(q_i)}$ until either threshold is met, or the available feature set is exhausted. Together with the decision threshold, $\alpha_l$ and $\alpha_u$ control trade-offs between FPR and FNR. In line with related AFA methods [55], we apply early stopping to ensure we stop expanding the feature set if the classification confidence is no longer improving significantly. We estimate the parameter for early stopping using grid search while maximizing the AUC for a given budget.

We define predictor $h^{ind}$ with individual-level information budgets as $h^{ind}(x_i) = f(x_i^{(q_i)})$, where $q_i$ is the feature set according to active inquiry in Algorithm 1.

---

**Algorithm 1:** Active inquiry at the individual level

1. **Input:** data $X$, model $f$, probabilities $(\alpha_l, \alpha_u)$, decision threshold $t$;
2. **for** $i = 1$ **to** $i = n$ **do**
   3. **while** $f(x_i^{(q_i)}) \leq \alpha_u$ and $f(x_i^{(q_i)}) \geq \alpha_l$ and $|q_i| < d$ and **not** $e$ **do**
      4. $j' \leftarrow$ Get next best feature $j' \notin q_i$;
      5. $x_i^{(q_i)} \leftarrow x_i^{(q_i)} \cup x_{ij'}$;
      6. $e \leftarrow$ early_stopping($x_i^{(q_i)}, x_{ij'}$);
   7. **end**
8. $\hat{y}_i = 1_{[t,1]}\left(f(x_i^{(q_i)})\right)$;
9. **end**
10. **return** $(q_i)_{i=1}^n, \hat{Y}$;
5.3.3 Random Forest Implementation

Implementation of active classifiers requires two elements: (1) a model \( f \), able to estimate \( P\{Y|X^{(q)}\} \) for arbitrary feature subsets \( X^{(q)} \), with \( q \in [0,d] \), and (2) a feature selection method for choosing expanding feature sets, either at the group- or individual-level.

**Probabilistic model.** We implement distribution-based classification with incomplete data based on a probabilistic random forest and extending related methods for dealing with incomplete data in trees [108, 112]. In particular, when given an arbitrarily incomplete feature vector \( x_i^{(q)} \), the algorithm traverses all possible paths of each tree according to the following rule: if value \( x_{ij} \) for the current decision node is available in \( q \), the search follows the path according to the node's decision function; otherwise, if the value is not available (\( j \notin q \)), the search follows both paths. We then compute classification probabilities as a weighted average of the leaf purity across all leaves landed on by the search. Finally, we compute the average predicted probability across all trees. Similar methods can be derived for adapting logistic regressions to admit arbitrarily incomplete feature vectors [135, 112].

**Static feature selection.** We first consider a static feature ranking for guiding the acquisition of additional features in Algorithm 1 based on feature importance derived from the random forest inter-trees variability. Hence, under static feature selection, given feature ranking \( R \), the group-level budget classifier uses the top-\( b_A \) variables in \( R \) for classification of any \( i \in A \), and the top-\( b_B \) variables in \( R \) for any \( i \in B \). Similarly, the individual-level budget classifier collects the top-\( b_i \) features in \( R \) in order to classify each subject \( i \).

**Dynamic feature selection.** In the same vein, we consider dynamic or personalized feature selection, given its potential for increased individual-level equity and overall cost-efficiency. For it we implemented a greedy feature selection algorithm, which, for each subject \( i \), and at each feature collection iteration, searches for the feature \( j' \notin q_i \) that maximizes the difference between the current predicted probability \( \hat{P} \) and the
expected probability given that an additional feature $j'$ is queried, given by:

$$j' = \{j: j \notin q_i, j \in [0, d]\} \left| \hat{P}\{y_i = 1|x_i^{(q_i \cup j')}\} - \hat{P}\{y_i = 1|x_i^{(q_i)}\} \right|$$

### 5.4 Datasets

We study these methods and compare them to randomization-based approaches on four real-world, public datasets. All results are computed using random 80%/20% train/test splits.

**Mexican poverty.** Targeted social programs are challenged with household poverty prediction in order to determine eligibility [70]. This dataset is extracted from the Mexican household survey 2016, which contains ground-truth household poverty levels, as well as a series of visible household features on which inferences are based. The dataset comprises a sample of 70,305 households in Mexico, with 183 categorical and continuous features, related to households’ observable attributes and other socio-demographic features. Classification is binary according to the country’s official poverty line, with 36% of the households having the label poor. We study fairness across groups defined by a) young and old families, split by the mean (where 53% are young), and b) across families living in urban and rural areas (where 64% are urban).

**Adult income.** The Adult Dataset from UCI Machine Learning Repository [83] comprises 14 demographic and occupational attributes for 49,000 individuals, with the goal of classifying whether a person’s income is above $50,000 (25% are above), and using ethnicity (whites v. non-whites) as sensitive attribute (where 86% are white).

**German credit.** The German Credit dataset from UCI Machine Learning Repository consists of 1000 instances, of which 70% correspond to credit-worthy applicants and 30% correspond to applicants to whom credit should not be extended. Each applicant is described by 24 attributes. The sensitive attribute describes whether people are below or above the mean age (60% is below).

**Heart health prediction.** The Heart Dataset from the UCI Machine Learning
Repository contains 17 features from 906 adults. The target is to accurately predict whether or not an individual has a heart condition (54% has a heart condition). The sensitive attribute is whether people are below or above the mean age (46% is below).

5.5 Achieving Equal Opportunity & Calibration

This section demonstrates how an active strategy with group-level budgets can be used to achieve calibration and single error parity, resulting in a higher efficiency and without resorting to naive randomization.

We follow [106] and study predictive performance in terms of the generalized false positive (GFPR) and false negative rates (GFNR), appropriate for contexts where risk scores themselves are the outputs of the algorithm (as opposed to fully automated classification). We aim at designing classifiers that satisfy calibration and error parity. As shown by [76], the GFNR and GFPR of all calibrated classifiers for a given group $A$ fall along the straight line with slope $(1 - \mu_A)/\mu_A$, where $\mu_A = P(Y = 1|A)$ is $A$’s base-rate, and origin in the perfect classifier with $(GFPR, GFNR) = (0, 0)$.

Panels A and D in Figure 5-1 show the space of calibrated classifiers achievable by naive randomization (method in [106]), for the Mexican poverty and adult income datasets described in Section 5.4. These replicate results from [106], showing how naive randomization of individuals in the advantaged group can, by eroding prediction performance, achieve calibration as well as either parity in false positives, parity in false negatives (but not both), or an equal cost generalization.

Similarly, panels B and E in Figure 5-1 demonstrate how calibration and either of the three parity objectives can be achieved by adjusting information budgets according to the groups’ needs, without resorting to naive randomization. Moreover, the right column in Figure 5-1 shows that classifiers with group-level budgets achieve these fairness goals with much higher efficiency in terms of information cost. In particular, we set an overall information budget restriction for both types of classifiers, equal to the minimum budget required by the naive random classifier to achieve equal opportunity. It is observed in panels C and F that the classifiers with group-level
budgets Pareto-dominate random classifiers by a wide margin, on both datasets, i.e.: for the same information budget, both population subgroups are better off, being exposed to substantially lower false positive and false negative errors.

5.6 Achieving Equal Odds

![Achieving parity in false positives and false negatives (equal odds) via group-level information budgets. Results correspond to the Mexican poverty dataset. Achievable regions of classifiers for each population subgroup are plotted in blue and yellow. The outer and inner FPR-FNR curves of each achievable region correspond to classifiers using maximum and minimum information budgets. Points along the curve correspond to different values of the decision threshold. It is observed that active classifiers with group-level budgets achieve parity in both FNR and FPR (equal odds). Moreover, they provide equal odds solutions anywhere on the overlap of the achievable regions for both subgroups and thus along the entire FNR-FPR trade-off.](image)

This section shows how active methods with group- and individual-level information budgets can be used to achieve parity in false positives and false negatives.
Figure 5-2 illustrates the achievable regions in FPR-FNR space for classifiers with group-level information budgets, for two subgroups in the Mexican poverty dataset. It is observed that urban households are more predictable than rural households (achievable regions closer to the origin). The yellow and purple areas comprise the achievable regions for urban and rural groups. A substantial overlap is observed, showing a wide-ranged achievable region for equal odds.

In a similar way, we can obtain the achievable region of active classifiers with individual-level information budgets, by varying parameters \( \alpha_l < \alpha_u \in [0, 1] \) of Algorithm 1 (see Section 5.3).

We ran experiments to compare the three methods—naive randomization (as in [65]), group-level budgets, and individual-level budgets—and their performance in achieving equal odds solutions along the FNR-FPR trade-off. In particular, we introduce an information budget constraint \( \bar{b} = \frac{1}{n} \sum_{vi} b_i < b_{\text{max}} \), and compare solutions sets that satisfy it. Solutions of individual- and group-level classifiers are discrete, due to finite sample sizes and features dimensionality.

Figure 5-3: Achievable space of equal odds classifiers. Comparison of active classifiers with group-level (green) and individual-level (yellow) information budgets, and naive randomization (blue, following [65]). Classifiers are constrained by an information budget \( \bar{b} \leq b_{\text{max}} \). Figure A corresponds to Mexican poverty, B to German credit (center) and C to Heart health under different budget constraints. It is observed that active classifiers yield equal odds solutions along the FPR-FNR trade-off, without resorting to randomization. Moreover, in a budget-constraint setting, the active classifiers are substantially more information-efficient. The individual-level budget Pareto-dominates the group-level classifiers which in turn dominate randomized classifiers in budget-constrained environments.
Figure 5-3 shows results for three real-world datasets: Mexican poverty, German credit, and Heart health datasets. We left out the Adult Income dataset used in Fig. 5-1 since there exists no overlap between achievable regions for both subgroups and therefore we cannot achieve equal odds. Points in the FNR-FPR space were filtered to include only classifier designs that satisfied equal odds and an overall information budget constraint $b_{\text{max}}$.

It is observed that both group-level and individual-level strategies yield equal odds solutions, covering a wide range along the FNR-FPR trade-off curve, and without resorting to naive randomization. Moreover, it is shown that both type of active classifiers are substantially more information-efficient than the randomized classifier—Pareto dominance along most of the FNR-FPR trade-off curve—leading to lower false positive and false negative errors in budget-constrained environments. Finally, active classifiers with individual-level budgets tend to dominate classifiers with group-level budgets, due to their more efficient use of information by means of personalized inquiry.

5.7 Conclusions

We have proposed and demonstrated methods for simultaneously achieving equal opportunity and calibration, as well as for achieving equal odds. In contrast to prior work, the active framework does not rely on naive randomization to reach these fairness notions, avoiding several known disadvantages of randomized approaches. Instead, a decision-maker acquires partial information sets according to the needs of different groups or individuals, allocating resources equitably in order to achieve balance in predictive performance. By leveraging this additional degree of freedom, active approaches can outperform randomization-based classifiers previously considered optimal. Moreover, classifiers with individual-level budgets dominated their group-level counterparts. Finally, the extent to which the former can as well reduce intra-group unfairness is a relevant question left to future work.

More broadly, this work illustrates how, by jointly considering information col-
lection, inference, and decision-making processes, we can design automated decision systems that more flexibly optimize social objectives, including fairness, accuracy, efficiency, and privacy. A natural direction for future work is to consider richer utility functions relevant to real-world decision systems. We expect future studies that generalize results here presented to contexts with varying feature costs; as well as to contexts with multi-stakeholder value functions, where the opportunity, privacy, and monetary costs that inquiry and decision-making bring to decision-subjects are jointly considered as part of the adaptive inquiry process.

Lastly, a relevant path forward is to allow observations with partial feature sets both during training and test phases. The current implementation of this work necessitates access to full-feature observations at training time. More efficient training and further model refinement could be achieved under schemes that can learn from partial feature vectors, or proactively collect features at training time; allowing to incorporate a wider set of features tailored to increase prediction accuracy over different types of individuals.
Chapter 6

Privacy-Utility Tradeoffs in the Use of High-Dimensional Data for Development
6.1 Introduction

Large-scale datasets of human behavior are likely to revolutionize the way we develop cities, fight disease and crime, and respond to natural disasters. However, these consist of sensitive information, such as citizens’ geo-location, purchasing behavior, and socialization patterns. Moreover, numerous studies have shown adversarial methods that can successfully associate sensitive information in anonymized datasets to individual identities—i.e., reidentification [38, 60, 118, 39, 24, 19, 98, 46, 137]. Hence, understanding and managing the risk to privacy of these datasets still preconditions their broad use and potential impact.

In this work, we consider mobile phone metadata as a paradigmatic example of what is colloquially referred to as ‘big data’. Due to its high granularity, high dimensionality, passive data generation process, and high potential value, mobile phone metadata represents most characteristic features of novel data types at the core of ‘big data’. Other such types include GPS tracks, web browsing, financial behavior, genetic data, and satellite imagery; which share in common a high potential societal value and concern for individuals’ privacy.

6.1.1 Mobile Phone Data for Development

Metadata is data about data. In the context of mobile phone usage, this represents a record that a call was made—including a time stamp and geographic location, with precision determined by the location of cell towers—but no information on the content of the call itself. Mobile phone metadata is most commonly referred to as CDRs (Call Detail Records). Table 6.1 shows dummy examples of call detail records of a couple of phone calls.

Relevant characteristics of CDRs are: 1) the caller and receiver identities are pseudonymized, i.e., their phone numbers are replaced by anonymous pseudonyms (e.g., through hashing); and 2) the geographic location of towers used for each communication provide an approximation of the user’s location.

CDRs are a particularly pervasive and relevant data source for development and
humanitarian response purposes. They are generated by standard telecommunication infrastructure, and collected by mobile phone companies on an ongoing basis. Moreover, handsets and airtime are becoming cheaper, leading to increased penetration and representativity, which by 2013 approached 89% in developing countries and 96% globally[125].

There are several ways in which the location information in mobile phone metadata is analyzed and used. For example, it is possible to build dynamic maps of population density and population mobility in real time, over areas as large as countries, and at high geographic and individual detail [43]. This information in turn has valuable applications in a wide range of development and humanitarian action domains, such as: disaster response upon earthquakes and floods [12, 124], epidemic analysis of malaria and influenza outbreaks [132, 52], socio-economic and poverty mapping in both the developed and developing worlds [44, 17, 119], transportation systems development [14], and improving national statistics [71].

6.1.2 Privacy Risk in Mobile Phone Metadata

In the recent past, privacy provided by pseudonymization coupled with institutional non-disclosure agreements (NDAs) has served as basis to allow sharing of large CDR datasets. However, research has recently shown adversarial methods that successfully associate sensitive information in the datasets to individuals’ identities, even under pseudonymization of all personal identifiers [38, 60, 118, 39, 24, 19, 98, 46, 137].

A seminal study on reidentification of CDRs analyzed mobility data from 1.5 million mobile phone subscribers in a small western country, where the location of an
individual was specified hourly and with a spatial resolution given by the geographic
distribution of the carrier’s antennae [33]. It demonstrated that outside knowledge
of just four random spatio-temporal points were enough to uniquely identify 95%
of individuals in the database. Furthermore, the study showed that data can be
coarsened in order to reduce the likelihood of re-identification. This coarsening, more
properly named *spatial and temporal generalization*, is a key technique applied to data
to preserve privacy, allowing companies, NGOs, and public organizations to balance
privacy risks with data’s potential for positive societal impact.

6.2 Methods

6.2.1 Assessing Privacy

**Concepts and Vocabulary**

Datasets contain attributes such as name, telephone, address, income, health status,
location, items purchased, and websites visited. These attributes can be classified as:
direct identifiers, quasi-identifiers, or sensitive attributes. For example, in an anony-
mous health database, where names and social security numbers (direct identifiers)
are pseudonymized, a prying third party with access to the database could attempt
to know the medical condition (sensitive attributes) of Jane by using auxiliary in-
formation of her ZIP code and age (quasi-identifiers) to single her out. We denote
the set of auxiliary information about quasi-identifiers of person $i$ by $a_i$; and refer to
the subset of individuals whose records match $a_i$ as the *equivalent class* of $i$ given $a_i$,
denoted by $E_i$. Jane is reidentified if $|E_i| = 1$.

Traditional measures to protect privacy have focused on guaranteeing that an
attacker, even with full knowledge of an individual’s quasi-identifiers, is unable to
reidentify her uniquely[79], or extract information about her[87, 82]. For example,
a privacy approach of widespread use in the last decade is *k-anonymity* [79], where
the granularity of quasi-identifiers is gradually reduced, thus increasing the size of
equivalent classes, until the requirement $\min_{\forall i} |E_i| \geq k$ is met. These approaches, however, are incapable of coping with most behavioral datasets due to their high-dimensionality\cite{102}, raising the need for appropriate ones.

**Privacy in High-Dimensional Data**

People’s online activity leaves a comprehensive data trace behind, which coupled with the advent of technologies for pervasive sensing, amount to an unprecedented instrumentation of our societal systems. Notably, data at the core of today’s “big data” is high-dimensional. Examples are human mobility data, banking and credit card data, consumer behavior, web browsing, online social networks, and genetic data.

High-dimensional datasets contain only a sparse sample of the space of possible records, which, similar to fingerprints, often entails that individual records are unique. To illustrate how sparsity can be exploited for reidentification, consider a very large database of song lyrics. The space of all possible song lyrics—permutations of a bounded number of words—is extremely large. Thus, given a sequence of only 3 or 4 words, we are likely to identify a song uniquely among thousands. In practice, research has shown high reidentifiability in varied high-dimensional datasets, from mobile phone records and credit card transactions to online movie reviews \cite{38,39,98}.

**Measures of Reidentifiability**

Measures for assessing reidentification risk in high-dimensional data must recognize sensitive attributes themselves as quasi-identifiers, and vice-versa, moving to a paradigm of partial adversary knowledge as basis of reidentification. One such measure is unicity \cite{38}. The unicity $u_p$ of database $D$ is calculated as the percentage of users in $D$ who are reidentified by using $p$ randomly selected data points from each user’s records; i.e., the percentage of users whose equivalent class satisfies $|E_i^p| = 1$. For instance, it was shown that outside knowledge of four calls was enough to reidentify 95% out of 1.5 million individuals in a CDR dataset ($u_4 = 95\%$).

Here we elaborate on previous work and propose the following two metrics of privacy in high-dimensional datasets. We aim at metrics that are meaningful and intuitive, as well as rooted in the formal framework of information theory.
**Information cost.** Similar in spirit to unicity, we define the information cost of reidentification in $D$ as the average quantity of outside information that suffices to reidentify users in $D$. Let $c_i$ denote the number of data points drawn from user $i$’s records needed to reidentify her, then information cost of $D$ is defined as $c = \frac{1}{n} \sum c_i$, where $n$ is the number of users.

**Information ratio.** In addition, we define the information ratio $r$ of $D$ as the average fraction of a user’s data that is required to reidentify her. Let $|d_i|$ with $d_i \in D$ denote the amount of $i$’s data in $D$, then the information ratio of $D$ is given by $r = \frac{1}{n} \sum \frac{c_i}{|d_i|}$. Relevantly, the information ratio summarizes not only the amount of information needed for reidentification, but also the amount of information gained by an adversary once a user is reidentified; where $1 - r$ is the average information gain. This feature of the information ratio is highly relevant, as it enables stakeholders to reflect over preferences accounting for both key elements of privacy risk: information requirement and information gain.

These measures connect with information theory through the core concept of average information content—i.e., the entropy of multivariate distributions [89]. In particular, the higher the entropy of a dataset, the higher the information content of any bit of adversary knowledge, and hence the fewer bits of information required for reidentification (lower information cost and ratio). Moreover, the measures convey a meaningful and intuitive interpretation, which may help a broader audience reflect upon and assess both the likelihood and potential harm that reidentification entails. Below we apply these measures to the case of CDRs at several spatio-temporal granularity levels.

### 6.2.2 Assessing Utility

In order to assess the usefulness of mobile phone data at the various spatial and temporal generalization levels, we collected data from a quantitative survey targeted to experts with experience in research and analysis of mobile phone data for development and humanitarian action. In particular, the survey’s population were experts
who took part in D4D-Senegal 2014, the open innovation data challenge based on anonymous records of Orange’s mobile phone users in Senegal [40]. Thirty two D4D experts—members of academia and research institutions around the globe—opted in to respond the survey. Notably, the pool represented a diversity of twenty five research institutions, across fourteen countries and five continents; and a spread of domain foci in health, transportation and urban planning, national statistics, and others (see Table 6.2).

Table 6.2: Experts data.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of experts</td>
<td>32</td>
</tr>
<tr>
<td>Number of institutions</td>
<td>25</td>
</tr>
<tr>
<td>Continents</td>
<td>North America, South America, Asia, Africa and Europe.</td>
</tr>
<tr>
<td>Countries</td>
<td>Belgium, Cameroon, Canada, Chile, China France, Germany, India, Italy, Japan, Spain, Sweden, United Kingdom, USA.</td>
</tr>
<tr>
<td>Domain focus of respondents</td>
<td>Health 20.5%, Transportation and Urban Planning 34%, National Statistics 20.5%, and Others 25%</td>
</tr>
</tbody>
</table>

The survey asked experts to consider a scenario in which they were provided with CDRs from a large metropolitan region in the developing world, including all
call communications of a large representative sample of the population in the region. Experts rated on a scale from 1 to 10 the usefulness of such data in their research domains, if provided generalized at the various spatio-temporal granularity levels shown in Figure 6-1 (screenshots in Additional File 2).

6.3 Results

6.3.1 Reidentification Results

We analyzed a mobile phone dataset $D$ comprising phone calls of 1.4M people across a large metropolitan region in the developing world over a month in 2013. From it we derived generalized datasets for each combination of spatial and temporal granularity levels $g \in \{ZIP, District, Municipality\} \times \{1h, 6h, 12h, 24h\}$.

The spatial granularity levels used were ZIP code, district, and municipality polygons, which partition the space in 56, 156, and 2130 polygons with average areas of 101, 36, and 3 km$^2$ respectively. The temporal granularity levels used were time slices with duration of 1 hour, 6 hours, 12 hours, and 24 hours. For example, under dataset $D_{ZIP6}$—generalized at ZIP code and 6 hours granularity—a call issued at 4 pm from ZIP code 02139 by one user is indistinguishable from a call issued at 7 pm by another user in the same ZIP code. We computed the information cost $c(D_g)$ and information ratio $r(D_g)$ of reidentification associated to each generalized dataset $D_g$. Figure 6-1 shows the results.

We observe that for the most granular dataset, $D_{ZIP1}$, it takes on average 2.6 bits—i.e., data points—to reidentify an individual, which represents 7% of that individual’s data ( $c(D_{ZIP1}) = 2.6$ and $r(D_{ZIP1}) = 7\%$ ). This means that a prying third party with outside knowledge of 7% of an individual’s data could reidentify her and obtain the remaining 93%. In contrast, we observe that for the least granular dataset $D_{M24}$, reidentification requires an average of 32 data points, or 51% of an individual’s data ( $c(D_{M24}) = 32$ and $r(D_{M24}) = 51\%$ ). Hence, if $D_{M24}$ is published, a prying third party requires on average outside knowledge of about 51% of an individual’s
data to reidentify her and obtain the remaining 49%.

6.3.2 Utility Results

Figure 6-2 shows the experts’ assessment of data utility for each granularity level. We observe that data usefulness decays as the data is generalized spatially and temporally, with values ranging from 9.3 to 4.0 for the most and least granular datasets ($D_{Z1}$ and $D_{M24}$).

![Utility results](image.png)

Figure 6-2: Utility results. Usefulness of mobile phone metadata per generalization profile. Grey bars denote bootstrapped 95% confidence intervals. Spatial granularity levels are $Z =$ ZIP code, $D =$ district, and $M =$ municipality.

6.3.3 Privacy-Utility Tradeoff

Figure 6-3 shows results on the privacy-utility tradeoff. Each point represents a generalized dataset $D_g$ assessed on usefulness and reidentification risk, where the optimal position corresponds to the top right corner—high usefulness and hard reidentification.

We observe a sharp tradeoff between usefulness and privacy. The most granular dataset $D_{Z1}$ is the most valuable, with usefulness score of 9.3; however, it is also the
dataset most prone to reidentification, where on average a third party with outside knowledge of only 7% of an individual’s data can reidentify the individual and gain the remaining 93% of personal information. Conversely, the least granular dataset $D_{M24}$ is the least valuable, with usefulness score of 4; yet it’s also the dataset least prone to reidentification, where on average a third party requires outside knowledge of 51% of an individual’s data to gain the remaining 49% of personal information.

Figure 6-3: **Privacy-Utility trade-off in mobile phone data.** Utility vs. reidentification risk in mobile phone data for development, across spatial and temporal granularities $\{ZIP, District, Municipality\} \times \{1h, 6h, 12h, 24h\}$. The more useful the dataset, the less auxiliary information is needed to reidentify its individuals. Conversely, while data generalization increasingly hinders reidentification, it strongly diminishes datasets’ value.

Figure 6-3 also shows that the tradeoff is not strict. Generalization levels such as $D_{24}$, $Z_{24}$, $M_{1}$, and $Z_{12}$ are Pareto-suboptimal, or dominated. For example, $D_{24}$ and $M_{24}$ have similar usefulness, however an adversary requires about 65% more outside information to reidentify an individual in $M_{24}$ than in $D_{24}$ ($r(M_{24}) = 51\%$ and
\[ r(D_{24}) = 29\% \).

The tradeoff in Figure 6-2 implies that, while generalization increasingly hinders reidentification, it strongly undermines data utility. This highlights the complimentary roles of coarsening and data sharing models in enabling use while controlling risks. For example, datasets most prone to reidentification—such as \( D1, Z6, \) and \( Z1, \) with reidentification information ratio \( r \leq 10\% \)—could be shared only under strict models, such as precomputed indicators, or the use of open algorithm platforms [64,2]. Figure 6-2 also implies that even highly coarse datasets can be vulnerable to reidentification, and hence should not be made fully public. However, we may want to share more broadly datasets posing more moderate reidentification risks—such as \( M24, \) with reidentification information ratio \( r > 50\% \)—, through models similar to those used in D4D challenges [40], where data is accessed by a limited number of semi-trusted parties under non-disclosure agreements (NDAs). Similarly, datasets posing moderate-high reidentification risks—such as those with \( 10\% < r < 50\% \)—, could be shared under additional control mechanisms, such as remote access with adjustable disclosure controls via Q&A architectures, and/or accountability and deterring incentive schemes [129]. See [35, 37] for details and discussion on modern data sharing models and protocols.

6.4 Conclusions

The present work shows for the first time the notorious trade-off between the societal value of mobile phone data for development and humanitarian action, and the reidentification risk to which individuals in it are exposed. Because data generalization directly erodes data’s value, it cannot be regarded as a silver bullet solution for preserving privacy in high-dimensional datasets [97]. Yet, coupled with data-sharing models that provide adjustable degrees of accountability and security, it may help find the right balance between privacy and utility.

This work assessed data utility as the value provided to experts in the analysis of mobile phone data for development and humanitarian action. This approach is
particularly germane when considering purpose-specific data sharing, such as in the
case of poverty mapping, transportation planning, or assisting response efforts upon
natural disasters. We anticipate future work focusing on the trade-offs of mobile
phone data usage in alternative domains, such as marketing and credit scoring.

The formal measures of reidentification risk here proposed can provide meaning-
ful and intuitive summaries of the information requirements, and information gains,
associated with reidentification. Ultimately, we hope this work helps promote par-
ticipation of broader audiences in reflecting upon data privacy tensions, as societal
preferences are indispensable inputs for resolving where systems should sit along the
privacy-utility spectrum.
Chapter 7

Conclusions and Future Work
The present doctoral dissertation has aimed at contributing to the design of hybrid decision systems—that sensibly integrate elements of human and artificial intelligence—towards materializing their positive real-world impact in critical areas related to sustainable development goals.

Looking forward, we’ll pursue exciting paths at the intersections of the domain areas and methodologies that the research projects of this thesis have focused on. In particular:

1) At the intersection of work in chapters 3, 4, and 5, we’ll work on developing intelligent systems that comprehensively look at chronic disease prevention, integrating: First, targeting systems that identify the population most likely to suffer a chronic disease, or a complication of the chronic disease. Moreover, coupling them with H+AI screening tools such as the diabetic retinopathy diagnosis system developed in Chapter 3 (see diagram in Figure ??). Finally, exploring the use of an adaptive information collection and active fairness paradigm to jointly optimize accuracy, fairness, and cost of the H+AI system.

2) Active fairness methods from Chapter 5 applied to achieve fairness in the context of targeting social policies (Chapter 4). In particular, work towards “fair poverty mapping”, beyond the status quo systems, where a fixed set of information is collected for all households, and into truly intelligent systems that scan the entire country based on cheap satellite and street view images, and then based on probabilistic models decide where to send people on the ground for household visits. Thereby, social institutions could move towards fair, exhaustive, and cost-efficient targeting.

3) Finally, at the intersection of chapters 5 and 6, the active fairness framework should be extended to incorporate costs that the information-collection process may imply, not to the decision maker, but to decision subjects. In particular, work is needed on considering differentiated privacy costs associated to the features collected, and incorporating them into the active feature acquisition function. Hence, advancing towards information-collection, inference, and decision-making systems that are privacy-sensitive.
Figure 7-1: Integrated system for early diagnosis of chronic disease. Targeting population at risk, coupled with H+AI screening, towards massifying early diagnostics in developing countries.


[27] François Chollet et al. Keras. [https://keras.io](https://keras.io), 2015.


136


