Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών "Επιστήμη Δεδομένων και Μηχανική Μάθηση"
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Global Counterfactual Explanations: Efficient Computation and Bias Auditing

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΩΝΣΤΑΝΤΙΝΟΣ Α. ΤΣΟΠΕΛΑΣ

Επιβλέπων : Δημήτριος Φωτάκης
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2024

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών "Επιστήμη Δεδομένων και Μηχανική Μάθηση"
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Global Counterfactual Explanations: Efficient Computation and Bias Auditing

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΩΝΣΤΑΝΤΙΝΟΣ Α. ΤΣΟΠΕΛΑΣ

Επιβλέπων : Δημήτριος Φωτάκης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 31η Οκτωβρίου 2024.

Δημήτριος Φωτάκης Καθηγητής Ε.Μ.Π.	Αθανάσιος Βουλόδημος Επικ. Καθηγητής Ε.Μ.Π.	Δημήτριος Σαχαρίδης Επικ. Καθηγητής Univ. Libre de Bruxelles

Αθήνα, Οκτώβριος 2024

Κωνσταντίνος Α. Τσόπελας

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Τα αντιθετικά επεξηγηματικά σενάρια (*counterfactual explanations*) αποκτούν ζωτική σημασία για την ενίσχυση της διαφάνειας και της αμεροληψίας (fairness) των μοντέλων μηχανικής μάθησης, ιδιαίτερα σε ευαίσθητες περιοχές εφαρμογής. Η παρούσα διπλωματική εργασία εξετάζει δύο υπάρχουσες εργασίες — το GLANCE (Global Actions in a Nutshell for Counterfactual Explainability) και το FACTS (Fairness-Aware Counterfactuals for Subgroups) — οι οποίες επικεντρώνονται στη δημιουργία καθολικών αντιθετικών επεξηγηματικών σεναρίων (global counterfactual explanations) και στον έλεγχο δικαιοσύνης μέσω τέτοιων αντιθετικών εξηγήσεων, αντίστοιχα.

Το GLANCE παράγει συνοπτικές, ερμηνεύσιμες καθολικές αντιθετικές εξηγήσεις μέσω της συσταδοποίησης (clustering) στιγμιότυπων και ενεργειών, αναζητώντας αντιπροσωπευτικές ενέργειες που πετυχαίνουν καλή ισορροπία ανάμεσα στην αποτελεσματικότητα, το κόστος και την επεξηγησιμότητα. Από την άλλη μεριά, το FACTS έχει σκοπό τον έλεγχο της δικαιοσύνης της δικαιοπραξίας (fairness of recourse auditing), αξιολογώντας αν διαφορετικές προστατευόμενες υποομάδες αντιμετωπίζουν ίσο βαθμό δυσκολίας στην αναζήτηση δικαιοπραξίας. Για το σκοπό αυτό το FACTS 'βλέπει' τα στιγμιότυπα και τις ενέργειες τόσο σε ατομικό επίπεδο (μικρο-προοπτική) όσο και σε επίπεδο ομάδας (μακροπροοπτική).

Η παρούσα διπλωματική συγκρίνει τις δύο αυτές εργασίες και τους αντίστοιχους αλγορίθμους που προτείνουν, αναπαράγοντας τα πειράματα και τα αποτελέσματά τους, και αναδεικνύοντας τα πλεονεκτήματα και τους περιορισμούς τους. Συνδέει επίσης τις βασικές τους συλλογιστικές πορείες, επιδεικνύοντας πώς οι έννοιες των ομάδων στιγμιότυπων και των καθολικών αντιθετικών εξηγήσεων μπορούν να αξιοποιηθούν και για σκοπούς επεξηγησιμότητας και για σκοπούς ελέγχου της δικαιοσύνης.

Μέσω της αναπαραγωγής και της επέκτασης των πειραμάτων των GLANCE και FACTS, παρέχουμε μια λεπτομερή αξιολόγηση των δύο εργασιών, απεικονίζοντας τα πλεονεκτήματα και τις αδυναμίες τους. Επιπλέον, συνδέουμε τις δύο αυτές προσεγγίσεις για να αναδείξουμε πώς οι καθολικές αντιθετικές εξηγήσεις και οι έλεγχοι δικαιοσύνης μπορούν να αλληλοσυμπληρώνονται, προωθώντας το διπλό στόχο της επεξηγησιμότητας και της δικαιοσύνης στα συστήματα μηχανικής μάθησης.

Λέξεις κλειδιά

Μηχανική Μάθηση, Επεξηγησιμότητα, Δικαιοσύνη, Αντιθετικές Εξηγήσεις, Δικαιοπραξία, Καθολικές Αντιθετικές Εξηγήσεις

Abstract

Counterfactual explanations are vital for enhancing the transparency and fairness of machine learning models, particularly in sensitive domains. This thesis explores two frameworks — GLANCE (Global Actions in a Nutshell for Counterfactual Explainability) and FACTS (Fairness-Aware Counterfactuals for Subgroups) — which address global counterfactual generation and fairness auditing via counterfactuals, respectively.

GLANCE generates concise, interpretable global counterfactuals by clustering instances and actions to find representative actions that balance effectiveness, cost, and interpretability. FACTS, meanwhile, audits fairness of recourse by evaluating whether different protected subgroups face equal difficulty when seeking recourse, using both individual-level (micro) and group-level (macro) perspectives.

This thesis compares these frameworks through experimental replication, highlighting their strengths and limitations. It also connects their narratives, demonstrating how the concepts of instance groups and global counterfactuals can be leveraged for both explainability and fairness auditing purposes.

By replicating and extending the experiments from GLANCE and FACTS, we provide a detailed evaluation of these frameworks, illustrating their respective strengths and weaknesses. Additionally, we connect these two narratives to showcase how global counterfactuals and fairness audits can complement each other, thereby advancing the dual goals of interpretability and fairness in machine learning systems.

Key words

Machine Learning, Explainability, Fairness, Counterfactual Explanations, Recourse, Global Counterfactual Explanations

Ευχαριστίες

Η ολοκλήρωση της διπλωματικής μου εργασίας σηματοδοτεί την λήξη των σπουδών μου στο μεταπτυχιακό πρόγραμμα σπουδών Επιστήμη Δεδομένων και Μηχανική Μάθηση της σχολής Ηλεκτρολόγων Μηχανικών του ΕΜΠ, και κατ' επέκταση τον οριστικό αποχωρισμό μου από τη σχολή αυτή μετά από δέκα συναπτά έτη. Για την επιτυχή ολοκλήρωση αυτής της πορείας οφείλω σίγουρα πολλά ευχαριστώ.

Πρώτα από όλα, στον καθηγητή μου κ Φωτάκη, ο οποίος με καθοδήγησε και συνεργαστήκαμε στενά τόσο στις προπτυχιακές μου σπουδές και στην εκεί διπλωματική μου εργασία όσο και στην παρούσα, και στον οποίο οφείλω, επιπλέον, και πολλά από τα ενδιαφέροντά μου στον τομέα της επιστήμης των υπολογιστών.

Στην παρούσα διπλωματική εργασία, ανεκτίμητη ήταν επίσης η βοήθεια και η καθοδήγηση της Ελένης Ψαρουδάκη, σε όλη τη διάρκειά της. Ταυτόχρονα δεν θα μπορούσα να μην αναφέρω και την εξαιρετική συνεργασία πρωτίστως με τον Λουκά Κάβουρα, αλλά και με όλη την ομάδα στο ΕΚ Αθηνά, προϊόν της οποίας ήταν και οι δύο εργασίες στις οποίες βασίστηκε η παρούσα διπλωματική.

Ακόμη, σημαντικότατη ήταν σε πολλά επίπεδα και η στήριξη όλου μου του φιλικού περίγυρου. Τόσο παλαιών φίλων τους οποίους γνωρίζω και αγαπώ εδώ και χρόνια, όσο και αναπάντεχων νέων τους οποίους γνώρισα, και στο πλαίσιο του μεταπτυχιακού και εκτός αυτού.

Τέλος, Θα ήθελα να επισημάνω, όσο προφανές και αν φανεί, την απαράμιλλη και ανεκτίμητη στήριξη της οικογένειάς μου σε όλες μου τις προσπάθειες. Χωρίς αυτήν, σίγουρα δεν θα είχα καταφέρει να ανελιχθώ στο σημείο που βρίσκομαι σήμερα.

Κωνσταντίνος Α. Τσόπελας,

Αθήνα, 31η Οκτωβρίου 2024

Περιεχόμενα

Εκτεταμένη Ελληνική Περίληψη

Στο κεφάλαιο αυτό, θα εκθέσουμε περιληπτικά τα δύο κεντρικά προβλήματα με τα οποία ασχολούμαστε, και στη συνέχεια τις εργασίες GLANCE (Kavouras et al., 2024) και FACTS (Kavouras et al., 2023) που μελετούν αυτά τα δύο προβλήματα και προτείνουν αντίστοιχες λύσεις, και τα κύρια αποτελέσματα των πειραμάτων τα οποία τρέξαμε για τους σκοπούς της παρούσας διπλωματικής.

Εισαγωγή

Την τελευταία δεκαετία περίπου, τα συστήματα μηχανικής μάθησης (ΜΜ) έχουν καταστεί αναπόσπαστο κομμάτι των διαδικασιών λήψης αποφάσεων σε κρίσιμους τομείς, όπως τα χρηματοοικονομικά (Boer et al., 2019), η υγεία (Kyrimi et al., 2020), η εκπαίδευση (Loftus et al., 2018) και η απασχόληση (Bogen and Rieke, 2018; Cohen et al., 2019). Τα συστήματα αυτά είναι ολοένα και περισσότερο υπεύθυνα για αποφάσεις που μπορούν να έχουν σημαντικό αντίκτυπο στη ζωή των ανθρώπων, όπως εγκρίσεις δανείων, προσλήψεις ή ιατρικές διαγνώσεις. Ωστόσο, καθώς αυξάνεται η εφαρμογή των μοντέλων ΜΜ σε αυτούς τους ευαίσθητους τομείς, αυξάνονται και οι ανησυχίες για τη *δικαιοσύνη* και τη *διαφάνεια* των συστημάτων αυτών.

Δικαιοσύνη Τα μοντέλα μηχανικής μάθησης είναι εκ φύσεως ευάλωτα σε *μεροληψία*, καθώς συχνά βασίζονται σε ιστορικά δεδομένα που ενδέχεται να αντικατοπτρίζουν υπάρχουσες κοινωνικές προκαταλήψεις. Εάν δεν ελεγχθούν, αυτές οι μεροληψίες μπορεί να οδηγήσουν σε ανεπιθύμητα αποτελέσματα, όπου στις αποφάσεις του εκάστοτε μοντέλου μειονεκτούν δυσανάλογα άτομα ή ομάδες με βάση ευαίσθητα χαρακτηριστικά, όπως η φυλή, το φύλο ή η ηλικία (Tashea, 2017; Raji and Buolamwini, 2019) και δυνητικά να ζημιώσουν την κοινωνία μας (Osoba and Welser IV, 2017). Αυτό καθιστά τη μελέτη της δικαιοσύνης στη μηχανική μάθηση κρίσιμο πεδίο έρευνας, με σκοπό την εξασφάλιση ότι τα αποτελέσματα των μοντέλων δεν επιζημιώνουν συστηματικά *προστατευόμενες υποομάδες*.

Η μεροληψία έναντι προστατευόμενων υποομάδων στην πιο 'κλασσική' βιβλιογραφία (Mehrabi et al., 2022) εντοπίζεται κατά βάση μέσω διαφόρων εννοιών *δικαιοσύνης στην πρόβλεψη*, π.χ. της έννοιας της στατιστικής ισοτιμίας (statistical parity), όπου όλες οι υποομάδες που ορίζονται από ένα προστατευόμενο χαρακτηριστικό πρέπει να έχουν την ίδια πιθανότητα να τους αποδοθεί η θετική (ευνοϊκή) κλάση πρόβλεψης. Οι ορισμοί αυτοί ποσοτικοποιούν την *άμεση* μεροληψία που αντικατοπτρίζεται στις προβλέψεις του μοντέλου. Ωστόσο, μια *έμμεση* μορφή μεροληψίας είναι η δυσκολία ή η *επιβάρυνση* (burden) (Sharma et al., 2020; Kuratomi et al., 2023) που αντιμετωπίζει ένα άτομο (ή μια ομάδα) για να επιτύχει *δικαιοπραξία (recourse)*, δηλαδή να προβεί στις απαραίτητες *ενέργειες* για να αλλάξει τα χαρακτηριστικά του με τέτοιο τρόπο, ώστε να αποδίδεται η ευνοϊκή απόφαση (Gupta et al., 2019; Ustun et al., 2019; Miller, 2019). Η δικαιοπραξία παρέχει εξηγήσεις - μέσω αντιθετικών επεξηγηματικών σεναρίων (counterfactual explanations) (Wachter et al., 2017) -

και τη δυνατότητα δράσης σε ένα επηρεαζόμενο άτομο (affected individual) και αποτελεί νομική αναγκαιότητα σε διάφορους τομείς. Για παράδειγμα, η νομοθετική πράξη "Equal Credit Opportunity Act" απαιτεί να μπορεί ένα άτομο να μάθει τους λόγους απόρριψης μίας αίτησης για δάνειο. Η *δικαιοσύνη της δικαιοπραξίας (fairness of recourse)* αντικατοπτρίζει την έννοια ότι οι προστατευόμενες υποομάδες πρέπει να αντιμετωπίζουν την ίδια επιβάρυνση στην προσπάθεια για δικαιοπραξία (Gupta et al., 2019; Karimi et al., 2022; von Kügelgen et al., 2022; Kuratomi et al., 2023).

Διαφάνεια - Επεξηγησιμότητα Οι εφαρμογές ΜΜ είναι συχνά συστήματα *'μαύρου κουτιού'*, που παρέχουν ελάχιστη γνώση για τις εσωτερικές τους διαδικασίες λήψης αποφάσεων. Αυτή η έλλειψη διαφάνειας δημιουργεί προκλήσεις για τους χρήστες και τα ενδιαφερόμενα μέρη που χρειάζονται κατανόηση και εμπιστοσύνη στις αποφάσεις αυτές. Η αυξανόμενη ζήτηση για *επεξηγησιμότητα (explainability)* (Burkart and Huber, 2021) έχει οδηγήσει στην ανάπτυξη του τομέα της *επεξηγήσιμης ΤΝ (XAI)*, που στοχεύει στην παροχή σαφών και ερμηνεύσιμων εξηγήσεων για τον τρόπο με τον οποίο τα μοντέλα ΜΜ καταλήγουν στις αποφάσεις τους.

Ειδικότερα, σε πολλές εφαρμογές με υψηλό αντίκτυπο (π.χ. εγκρίσεις δανείων), γίνεται ολοένα και πιο αναγκαίο για τους χρήστες να κατανοούν πώς τα χαρακτηριστικά τους επηρεάζουν τα αποτελέσματα των αυτόματων συστημάτων απόφασης και πώς μπορούν να τα τροποποιήσουν για να επιτύχουν ευνοϊκά αποτελέσματα. Αυτή είναι η ίδια έννοια που συζητήσαμε προηγουμένως, γνωστή ως *δικαιοπραξία* (Miller, 2019). Τα αντιθετικά επεξηγηματικά σενάρια έχουν συγκεντρώσει μεγάλη προσοχή λόγω της καταλληλότητάς τους στην επίτευξη δικαιοπραξίας (Karimi et al., 2020), την ερμηνευσιμότητά τους (Wachter et al., 2017), τη δυνατότητα δράσης που προσφέρουν (Ustun et al., 2019), τη χρησιμότητά τους σε ελέγχους δικαιοσύνης (fairness audits) (Sharma et al., 2020; Kavouras et al., 2023) κ.λπ. Μια αντιθετική ενέργεια (counterfactual action), ή απλά μια *ενέργεια*, ορίζει τις συγκεκριμένες αλλαγές χαρακτηριστικών που χρειάζονται για τη μετατροπή μιας μη ευνοϊκής απόφασης σε ευνοϊκή.

Οι προκλήσεις της δικαιοσύνης και της επεξηγησιμότητας είναι βαθιά διασυνδεδεμένες. Σε πολλά πραγματικά πλαίσια, η διασφάλιση της δικαιοσύνης απαιτεί επίσης να παρέχεται στα επηρεαζόμενα άτομα η δυνατότητα δικαιοπραξίας — η δυνατότητα να κατανοήσουν γιατί ελήφθη μια απόφαση και ποιες συγκεκριμένες ενέργειες μπορούν να εκτελέσουν για να αλλάξουν ένα μη ευνοϊκό αποτέλεσμα. Αυτό όπως προείπαμε δεν αποτελεί μόνο επιθυμητό χαρακτηριστικό αλλά και νομική υποχρέωση σε διάφορους τομείς.

Αντικείμενα Μελέτης

Αυτή η διπλωματική εξετάζει τη διπλή πρόκληση της δικαιοσύνης και της επεξηγησιμότητας μέσω μιας λεπτομερούς μελέτης δύο βασικών εργασιών: του *FACTS* (Kavouras et al., 2023) και του *GLANCE* (Kavouras et al., 2024). Κάθε εργασία αντιμετωπίζει μια από αυτές τις βασικές προκλήσεις, όμως προσφέρουν, κατά μία έννοια, και συμπληρωματικές προοπτικές στη χρήση *αντιθετικών επεξηγηματικών σεναρίων* σε *ομάδες στιγμιοτύπων* (instances) για την αξιολόγηση και τη βελτίωση των συστημάτων μηχανικής μάθησης.

- Το *GLANCE* (Global Actions in a Nutshell for Counterfactual Explainability) αντιμετωπίζει την πρόκληση της *επεξηγησιμότητας* δημιουργώντας *καθολικά αντιθετικά επεξηγηματικά σενάρια* (global counterfactual explanations). Διατυπώνει, αρχικά, το

πρόβλημα της εύρεσης επεξηγήσιμων, εφαρμόσιμων και αποτελεσματικών καθολικών αντιθετικών εξηγήσεων που λειτουργούν επιτυχώς σε μεγάλο μέρος του χώρου εισόδου. Από εκεί και πέρα, για την δική του προτεινόμενη προσέγγιση, το GLANCE βασίζεται στην ιδέα ότι *παρόμοια στιγμιότυπα μπορούν να εξηγηθούν με παρόμοιες ενέργειες* και χρησιμοποιεί τεχνικές συσταδοποίησης (clustering) για την παραγωγή συνοπτικών συνόλων αντιθετικών ενεργειών. Διαμερίζοντας τον χώρο εισόδου και δημιουργώντας αντιπροσωπευτικές ενέργειες για κάθε τμήμα, το GLANCE ενισχύει τόσο την ερμηνευσιμότητα όσο και την πρακτικότητα των παραγόμενων εξηγήσεων.

- Το *FACTS* (Fairness-Aware Counterfactuals for Subgroups) επαναλαμβάνει και ξανατονίζει, όπως και προηγούμενες εργασίες, ότι η δικαιοσύνη της δικαιοπραξίας αποτελεί σημαντική και ξεχωριστή έννοια της αλγοριθμικής δικαιοσύνης. Ωστόσο, επισημαίνει ότι αυτή η έννοια είναι πιο περίπλοκη από ό,τι έχει διερευνηθεί μέχρι στιγμής στη βιβλιογραφία. Το FACTS διακρίνει δύο εννοιολογικές οπτικές γωνίες: τη *μικρο-προοπτική (micro viewpoint)*, που εξετάζει σε ατομικό επίπεδο το κόστος για την επίτευξη δικαιοπραξίας (όπως εστιάζεται και σε προηγούμενη έρευνα δηλαδή), και τη νέα *μακρο-προοπτική (macro viewpoint)*, που αξιολογεί πώς οι ενέργειες επηρεάζουν συλλογικά ομάδες ατόμων. Το FACTS εξερευνά συστηματικά το χώρο χαρακτηριστικών για να εντοπίσει υποχώρους όπου υπάρχει μεροληψία στη δικαιοπραξία και ποσοτικοποιεί αυτή τη μεροληψία χρησιμοποιώντας *κατανομές κόστους-αποτελεσματικότητας (effectiveness-cost distributions)* και ένα σύνολο *ορισμών δικαιοσύνης (fairness definitions)*, καθένας από τους οποίους σχεδιάστηκε ώστε να αποτυπώνει μια συγκεκριμένη πτυχή της μεροληψίας δικαιοπραξίας. Ταξινομώντας, στη συνέχεια, τους υποχώρους σύμφωνα με την μεροληψία την οποία υπολόγισε (κατά φθίνουσα σειρά), εστιάζει σε υποχώρους που παρουσιάζουν μεγάλες αποκλίσεις στην κατανομή κόστους-αποτελεσματικότητας μεταξύ των προστατευόμενων υποομάδων. Τέλος, εξάγει *Συγκριτικές Αντιθετικές Επεξηγήσεις Υποομάδων (Comparative Subgroup Counterfactuals - CSCs)*, για να παρέχει ερμηνεύσιμες περιλήψεις των ευρημάτων αυτών.

Παρακάτω παρουσιάζουμε μια πιο εκτενή εισαγωγή στον τρόπο με τον οποίο καθεμία από τις δύο εργασίες προσεγγίζει το αντίστοιχο πρόβλημα και θα προσπαθήσουμε να δικαιολογήσουμε διαισθητικά την προσέγγιση αυτή.

Σύντομη εισαγωγή στο GLANCE και στον Αλγόριθμο Augmented Space

Τα αντιθετικά επεξηγηματικά σενάρια παραδοσιακά εστιάζουν στην παροχή *τοπικής επεξηγησιμότητας*, δηλαδή προσαρμόζουν τις εξηγήσεις σε συγκεκριμένα *επηρεαζόμενα στιγμιότυπα* (affected instances ή individuals, αυτά δηλαδή τα άτομα που 'απορρίπτονται' από το εκάστοτε μοντέλο). Ωστόσο, σε πολλές πρακτικές εφαρμογές, τα *καθολικά αντιθετικά επεξηγηματικά σενάρια* είναι πιο πολύτιμα, καθώς προσφέρουν ευρύτερες εξηγήσεις που εφαρμόζονται σε ολόκληρο τον επηρεαζόμενο πληθυσμό. Φυσικά, είναι δυνατόν να καλυφθούν όλα τα άτομα δημιουργώντας πολλαπλά τοπικά αντιθετικά σενάρια (ξεχωριστά για κάθε άτομο), ωστόσο αυτή η προσέγγιση μειώνει σημαντικά την ερμηνευσιμότητα, που αποτελεί βασικό στοιχείο της καθολικής επεξηγησιμότητας (global explainability).

Ορμώμενο από προηγούμενες εργασίες, όπως οι Rawal and Lakkaraju (2020); Kanamori et al. (2022), το Kavouras et al. (2024) ορίζει τα Καθολικά Αντιθετικά Επεξηγηματικά Σενά-

ρια (Global Counterfactual Explanations - GCE) ως ένα μικρό σύνολο *καθολικών ενεργειών* σχεδιασμένων για την αποτελεσματική παροχή δικαιοπραξίας στον επηρεαζόμενο πληθυσμό (affected population). Μια επιτυχημένη λύση καθολικών αντιθετικών ενεργειών θα έπρεπε να ικανοποιεί τρεις βασικούς στόχους, στους οποίους δίνει έμφαση και μεγάλο μέρος της προηγούμενης βιβλιογραφίας:

1. να αποτελείται από μικρό αριθμό ενεργειών για τη διατήρηση της ερμηνευσιμότητας (size - μέγεθος),

2. να ελαχιστοποιεί το κόστος εφαρμογής των ενεργειών αυτών (cost - κόστος),

3. να παρέχει δικαιοπραξία σε όσο το δυνατόν περισσότερα επηρεαζόμενα άτομα (effectiveness - αποτελεσματικότητα).

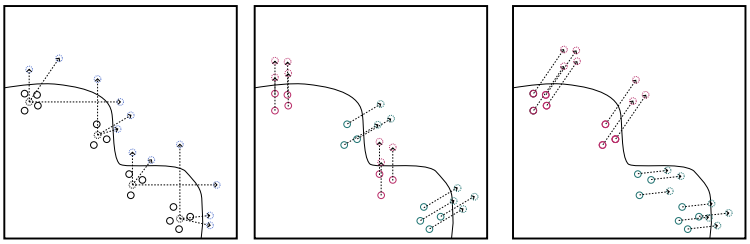
Παρατηρείστε ότι οι τρεις αυτοί στόχοι μαζί συνιστούν ένα πρόβλημα βελτιστοποίησης πολλαπλών αντικειμενικών συναρτήσεων (multi-objective optimization problem). Όπως επισημαίνεται πχ στο Branke (2008), οι σχέσεις μεταξύ πολλαπλών στόχων βελτιστοποίησης μπορεί να είναι περίπλοκες, και η αναγωγή τους σε έναν μόνο στόχο, όπως γίνεται συχνά στην πράξη (πχ Rawal and Lakkaraju, 2020), μπορεί να είναι προβληματική, καθώς αυτές οι ποσότητες κατά βάση δεν είναι συγκρίσιμες. Θέτοντας τις GCEs ως ένα πρόβλημα πολλαπλών στόχων βελτιστοποίησης, οι Kavouras et al. (2024) είναι σε θέση να εξερευνήσουν τους εγγενείς *συμβιβασμούς (trade-offs)* μεταξύ αποτελεσματικότητας και κόστους, ιδιαίτερα από τη στιγμή που η ερμηνευσιμότητα επιβάλλει περιορισμούς στο πλήθος των ενεργειών που μπορεί να περιέχει η τελική λύση - σύνολο.

Η εύρεση καθολικών αντιθετικών εξηγήσεων παρουσιάζει μια μοναδική πρόκληση βελτιστοποίησης σε σύγκριση με την εύρεση τοπικών εξηγήσεων. Από μία άλλη οπτική γωνία, παρατηρείστε ότι, ακόμη και αν υποθέσουμε ότι για κάθε άτομο βρίσκεται (με κάποιον τρόπο) η βέλτιστη τοπική ενέργεια, και ως καθολικές ενέργειες επιλεγούν ένα υποσύνολο όλων των τοπικών ενεργειών, οι GCEs μπορεί να παραμένουν μη βέλτιστες. Για παράδειγμα, η βέλτιστη καθολική ενέργεια μπορεί να μην είναι η καλύτερη τοπική ενέργεια κανενός μεμονωμένου στιγμιότυπου, αλλά να αποτελεί συμβιβασμό μεταξύ όλων των βέλτιστων τοπικών ενεργειών. Συνεπώς, η παραγωγή αποτελεσματικών GCEs απαιτεί μια πιο εκλεπτυσμένη εξερεύνηση του χώρου ενεργειών.

Η πρόταση του GLANCE Το GLANCE (Kavouras et al., 2024; Emiris et al., 2024) εισάγει διάφορες προσεγγίσεις για την αντιμετώπιση αυτού του περίπλοκου προβλήματος, οι οποίες μοιράζονται ένα κοινό χαρακτηριστικό: βασίζονται στην *ομαδοποίηση ατόμων*, χρησιμοποιώντας μια διαδικασία που με κάποιο τρόπο λαμβάνει υπόψη τόσο τον χώρο χαρακτηριστικών όσο και τον *χώρο ενεργειών*.

Στην παρούσα εργασία, εστιάζουμε σε μία συγκεκριμένη προσέγγιση εξ αυτών του GLANCE — τον αλγόριθμο *Augmented Space* συγκεκριμένα — για δύο βασικούς λόγους. Πρώτον, αποτέλεσε το κύριο αντικείμενο των προσπαθειών και της συνεισφοράς του γράφοντος κατά την ανάπτυξη του GLANCE. Δεύτερον, αυτός ο συγκεκριμένος αλγόριθμος δεν εξετάστηκε περαιτέρω στη δεύτερη εργασία για το GLANCE (Kavouras et al., 2024), κι έτσι έχει ένα ενδιαφέρον για περαιτέρω διερεύνηση.

Όπως και άλλες προσεγγίσεις εντός του GLANCE, ο αλγόριθμος Augmented Space (AS) ξεκινά με τη διενέργεια συσταδοποίησης των ατόμων στον χώρο χαρακτηριστικών, η οποία είναι αρκετά 'λεπτομερής', δηλαδή δημιουργεί πολλές και μικρές συστάδες (clusters). Για κάθε συστάδα, υπολογίζει ένα σύνολο ποικίλων ενεργειών που μπορούν να προσφέρουν δικαιοπραξία στο κεντροειδές (centroid) της συστάδας. Η βασική υπόθεση εδώ είναι



(a) (b) (c)

Σχήμα G.1: Διαίσθηση πίσω από τις μεθόδους συσταδοποίησης (όπως παρουσιάζεται στο Kavouras et al. (2024)). (α) Το GLANCE ξεκινά δημιουργώντας ποικίλες υποψήφιες ενέργειες από τα κεντρικά σημεία των συστάδων που σχηματίζονται στον χώρο χαρακτηριστικών. (β) Οι τελικές συστάδες σχηματίζονται βάσει της ομοιότητας τόσο στον χώρο χαρακτηριστικών όσο και στις ενέργειες, ομαδοποιώντας περιπτώσεις που μπορεί να είναι απομακρυσμένες στον χώρο χαρακτηριστικών αλλά να μοιράζονται παρόμοιες ενέργειες. (γ) Αντίθετα, η συσταδοποίηση αποκλειστικά βάσει ομοιότητας χαρακτηριστικών μπορεί να οδηγήσει σε μη βέλτιστες ενέργειες, είτε επιτυγχάνοντας υψηλή αποτελεσματικότητα με υψηλό κόστος είτε χαμηλό κόστος με μικρή αποτελεσματικότητα.

ότι *άτομα με υψηλή ομοιότητα* είναι πιθανό να μοιράζονται παρόμοια 'οπτική γωνία' του ορίου απόφασης του μοντέλου και, συνεπώς, μπορούν να επιτύχουν δικαιοπραξία μέσω *παρόμοιων, καλού κόστους ενεργειών*. Το Σχήμα G.1a παρουσιάζει ένα παράδειγμα όπου τέσσερις μικρές συστάδες συνδέονται με τρεις διαφορετικές ενέργειες η καθεμία, που προκύπτουν από τα κεντρικά σημεία τους. Αυτές οι ενέργειες προσφέρουν διαφορετικές διαδρομές για την επίτευξη δικαιοπραξίας στα μέλη της συστάδας, δίνοντας στο GLANCE την ευελιξία να επιλέξει την καταλληλότερη ενέργεια αργότερα.

Στο τέλος αυτής της αρχικής φάσης, λοιπόν, το GLANCE συνδέει κάθε ομάδα ατόμων με ένα σύνολο πιθανών ενεργειών. Σε αυτό το σημείο, οι διάφορες προσεγγίσεις του GLANCE αποκλίνουν, και από εδώ και στο εξής, θα επικεντρωθούμε αποκλειστικά στον αλγόριθμο Augmented Space. Στη δεύτερη φάση, ο αλγόριθμος αντιστοιχεί σε κάθε στιγμιότυπο την καλύτερη δυνατή ενέργεια[1] από το σύνολο ενεργειών της αντίστοιχης ομάδας. Αυτό ουσιαστικά απεικονίζει τα στιγμιότυπα σε έναν χώρο υψηλότερων διαστάσεων — τον *επαυξημένο χώρο (augmented space)* — που κωδικοποιεί τόσο τα χαρακτηριστικά όσο και τις σχετικές ενέργειες των στιγμιοτύπων.

Στη συνέχεια, ο αλγόριθμος διενεργεί συσταδοποίηση σε αυτά τα επαυξημένα στιγμιότυπα, σχηματίζοντας t νέες συστάδες στον επαυξημένο χώρο. Στο τέλος, από κάθε συστάδα επιλέγεται μία τελική ενέργεια με βάση μια απλή, άπληστη (greedy) διαδικασία. Η κύρια καινοτομία, ωστόσο, του αλγορίθμου Augmented Space έγκειται σε αυτή την συσταδοποίηση εντός του επαυξημένου χώρου χαρακτηριστικών και ενεργειών. Το σκεπτικό για τη συμπερίληψη της ομοιότητας χαρακτηριστικών είναι ότι ομαδοποιεί παρόμοια άτομα, με την ίδια λογική που εκθέσαμε και για την αρχική φάση. Το κίνητρο για τη συμπερίληψη και των ενεργειών, όμως, είναι ότι αν παρόμοιες ενέργειες προσφέρουν δικαιοπραξία, είναι ευκολότερο να εντοπιστεί μια κοινή, χαμηλού κόστους και υψηλής αποτελεσματικότητας ενέργεια, ακόμη και αν τα άτομα διαφέρουν κάπως ως προς τα χαρακτηριστικά τους. Το Σχήμα G.1b δείχνει ένα παράδειγμα όπου τα άτομα ομαδοποιούνται βάσει τόσο της ομοιό-

[1] Με απλή λογική, όπως θα ακολουθούνταν για μια τοπική αντιθετική εξήγηση. Δηλαδή, από τις ενέργειες της συστάδας στην οποία ανήκει το στιγμιότυπο, κρατώνται εκείνες που επιτυγχάνουν δικαιοπραξία για το συγκεκριμένο στιγμιότυπο, και από αυτές επιλέγεται η φθηνότερη (με το μικρότερο κόστος).

τητας χαρακτηριστικών όσο και της ομοιότητας ενεργειών.

Όπως επισημαίνεται και στο Kanamori et al. (2022), το να βασίζεται κανείς αποκλειστικά σε μία γενική, αδρομερή συσταδοποίηση στον χώρο χαρακτηριστικών, καθώς και η ανάθεση μιας μόνο ενέργειας σε κάθε συστάδα μπορεί να οδηγήσει σε υπο-βέλτιστες καθολικές ενέργειες, υποβιβάζοντας είτε το κόστος είτε την αποτελεσματικότητα. Για παράδειγμα, στο Σχήμα G.1c, τα άτομα συσταδοποιούνται βάσει της ομοιότητας των χαρακτηριστικών τους, και επιλέγεται μία ενέργεια για ολόκληρη την ομάδα. Στην μπλε ομάδα, οι ενέργειες παρουσιάζουν αρκετή ποικιλία, κι έτσι η επιλογή μιας μόνο ενέργειας επιβάλλει κάποιον συμβιβασμό μεταξύ αποτελεσματικότητας (όπως απεικονίζεται στο σχήμα) και κόστους. Αυτό υπογραμμίζει την αξία της ομαδοποίησης που λαμβάνει υπόψη τόσο τα χαρακτηριστικά όσο και τις ενέργειες.

Σύντομη εισαγωγή στο FACTS

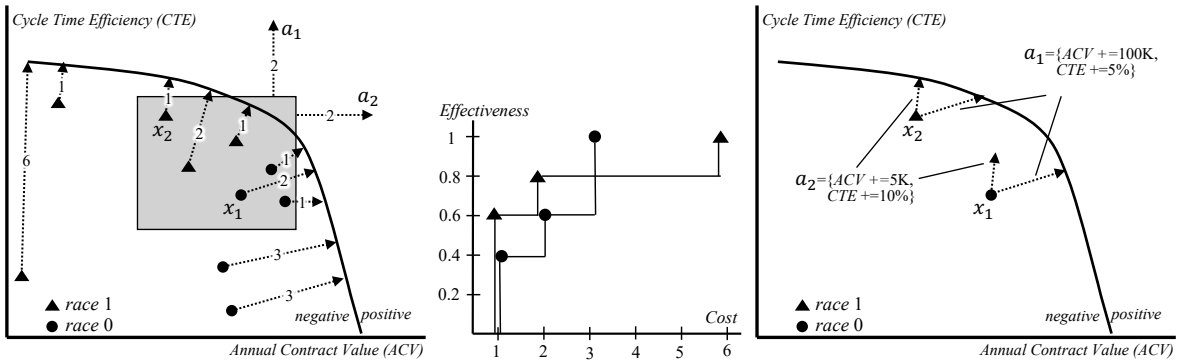
Ως εισαγωγικό παράδειγμα της δικαιοσύνης στη δικαιοπραξία (*fairness of recourse*), ας εξετάσουμε ένα υποτιθέμενο σύστημα AI μιας εταιρείας που χρησιμοποιείται για αποφάσεις προαγωγών, το οποίο βασίζεται σε δύο χαρακτηριστικά: CTE (cycle time efficiency) και ACV (annual contract value). Στο Σχήμα G.2a εμφανίζονται υπάλληλοι που ανήκουν στην αρνητική προβλεπόμενη κατηγορία, ενώ επιπλέον εμφανίζεται ένα υποτιθέμενο χαρακτηριστικό φυλής, που απεικονίζεται με κύκλους και τρίγωνα. Τα βέλη υποδεικνύουν τις συντομότερες ενέργειες για κάθε υπάλληλο για να επιτύχει δικαιοπραξία (δηλαδή, τοπικές αντιθετικές ενέργειες - local counterfactual actions), και οι αριθμοί στα βέλη αντιπροσωπεύουν το μήκος τους, το οποίο θεωρείται ως το κόστος της αντίστοιχης ενέργειας. Για παράδειγμα, το x_1 μπορεί να βελτιωθεί κυρίως μέσω της απόκτησης έργων υψηλής αξίας (ACV), και η συγκεκριμένη ενέργεια έχει κόστος ίσο με 2 (το μήκος του βέλους).

Αν εξετάσουμε το μέσο κόστος, αυτό διαφέρει ανά φυλή: 2 για τη φυλή 0 και 2.2 για τη φυλή 1, υποδεικνύοντας αδικία στη δικαιοπραξία κατά της φυλής 1 [2]. Ωστόσο, οι συγγραφείς του Kavouras et al. (2023) υποστηρίζουν ότι η δικαιοσύνη στη δικαιοπραξία είναι πιο περίπλοκη από το απλό μέσο κόστος δικαιοπραξίας που εξετάζεται στη σχετική βιβλιογραφία (π.χ. Gupta et al., 2019; Sharma et al., 2020; von Kügelgen et al., 2022; Kuratomi et al., 2023), και φέρνουν μία σειρά από επιχειρήματα.

Πρώτον, το μέσο κόστος παρέχει μια περιορισμένη εικόνα για τα ατομικά κόστη δικαιοπραξίας και μπορεί να παρερμηνεύσει τις ανισότητες. Για παράδειγμα, στο Σχήμα G.2a, η φυλή 1 έχει έναν outlier με υψηλό κόστος 6, που επηρεάζει έντονα τον μέσο όρο. Για να προσφέρουμε μια πιο καθαρή εικόνα, η συνάρτηση κατανομής των κοστών, που ονομάζεται *κατανομή κόστους-αποτελεσματικότητας* (ecd) στο Kavouras et al. (2023), δείχνει την αντιστάθμιση μεταξύ δικαιοπραξίας και κόστους (Σχήμα G.2b). Βασιζόμενοι στην ecd, οι συγγραφείς του Kavouras et al. (2023) διαμορφώνουν νέους ορισμούς/μετρικές της δικαιοσύνης στη δικαιοπραξία, οι οποίοι βασίζονται σε άλλα κριτήρια, πέρα από το μέσο κόστος. Για παράδειγμα, αν σε μια συγκεκριμένη εφαρμογή οι ενέργειες με κόστος πάνω από 2 θεωρούνται ανέφικτες, μόνο το 60% της φυλής 0 μπορεί να επιτύχει δικαιοπραξία, σε σύγκριση με το 80% της φυλής 1, αντιστρέφοντας την μεροληψία που προέκυψε από τον μέσο όρο. Αυτός ο ορισμός αναφέρεται ως *equal effectiveness within budget* στο Kavouras et al. (2023).

Επιπλέον, η παροχή δικαιοπραξίας ατομικά στο κάθε στιγμιότυπο πριν από τον υπολογισμό του μέσου όρου (όπως στο Σχήμα G.2a) μπορεί να μην είναι χρήσιμη σε ορισμένα πρακτικά σενάρια. Η *μακρο-προοπτική* που προτείνεται στο Kavouras et al. (2023) αντιμε-

[2] Σημειώστε ότι η δικαιοσύνη στην πρόβλεψη με όρους στατιστικής ισοτιμίας ικανοποιείται.



(a) (b) (c)

Σχήμα G.2: (εικόνα από Kavouras et al. (2023)) (α) Μια απεικόνιση των επηρεαζόμενων ατόμων, του ορίου απόφασης, των ενεργειών και μιας υποομάδας (που επισημαίνεται με τη σκιασμένη περιοχή) εντός του χώρου χαρακτηριστικών (β) Η σωρευτική συνάρτηση κατανομής των κοστών δικαιοπραξίας για τα άτομα στο (α) (γ) Σύγκριση δύο ενεργειών που επιτρέπουν δικαιοπραξία για δύο άτομα.

τωπίζει μια ομάδα ατόμων συλλογικά, αξιολογώντας ενέργειες που επηρεάζουν όλα τα μέλη, που θα μπορούσαν να αντιπροσωπεύουν κοινωνικές ή οργανωτικές παρεμβάσεις (π.χ., φορολογικές ελαφρύνσεις ή εκπαιδευτικά προγράμματα). Αυτό απεικονίζεται στο Σχήμα G.2a με τα άτομα στην σκιασμένη περιοχή, τα οποία μπορούν να δεχτούν μια οριζόντια ενέργεια όπως η a_1 ή η a_2. Η ενέργεια a_1 καταδεικνύει την αξία αυτής της προοπτικής: ενώ οι δύο φυλετικές υποομάδες θα είχαν παρόμοια κόστη στη μικρο-προοπτική, η ενέργεια a_1 ωφελεί δυσανάλογα τη φυλή 0 σε σχέση με τη φυλή 1. Αυτή η προοπτική αναδεικνύει συστημικές προκαταλήψεις που μπορεί να παραβλέπουν οι αξιολογήσεις σε ατομικό επίπεδο.

Τέλος, η παραδοσιακή δικαιοσύνη στη δικαιοπραξία εξαρτάται από μια συνάρτηση κόστους, η οποία εισάγει πρακτικούς περιορισμούς. Ο προσδιορισμός των χαρακτηριστικών που είναι εφαρμόσιμο ή ηθικά αποδεκτό να τροποποιηθούν περιπλέκει την ανάλυση (Rawal and Lakkaraju, 2020; Ustun et al., 2019). Για να αντιμετωπιστεί αυτό, το Kavouras et al. (2023) προτείνει και ορισμούς δικαιοσύνης ανεξάρτητους από το κόστος, όπως ο *equal choice for recourse*, που λαμβάνει υπόψη τον αριθμό διαθέσιμων ενεργειών αντί για το κόστος. Για παράδειγμα, το x_1 έχει μόνο μία επιλογή για δικαιοπραξία, ενώ το x_2 έχει δύο, γεγονός που υποδεικνύει αδικία, ακόμη και αν και οι δύο έχουν ίσα κόστη με την ενέργεια a_1.

Βάσει των παραπάνω παρατηρήσεων, το Kavouras et al. (2023) προτείνει επίσης το FACTS (Fairness-Aware Counterfactuals for Subgroups), έναν αλγόριθμο που σχεδιάστηκε για την αποδοτική αξιολόγηση της δικαιοσύνης στη δικαιοπραξία. Το FACTS υπολογίζει τη προαναφερθείσα κατανομή κόστους-αποτελεσματικότητας, η οποία αποτυπώνει την ισορροπία μεταξύ κόστους και δικαιοπραξίας τόσο στη μικρο-προοπτική όσο και στη μακρο-προοπτική. Στη συνέχεια, αξιολογεί αποδοτικά τη δικαιοσύνη στη δικαιοπραξία, εξερευνώντας το χώρο χαρακτηριστικών και εντοπίζοντας υποομάδες με μεροληψία στη δικαιοπραξία, βάσει ενός από τους διάφορους ορισμούς δικαιοσύνης που δίνει το Kavouras et al. (2023) (και που υπολογίζονται με τη βοήθεια της `ecd`). Το FACTS κατατάσσει μεροληπτικούς υποχώρους και παρέχει ερμηνεύσιμες περιλήψεις, προσφέροντας έναν συστηματικό τρόπο εξέτασης της δικαιοσύνης τόσο στη μικρο-προοπτική όσο και στη

μακρο-προοπτική.

Εν κατακλείδι, το Kavouras et al. (2023) έχει σημαντική εννοιολογική αλλά και τεχνική συνεισφορά. Σε εννοιολογικό επίπεδο, διακρίνει μεταξύ της *μικρο-προοπτικής*, η οποία εστιάζει στα ατομικά κόστη, και της *μακρο-προοπτικής*, που αξιολογεί τις επιπτώσεις σε επίπεδο ομάδας. Εισάγονται επίσης νέες έννοιες της δικαιοσύνης στη δικαιοπραξία, που επεκτείνονται πέρα από τους περιορισμούς του μέσου κόστους. Σε τεχνικό επίπεδο, προσφέρει τον αλγόριθμο FACTS, ο οποίος αξιολογεί αποδοτικά τη δικαιοσύνη στη δικαιοπραξία εξερευνώντας το χώρο χαρακτηριστικών και εντοπίζοντας υποομάδες με μεροληψία στη δικαιοπραξία.

Βασική Συνεισφορά της Διπλωματικής Εργασίας

Αυτή η διατριβή εξετάζει τη διπλή πρόκληση της δικαιοσύνης και της επεξηγησιμότητας μέσω μιας λεπτομερούς μελέτης δύο βασικών εργασιών: *FACTS* (Kavouras et al., 2023) και *GLANCE* (Kavouras et al., 2024). Κάθε εργασία αντιμετωπίζει μία από αυτές τις βασικές προκλήσεις, ωστόσο προσφέρουν και συμπληρωματικές προοπτικές για τη χρήση αντιθετικών επεξηγηματικών σεναρίων στην αξιολόγηση και βελτίωση συστημάτων μηχανικής μάθησης.

Παρότι κάθε πλαίσιο έχει διακριτή εστίαση — το FACTS στον έλεγχο δικαιοσύνης και το GLANCE στις καθολικές αντιθετικές εξηγήσεις — αυτή η διατριβή θα επιχειρήσει επίσης να διερευνήσει τις διασυνδέσεις μεταξύ αυτών των δύο τομέων. Αναπαράγοντας τις μεθοδολογίες των FACTS και GLANCE, διεξάγοντας επιπλέον πειράματα και εξετάζοντας τα αποτελέσματα σε διαφορετικά πλαίσια, αυτή η εργασία στοχεύει να αναδείξει πώς η δικαιοσύνη και η επεξηγησιμότητα μπορούν να αλληλοσυμπληρώνονται σε πρακτικές εφαρμογές.

Αντιθετικές Εξηγήσεις: Βασικοί Ορισμοί, Μέθοδοι και Εφαρμογές

Σε όλη τη διάρκεια της εργασίας μας, ξεκινάμε με την υπόθεση ότι μας δίνεται ένα μοντέλο ΜΜ το οποίο επιλύει ένα πρόβλημα ταξινόμησης και εκφράζεται μέσω μιας συνάρτησης $h : \mathcal{X} \to \mathcal{Y}$ που απεικονίζει στιγμιότυπα σε ετικέτες. Κάθε στιγμιότυπο έχει έναν αριθμό από χαρακτηριστικά (features), που αντιστοιχούν στις συνιστώσες - διαστάσεις του χώρου $\mathcal{X}$.

Με δεδομένο αυτό, στην παρούσα ενότητα θα δούμε το τι είναι οι αντιθετικές εξηγήσεις, κάποιους ορισμούς τους δηλαδή, τόσο σε τοπικό όσο και σε καθολικό επίπεδο. Επίσης, θα μελετήσουμε επιγραμματικά προηγούμενες εργασίες πάνω σε αυτά τα ζητήματα, και τέλος θα συζητήσουμε περιληπτικά την πολύ σημαντική εφαρμογή των αντιθετικών εξηγήσεων που είναι στον έλεγχο της δικαιοσύνης των μοντέλων ΜΜ.

Τοπικές Αντιθετικές Εξηγήσεις

Οι αντιθετικές εξηγήσεις (counterfactual explanations) είναι σημαντικές για την κατανόηση αιτιωδών σχέσεων, απαντώντας σε ερωτήσεις τύπου «τι θα συνέβαινε αν;». Στα συστήματα μηχανικής μάθησης, αυτές οι εξηγήσεις βοηθούν στο να κατανοηθεί πώς οι αλλαγές στα χαρακτηριστικά εισόδου θα μπορούσαν να μεταβάλουν την πρόβλεψη ενός μοντέλου. Για παράδειγμα, εάν ένα μοντέλο απορρίπτει μια αίτηση δανείου, μια αντιθετική

εξήγηση θα μπορούσε να υποδείξει ότι μια αύξηση στο ετήσιο εισόδημα θα άλλαζε την απόφαση προς όφελος του αιτούντος. Σημαντικό τμήμα της διαδικασίας αυτής αποτελεί η εύρεση των *ελαχίστων* απαιτούμενων αλλαγών στα χαρακτηριστικά ενός στιγμιότυπου.

Οι τοπικές αντιθετικές εξηγήσεις (local counterfactuals), λοιπόν, αναζητούν το πλησιέστερο στιγμιότυπο που θα προκαλέσει αλλαγή στην πρόβλεψη του μοντέλου. Στην απλή τους μορφή, οι τοπικές αντιθετικές εξηγήσεις επιδιώκουν να βρουν ελάχιστες αλλαγές στο αρχικό στιγμιότυπο ώστε το μοντέλο να καταλήξει σε μια διαφορετική, επιθυμητή πρόβλεψη. Το έργο των Wachter et al. (2017) έθεσε ένα θεμέλιο ορίζοντας το πρόβλημα ως ελαχιστοποίηση μιας συνάρτησης κόστους που ισορροπεί μεταξύ της επιθυμητής πρόβλεψης και της ελάχιστης απόστασης από το αρχικό στιγμιότυπο (βλ. Εξίσωση 2.2).

Προκειμένου να δημιουργηθούν καλές αντιθετικές εξηγήσεις, υπάρχουν διάφορα κριτήρια που πρέπει να ληφθούν υπόψη, όπως η μικρότερη δυνατή αλλαγή, η ρεαλιστικότητα και η δυνατότητα δράσης των προτεινόμενων αλλαγών. Μέθοδοι όπως το DiCE (Mothilal et al., 2019) προσθέτουν και την έννοια της ποικιλίας στις αντιθετικές εξηγήσεις, επιτρέποντας πολλαπλές και διαφορετικές επιλογές για την επίτευξη του στόχου. Το DiCE χρησιμοποιεί διαδικασίες τυχαίας επιλογής και κατάλληλους περιορισμούς για να παράγει ένα σύνολο αντιθετικών παραδειγμάτων που διατηρούν την ερμηνευσιμότητα και τη δυνατότητα δράσης[3].

Εκτός από το DiCE, υπάρχουν και άλλες μέθοδοι όπως το FACE (Poyiadzi et al., 2020), που διασφαλίζει ότι οι αντιθετικές εξηγήσεις είναι εφικτές και αντιπροσωπεύουν ρεαλιστικά μονοπάτια αλλαγών, το FOCUS (Lucic et al., 2021), που επιτρέπει την παραγωγή αντιθετικών εξηγήσεων για μη διαφορίσιμα μοντέλα, και το Growing Spheres (Laugel et al., 2017), το οποίο αναζητά την ελάχιστη διαταραχή του αρχικού στιγμιότυπου για να επιτύχει την επιθυμητή αλλαγή στην πρόβλεψη.

Η παραγωγή τοπικών αντιθετικών εξηγήσεων μπορεί να είναι δύσκολη και απαιτητική, ιδιαίτερα σε πλαίσια όπου εκφράζεται ως κάποιο πολυπαραμετρικό πρόβλημα βελτιστοποίησης, όπου οι στόχοι συχνά συγκρούονται. Αυτές οι προκλήσεις δείχνουν ότι η παραγωγή τοπικών αντιθετικών εξηγήσεων απαιτεί μια προσεκτική ισορροπία μεταξύ της επιθυμητής πρόβλεψης, της μικρότερης δυνατής αλλαγής και άλλων περιορισμών, όπως η δυνατότητα δράσης και η ποικιλία.

Καθολικές Αντιθετικές Εξηγήσεις

Η πιο πρόσφατη έρευνα στις αντιθετικές εξηγήσεις έχει επεκταθεί από την τοπική ανάλυση σε μια νέα κατηγορία: τις *καθολικές αντιθετικές εξηγήσεις* (global counterfactuals). Αυτές στοχεύουν στην παραγωγή εξηγήσεων που καλύπτουν όχι μόνο μεμονωμένα στιγμιότυπα αλλά ομάδες ή και ολόκληρο τον πληθυσμό, επιτρέποντας την κατανόηση της συμπεριφοράς του μοντέλου σε επίπεδο ομάδων. Σε ένα σύστημα αυτόματης έγκρισης δανείων, για παράδειγμα, μια καθολική αντιθετική εξήγηση μπορεί να αναφέρει ότι η αύξηση του ετησίου μισθού κατά 50.000 δολάρια για άτομα με λιγότερα από 10 χρόνια προϋπηρεσίας θα οδηγήσει σε έγκριση, ενώ για περισσότερα χρόνια αρκεί μια αύξηση 30.000 δολαρίων.

AReS (Actionable Recourse Summaries) Το *AReS* (Rawal and Lakkaraju, 2020) είναι από τις πρώτες εργασίες στις καθολικές αντιθετικές εξηγήσεις και στοχεύει στην παραγωγή ερμηνεύσιμων συνόψεων κανόνων για τη δικαιοπραξία. Επιδιώκει να βρει κανόνες του

[3] Το DiCE το χρησιμοποιεί και η υλοποίηση του GLANCE ως υπορουτίνα για την παραγωγή εξηγήσεων, για αυτό το λόγο του δίνουμε λίγο παραπάνω προσοχή.

τύπου: "αν $f_1 \lessgtr v_1$ και $f_2 \lessgtr v_2$, τότε αλλάξτε το f_1 σε v_1' και το f_2 σε v_2'". Το AReS διατυπώνει τον στόχο του με μια συνάρτηση βελτιστοποίησης που συνδυάζει την κάλυψη, την αποτελεσματικότητα και το κόστος:

$$\lambda_1 \cdot \text{Cov} + \lambda_2 \cdot \text{Eff} - \lambda_3 \cdot \text{Cost}_1 - \lambda_4 \cdot \text{Cost}_2,$$

όπου:

- Cov και Eff αντιπροσωπεύουν την κάλυψη και την αποτελεσματικότητα των κανόνων,
- Cost_1 και Cost_2 είναι οι διαφορετικές διαστάσεις του κόστους που συνδέονται με την εφαρμογή των κανόνων.

Οι παράμετροι λ καθορίζουν τις προτεραιότητες, ενώ εισάγονται και περιορισμοί που στοχεύουν να διατηρείται ο αριθμός και το εύρος των κανόνων κάτω από κάποιο κατώφλι, για διατήρηση της ερμηνευσιμότητας. Αν και το AReS αποδεικνύεται αποτελεσματικό στις περισσότερες περιπτώσεις, παρουσιάζει προβλήματα ταχύτητας σε μεγαλύτερα σύνολα δεδομένων, και η αποτελεσματικότητά του έχει αρκετή διακύμανση σε ορισμένα παραδείγματα (όπως θα δούμε και στη συνέχεια), καθιστώντας το λιγότερο αξιόπιστο σε κάποιες εφαρμογές.

GLOBE-CE Μία πιο πρόσφατη μέθοδος για καθολικές αντιθετικές εξηγήσεις είναι το *GLOBE-CE* (Ley et al., 2023). Το GLOBE-CE βελτιώνει τις καθολικές αντιθετικές εξηγήσεις, επιτρέποντας σε διαφορετικά άτομα να έχουν προσαρμοσμένα μεγέθη αλλαγών, αλλά με τον περιορισμό όλες οι αλλαγές να είναι στην ίδια κατεύθυνση, προσαρμόζοντας την κάλυψη και το κόστος σύμφωνα με τις ανάγκες του πληθυσμού. Αντί να περιορίζεται σε σταθερές αλλαγές για όλους, το GLOBE-CE εφαρμόζει ένα προσαρμόσιμο μέγεθος ενεργειών, κάνοντας τις αντιθετικές εξηγήσεις πιο αποτελεσματικές και οικονομικές. Η χρήση κανόνων If/Then επιτρέπει την ενσωμάτωση κατηγορικών χαρακτηριστικών, κάνοντας τη μέθοδο πιο προσιτή για ποικιλόμορφα δεδομένα.

Αν και δεν διατυπώνει ρητά κάποιο πρόβλημα βελτιστοποίησης, το *GLOBE-CE* στοχεύει στην εξισορρόπηση της κάλυψης με το ελάχιστο κόστος, όπου κάθε άτομο μπορεί να επηρεάζεται διαφορετικά ανάλογα με την προσαρμογή του μεγέθους. Παρότι αυτό βελτιώνει την κάλυψη και το μέσο κόστος, μειώνει την ερμηνευσιμότητα σε περιπτώσεις όπου η ποικιλία των μεγεθών αυξάνεται πολύ.

CET (Counterfactual Explanation Tree) Το CET (Counterfactual Explanation Tree) συνδυάζει τη *διαφάνεια* και τη *συνέπεια* μέσω ενός δέντρου αποφάσεων, αποδίδοντας αντιθετικές ενέργειες σε ομάδες ατόμων με παρόμοια χαρακτηριστικά. Η *δομή δέντρου* επιτρέπει τη σαφή και συνεπή εξήγηση, ειδικά σε περιπτώσεις οργανισμών όπου η διαφάνεια είναι απαραίτητη. Το CET προσπαθεί να ελαχιστοποιήσει το κόστος ενώ ταυτόχρονα μεγιστοποιεί την αποτελεσματικότητα και τη διατηρεί όσο το δυνατόν απλούστερη με βάση τον αριθμό των φύλλων του δέντρου.

Το πρόβλημα βελτιστοποίησης που προσπαθεί να λύσει το CET διαμορφώνεται ως ελαχιστοποίηση της ακόλουθης αντικειμενικής συνάρτησης:

$$\text{Cost} - \gamma\text{Eff} + \lambda(\#\text{Leaves}),$$

όπου:

- **Cost** είναι το μέσο κόστος δικαιοπραξίας για τα άτομα,
- **Eff** είναι η αποτελεσματικότητα των αντιθετικών δράσεων, και
- **# Leaves** είναι το πλήθος των φύλλων, που αντικατοπτρίζει την πολυπλοκότητα του δέντρου.

Η μέθοδος CET εξισορροπεί τα αντικρουόμενα κριτήρια, με την πολυπλοκότητα του δέντρου να μειώνει την ερμηνευσιμότητα και να επιβάλλει υψηλές υπολογιστικές απαιτήσεις. Παρά την επιτυχία του CET στην παραγωγή συνεπών και διαφανών εξηγήσεων, η υπολογιστική του απαίτηση είναι εξαιρετικά υψηλή και η αποτελεσματικότητα είναι σχετικά χαμηλή σε μεγάλα σύνολα δεδομένων.

Διατύπωση Προβλήματος των Καθολικών Αντιθετικών Εξηγήσεων από GLANCE Για τον ορισμό των καθολικών αντιθετικών εξηγήσεων, το *GLANCE* εμπνέεται από τις εργασίες που συζητήσαμε παραπάνω και θέτει τρία βασικά κριτήρια: χαμηλό κόστος, υψηλή αποτελεσματικότητα και περιορισμένο πλήθος ενεργειών. Έτσι, μια καθολική αντιθετική εξήγηση πρέπει να επιλέγει το μικρότερο δυνατό σύνολο ενεργειών, να ελαχιστοποιεί το κόστος αυτών των ενεργειών και να μεγιστοποιεί την κάλυψη των στιγμιότυπων. Αυτά τα κριτήρια διαμορφώνονται στο εξής multiobjective πρόβλημα βελτιστοποίησης:

$$\begin{aligned} &\underset{S \subseteq \mathcal{A}}{\text{minimize}}\ \mathsf{size}(S) \\ &\underset{S \subseteq \mathcal{A}}{\text{minimize}}\ \mathsf{avc}(S, \mathcal{D}_{aff}) \\ &\underset{S \subseteq \mathcal{A}}{\text{maximize}}\ \mathsf{eff}(S, \mathcal{D}_{aff}) \end{aligned} \tag{G.1}$$

όπου:

- $\mathsf{size}(S)$ αντιπροσωπεύει το πλήθος των δράσεων στο σύνολο S,
- $\mathsf{avc}(S, \mathcal{D}_{aff})$ το μέσο κόστος δικαιοπραξίας για τα επηρεαζόμενα στιγμιότυπα (affected instances), και
- $\mathsf{eff}(S, \mathcal{D}_{aff})$ την αποτελεσματικότητα των δράσεων για τις περιπτώσεις που επηρεάζονται αρνητικά.

Στόχος είναι να επιτευχθεί η βέλτιστη ισορροπία μεταξύ αυτών των αντικρουόμενων στόχων, καθώς η αύξηση της κάλυψης και η μείωση του κόστους συνήθως απαιτούν αύξηση των δράσεων. Το GLANCE χρησιμοποιεί συσταδοποίηση για να επιλέξει ένα σύνολο αντιπροσωπευτικών δράσεων που καλύπτουν μεγάλες ομάδες στιγμιότυπων με παρόμοιες ανάγκες.

Δικαιοσύνη της Δικαιοπραξίας: μία Σημαντική Εφαρμογή των Αντιθετικών Εξηγήσεων

Οι αντιθετικές εξηγήσεις παρέχουν στους χρήστες καθοδήγηση για το πώς να αλλάξουν χαρακτηριστικά τους ώστε να πετύχουν ευνοϊκά αποτελέσματα, όπως έγκριση δανείου ή προαγωγή. Ωστόσο, εκτός από την επεξηγησιμότητα, οι αντιθετικές εξηγήσεις μπορούν να χρησιμοποιηθούν και για τη δικαιοσύνη στη δικαιοπραξία, διασφαλίζοντας ότι ομάδες που θα πρέπει να προστατεύονται από μεροληψία, όπως φυλή ή φύλο, δεν αντιμετωπίζουν άνιση επιβάρυνση για να επιτύχουν ένα θετικό αποτέλεσμα.

Παραδοσιακά, η δικαιοσύνη στη μηχανική μάθηση έχει επικεντρωθεί στη δικαιοσύνη της πρόβλεψης (Mehrabi et al., 2022), η οποία στοχεύει στο να μην επηρεάζονται δυσανάλογα οι προβλέψεις του μοντέλου από χαρακτηριστικά όπως η φυλή ή το φύλο. Ωστόσο, η προσέγγιση αυτή δεν εξετάζει τη δυσκολία που μπορεί να αντιμετωπίζουν συγκεκριμένες ομάδες στην επίτευξη του ευνοϊκού αποτελέσματος, δηλαδή το *κόστος δικαιοπραξίας*. Η δικαιοσύνη της δικαιοπραξίας προχωρά πέρα από τη δικαιοσύνη της πρόβλεψης, εστιάζοντας στην εξασφάλιση ίσων ευκαιριών αλλαγής των χαρακτηριστικών μεταξύ προστατευόμενων ομάδων.

Για να αξιολογηθεί η δικαιοσύνη της δικαιοπραξίας, γίνεται χρήση αντιθετικών εξηγήσεων, που παρέχουν τις απαραίτητες ενέργειες για αλλαγή απόφασης, οι οποίες φέρουν κόστος για το άτομο. Η μέση τιμή του κόστους αυτού, γνωστή ως *επιβάρυνση (burden)* (Sharma et al., 2020; Kuratomi et al., 2023), υποδεικνύει πόσο εύκολο είναι για μια ομάδα να επιτύχει δικαιοπραξία. Εάν τα μέλη μιας προστατευόμενης ομάδας αντιμετωπίζουν υψηλότερο κόστος από μια άλλη για την ίδια αλλαγή απόφασης, το σύστημα θεωρείται άδικο. Ωστόσο, η μέση τιμή του κόστους μπορεί να καλύψει ή να παραποιήσει ανισότητες. Για αυτό το λόγο, τόσο στο Kavouras et al. (2023) όσο και στην παρούσα εργασία, χρησιμοποιούνται κατανομές κόστους, που προσφέρουν πιο λεπτομερείς ενδείξεις.

Επίσης, η δικαιοσύνη της δικαιοπραξίας μπορεί να αξιολογηθεί και σε συλλογικό επίπεδο. Η οπτική αυτή εξετάζει το κόστος που προκύπτει για ολόκληρες ομάδες ατόμων και αποκαλύπτει συστημικές προκαταλήψεις, οι οποίες δεν είναι πάντοτε εμφανείς με ατομικές αναλύσεις (Kavouras et al., 2023). Επιπλέον, οι συγγραφείς στο Kavouras et al. (2023) προτείνουν, συν τοις άλλοις, μεθόδους-ορισμούς που αγνοούν το κόστος, και επικεντρώνονται στον αριθμό των διαθέσιμων ενεργειών για κάθε ομάδα ('equal choice for recourse').

Η δικαιοσύνη της δικαιοπραξίας, λοιπόν, είναι μία σημαντική εφαρμογή των αντιθετικών εξηγήσεων, η οποία επεκτείνει τη δικαιοσύνη στη μηχανική μάθηση από τα αποτελέσματα στις δυνατότητες αλλαγής, εξασφαλίζοντας ότι όλες οι υποομάδες έχουν ίσες δυνατότητες να επηρεάσουν τις αποφάσεις του συστήματος μέσω αντιθετικών εξηγήσεων.

GLANCE: Προσέγγιση Συσταδοποίησης για την Γέννεση Καθολικών Αντιθετικών Εξηγήσεων

Για την εύρεση λύσης στο Πρόβλημα G.1, το Kavouras et al. (2024) ανέπτυξε μία σειρά από μεθόδους, μία εκ των οποίων είναι και ο αλγόριθμος *Augmented Space* που θα εξετάσουμε εδώ. Εν συντομία, η μέθοδος Augmented Space είναι ένας αλγόριθμος ο οποίος βασίζεται στην συσταδοποίηση (clustering) και στην απεικόνιση των δεδομένων σε έναν επαυξημένο χώρο που συνδυάζει τις τιμές των χαρακτηριστικών με τις προτεινόμενες ενέργειες.

Η διαδικασία λειτουργεί σε τρεις κύριες φάσεις (για περισσότερες λεπτομέρειες, παραπέμπουμε τον αναγνώστη και στον ψευδοκώδικα 1 και την αντίστοιχη ενότητα):

1. **Δημιουργία Αρχικών Συστάδων και Ενεργειών**: Στην αρχή, ο χώρος χαρακτηριστικών διαιρείται σε k μικρές συστάδες χρησιμοποιώντας μεθόδους συσταδοποίησης όπως το k-means. Στη συνέχεια, για το κεντροειδές κάθε συστάδας, δημιουργούνται m αντιθετικές ενέργειες, συνήθως με τοπικές μεθόδους όπως το DiCE, ώστε να διερευνηθούν διάφορες ενέργειες που περνούν το όριο απόφασης του μοντέλου.

2. **Δημιουργία του Επαυξημένου Χώρου**: Για κάθε στιγμιότυπο, επιλέγεται η πιο οικονομική ενέργεια από την συστάδα στην οποία αυτό ανήκει και συνδυάζεται με το

```
ACTIONS:
 (1) NumInqLast6Mexcl7days + 9.1
 (2) NumInqLast6Mexcl7days + 13.87 & NumSatisfactoryTrades - 0.75
 (3) NumInqLast6Mexcl7days + 10.27
```

Σχήμα G.3: Προτεινόμενες ενέργειες (Algorithm 1) για το σύνολο δεδομένων HELOC και μοντέλο λογιστικής παλινδρόμησης.

Πίνακας G.1: Μέγεθος, κόστος και αποτελεσματικότητα των συνόλων ενεργειών που υπολογίζονται από διαφορετικές μεθόδους για το σύνολο δεδομένων HELOC. Η γραμμή "LOCAL CE" αναφέρεται σε απλές τοπικές αντιθετικές εξηγήσεις (κατασκευασμένες με χρήση του DiCE (Mothilal et al., 2019)).

Method	Actions	Cost	Effectiveness
Algorithm 1	3	1.38	100.00%
Globe-CE	588	0.41	99.90%
Local CE	1045	1.16	100.00 %

αρχικό στιγμιότυπο σε έναν επαυξημένο χώρο που περιλαμβάνει και τις τιμές των χαρακτηριστικών και την ενέργεια. Έτσι, κάθε στιγμιότυπο αναπαρίσταται πλέον από το συνδυασμένο του προφίλ χαρακτηριστικών και την καλύτερη ενέργεια.

3. **Τελική Συσταδοποίηση στον Επαυξημένο Χώρο**: Στον τελικό επαυξημένο χώρο, τα στιγμιότυπα ομαδοποιούνται εκ νέου σε t νέες συστάδες. Από κάθε νέα ομάδα επιλέγεται μία αντιπροσωπευτική ενέργεια που να εφαρμόζεται σε όλα τα στιγμιότυπα της ομάδας, εξασφαλίζοντας ότι το τελικό σύνολο ενεργειών είναι οικονομικό και αντιπροσωπευτικό.

Ο αλγόριθμος *Augmented Space* δίνει ιδιαίτερη έμφαση στη μείωση του κόστους δικαιοπραξίας, καθώς κάθε στιγμιότυπο συνδέεται με την ενέργεια χαμηλότερου κόστους από την αρχική του ομάδα. Έτσι, η τελική επιλογή των t ενεργειών επιτυγχάνει έναν καλό συμβιβασμό μεταξύ κόστους και αποτελεσματικότητας. Ο αλγόριθμος εμφανίζει ταχύτητα και προσαρμοστικότητα, ενώ η ποιότητα των λύσεών του είναι συγκρίσιμη με τοπικές μεθόδους, όπως φαίνεται στα αποτελέσματα των πειραμάτων του GLANCE.

Ένα παράδειγμα αυτών των αποτελεσμάτων, συγκεκριμένα για το σύνολο δεδομένων HELOC, παρουσιάζεται στο Σχήμα G.3 (όπου φαίνεται η έξοδος του αλγορίθμου) και στον Πίνακα G.1 (όπου συγκρίνεται το αποτέλεσμα του αλγορίθμου AS με δύο άλλες βασικές προσεγγίσεις, με όρους κόστους, αποτελεσματικότητας και πλήθους ενεργειών).

Στο παράδειγμα αυτό, ο αλγόριθμος Augmented Space παράγει τρεις συνοπτικές και ερμηνεύσιμες ενέργειες με υψηλή αποτελεσματικότητα (100%) και ελάχιστα αυξημένο κόστος, σε αντίθεση με τις 588 ενέργειες της GLOBE-CE και τις 1045 τοπικές αντιθετικές εξηγήσεις, αποδεικνύοντας την αποδοτικότητά της στην απλοποίηση των καθολικών αντιθετικών εξηγήσεων.

Πειραματική Αξιολόγηση

Σε αυτή την ενότητα θα δούμε τα βασικά αποτελέσματα και αντίστοιχα συμπεράσματα που προέκυψαν από τα πειράματα με τον αλγόριθμο Augmented Space. Για λεπτο-

Πίνακας G.2: Σύγκριση της αποτελεσματικότητας και του κόστους του αλγορίθμου Augmented Space με τις μεθόδους Fast AReS, CET, Group-CF, GLOBE-CE και dGLOBE-CE για την επίλυση του προβλήματος G.1 με τον περιορισμό $\mathsf{size}(S) = 4$. Οι λύσεις με αποτελεσματικότητα κάτω από το όριο εφαρμοσιμότητας του 80% επισημαίνονται με μωβ. Οι μη αξιόπιστες GCE — δηλαδή εκείνες με τυπική απόκλιση (std) στην αποτελεσματικότητα άνω του 5% ή με std στο κόστος μεγαλύτερη από το μισό του μέσου κόστους — επισημαίνονται με κόκκινο.

Models	Algorithms	Datasets									
		Adult		COMPAS		Default Credit		German Credit		HELOC	
		eff	cost	eff	cost	eff	cost	eff	cost	eff	cost
dnn	Fast AReS	12.39 ± 1.06	1.0 ± 0.0	55.0 ± 0.86	1.21 ± 0.09	18.88 ± 2.16	1.0 ± 0.0	52.39 ± 1.63	1.0 ± 0.0	12.19 ± 0.58	1.03 ± 0.05
	CET	NaN	NaN	63.62 ± 10.35	0.96 ± 0.24	98.87 ± 0.62	6.32 ± 2.28	97.3 ± 2.46	1.58 ± 0.54	86.78 ± 10.62	8.67 ± 3.25
	GroupCF	100.0 ± 0.0	10.08 ± 0.03	100.0 ± 0.0	4.48 ± 2.53	79.6 ± 20.79	1.53 ± 0.62	97.8 ± 4.4	1.85 ± 0.13	80.4 ± 10.17	3.09 ± 0.91
	GLOBE-CE	99.92 ± 0.0	4.24 ± 0.42	100.0 ± 0.0	4.54 ± 3.31	76.94 ± 37.55	5.14 ± 0.35	93.31 ± 3.48	2.0 ± 1.55	42.72 ± 46.97	11.77 ± 15.87
	dGLOBE-CE	99.92 ± 0.0	10.89 ± 1.37	100.0 ± 0.0	7.96 ± 3.91	87.38 ± 18.69	5.96 ± 4.14	97.36 ± 0.82	2.49 ± 0.27	99.96 ± 0.05	11.07 ± 8.6
	Augmented Space	100.0 ± 0.0	5.92 ± 0.66	100.0 ± 0.0	4.24 ± 1.83	99.62 ± 0.29	1.36 ± 0.33	98.11 ± 1.68	1.05 ± 0.04	99.36 ± 0.86	11.96 ± 2.44
lr	Fast AReS	11.74 ± 2.4	1.0 ± 0.0	62.5 ± 1.82	1.24 ± 0.14	10.85 ± 5.45	1.07 ± 0.13	75.27 ± 2.96	1.0 ± 0.0	9.23 ± 1.24	1.12 ± 0.1
	CET	NaN	NaN	73.18 ± 4.34	1.24 ± 0.15	100.0 ± 0.0	3.79 ± 1.31	96.5 ± 2.85	2.42 ± 0.24	100.0 ± 0.0	3.57 ± 1.48
	GroupCF	100.0 ± 0.0	1.71 ± 0.39	100.0 ± 0.0	3.97 ± 2.38	95.4 ± 9.2	1.94 ± 1.2	97.6 ± 2.94	9.34 ± 3.85	90.6 ± 3.93	2.4 ± 1.38
	GLOBE-CE	99.92 ± 0.0	2.68 ± 0.17	95.74 ± 8.52	5.14 ± 3.77	99.94 ± 0.07	3.42 ± 1.99	57.09 ± 20.03	0.75 ± 1.04	99.9 ± 0.0	0.6 ± 0.54
	dGLOBE-CE	99.92 ± 0.0	5.91 ± 0.93	100.0 ± 0.0	6.71 ± 0.23	99.94 ± 0.07	10.38 ± 7.76	69.89 ± 15.35	2.47 ± 0.23	99.9 ± 0.0	1.63 ± 0.35
	Augmented Space	100.0 ± 0.0	1.47 ± 0.44	100.0 ± 0.0	5.49 ± 0.92	100.0 ± 0.0	1.22 ± 0.26	99.55 ± 0.91	1.64 ± 0.32	100.0 ± 0.0	2.37 ± 1.03
xgb	Fast AReS	6.13 ± 0.42	1.0 ± 0.0	59.83 ± 3.12	1.1 ± 0.05	31.86 ± 5.12	1.05 ± 0.04	51.27 ± 1.57	1.0 ± 0.0	8.49 ± 1.32	1.16 ± 0.13
	CET	NaN	NaN	58.4 ± 9.3	1.06 ± 0.24	86.29 ± 9.94	4.5 ± 2.64	100.0 ± 0.0	2.73 ± 0.49	86.78 ± 6.7	12.51 ± 2.75
	GroupCF	96.8 ± 1.72	1.41 ± 0.54	100.0 ± 0.0	4.06 ± 2.1	95.2 ± 1.6	1.41 ± 0.64	100.0 ± 0.0	5.78 ± 4.11	78.4 ± 5.82	5.63 ± 1.93
	GLOBE-CE	82.87 ± 12.14	30.1 ± 10.39	87.13 ± 11.14	9.75 ± 7.2	82.7 ± 7.26	20.82 ± 1.73	77.05 ± 11.26	1.14 ± 1.24	27.66 ± 5.06	12.52 ± 32.48
	dGLOBE-CE	93.76 ± 1.98	64.76 ± 1.29	99.84 ± 0.31	12.46 ± 3.42	97.47 ± 0.82	42.58 ± 3.57	86.96 ± 9.79	2.66 ± 0.77	77.64 ± 11.51	128.0 ± 0.0
	Augmented Space	99.78 ± 0.14	6.15 ± 3.05	98.77 ± 1.17	3.54 ± 1.04	95.33 ± 1.44	2.52 ± 0.55	99.43 ± 1.14	1.06 ± 0.08	98.33 ± 1.42	23.94 ± 1.66

μερείς πληροφορίες σχετικά με την πειραματική μας διάταξη (σύνολα δεδομένων, μοντέλα κλπ), παραπέμπουμε τον αναγνώστη στην ενότητα 3.2 του αγγλόφωνου μέρους της εργασίας.

Για την πειραματική αξιολόγηση του Augmented Space αλγορίθμου στο πλαίσιο του προβλήματος καθολικών αντιθετικών εξηγήσεων G.1, παραθέτουμε και σχολιάζουμε τα αποτελέσματα του αλγορίθμου σε σύγκριση με κάποιες σημαντικές υπάρχουσες μεθόδους:

- AReS, χρησιμοποιώντας την υλοποίηση Fast-AReS του Ley et al. (2022)
- CET (Kanamori et al., 2022)
- GroupCF (Warren et al., 2023)
- GLOBE-CE και την ελαφρά παραλλαγή του dGLOBE-CE (Ley et al., 2023)

Βασικά κριτήρια σύγκρισης αποτελούν τα μεγέθη του ορισμού, δηλαδή: αποτελεσματικότητα, κόστος, αριθμός ενεργειών. Φυσικά, σε αυτά προστίθεται και ο απαιτούμενος χρόνος εκτέλεσης.

Επισκόπηση Αποτελεσμάτων

Στον Πίνακα G.2 συνοψίζονται τα βασικά αποτελέσματα όλων των μεθόδων. Ο αλγόριθμος Augmented Space επιτυγχάνει υψηλή αποτελεσματικότητα (πάνω από 80%) και σταθερά καλό κόστος στις περισσότερες περιπτώσεις, ενώ οι άλλες μέθοδοι παρουσιάζουν συχνά αποτελεσματικότητα κάτω από 80% ή υψηλότερο κόστος.

Πίνακας G.3: Αξιολόγηση της κυριαρχίας Pareto μεταξύ λύσεων για το Πρόβλημα 2.6 με τον περιορισμό $\mathsf{size}(S) = 4$. Ο πίνακας δείχνει τη συχνότητα (αριθμός συγκρίσεων επί του συνόλου των δυνατών συγκρίσεων) στην οποία ο αλγόριθμος Augmented Space υπερισχύει έναντι των ανταγωνιστών (στήλη DOMINATES) καθώς και τη συχνότητα στην οποία υπερισχύουν οι ανταγωνιστές έναντι αυτού (στήλη IS DOMINATED).

	Augmented Space	
	Dominates	is Dominated
Fast AReS	0/15	0/15
CET	6/12	0/12
GroupCF	8/15	1/15
GLOBE-CE	9/15	0/15
dGLOBE-CE	11/15	1/15
competitors	34/72	2/72

Θα πρέπει να σημειωθεί ότι στα συγκεκριμένα αποτελέσματα όλες οι μέθοδοι εκτελούνται υπό περιορισμό μέγιστου αριθμού ενεργειών ($\mathsf{size}(S) = 4$), προκειμένου να διασφαλιστεί η συγκρισιμότητα των αποτελεσμάτων. Η σύγκριση γίνεται τόσο σε μικρά όσο και σε μεγάλα σύνολα δεδομένων, καθώς και σε τρεις διαφορετικούς τύπους μοντέλων.

Κυριαρχία Pareto Ένας τρόπος με τον οποίο αξιολογούμε την απόδοση των μεθόδων είναι εξετάζοντας εάν μία λύση υπερισχύει έναντι κάποιας άλλης με βάση την αποτελεσματικότητα, το κόστος και το μέγεθος. Συγκεκριμένα, *μία λύση S για το Πρόβλημα G.1* κυριαρχεί κατά Pareto *σε μία άλλη λύση S', εάν ισούται ή υπερνικά το S' σε αποτελεσματικότητα και κόστος, έχει ίσο ή μικρότερο μέγεθος και είναι αυστηρά καλύτερη σε τουλάχιστον έναν από αυτούς τους στόχους.*[4]

Ο Πίνακας G.3 δείχνει ότι ο αλγόριθμος Augmented Space κυριαρχεί έναντι άλλων μεθόδων σε 34 από τις 72 περιπτώσεις (47%) και υπερτερείται μόνο μία φορά από τον GroupCF (στο COMPAS-LR, βλ. Πίνακα G.2) και μία φορά από τον dGLOBE-CE (στο HELOC-DNN). Αξιοσημείωτο είναι ότι, σε αυτές τις δύο περιπτώσεις, η αποτελεσματικότητα του Augmented Space είναι πολύ κοντά σε αυτήν των μεθόδων που υπερισχύουν.

Πρακτικότητα Λύσεων Στην αξιολόγηση του Pareto dominance (Πίνακας G.3), συγκρίνουμε λύσεις με βέλτιστη ή σχεδόν βέλτιστη αποτελεσματικότητα με άλλες που εμφανίζουν σαφώς χαμηλότερη αποτελεσματικότητα. Αυτές οι χαμηλότερης αποτελεσματικότητας λύσεις θεωρούνται μη πρακτικές για τον ρόλο των καθολικών αντιθετικών εξηγήσεων (GCEs) από τους συγγραφείς του Kavouras et al. (2024), καθώς στόχος είναι η παροχή δικαιοπραξίας σε σημαντικό μέρος του πληθυσμού. Οι λύσεις με χαμηλή αποτελεσματικότητα δεν πετυχαίνουν αυτόν τον στόχο, μειώνοντας τη σημασία τους για εφαρμογές στον πραγματικό κόσμο. Αν μια λύση αφήνει μεγάλο τμήμα των ατόμων χωρίς βιώσιμες επιλογές για δικαιοπραξία, αποτυγχάνει στον θεμελιώδη σκοπό των GCEs, όπως επισημαίνεται και από το Ley et al. (2023). Επιπλέον, η επίτευξη χαμηλού κόστους δικαιοπραξίας είναι πιο εφικτή για μικρότερες υποομάδες, ιδιαίτερα εκείνες κοντά στο όριο απόφασης. Ακολουθώντας το Kavouras et al. (2024), ορίζουμε μια λύση ως πρακτική εφόσον επιτυγχάνει

[4] Η κυριαρχία Pareto είναι μια ευρέως καθιερωμένη έννοια στη βελτιστοποίηση πολλαπλών στόχων. Για περισσότερες πληροφορίες, βλέπε Branke (2008).

Πίνακας G.4: Αξιολόγηση της κυριαρχίας Pareto μεταξύ των λύσεων για το Πρόβλημα G.1 με περιορισμένο $\mathsf{size}(S) = 4$, εφαρμόζοντας το όριο $\mathsf{eff} > 80\%$. Ο πίνακας δείχνει τη συχνότητα (αριθμός συγκρίσεων εκ του συνολικού) κατά την οποία ο αλγόριθμος Augmented Space κυριαρχεί έναντι των άλλων μεθόδων (στην στήλη DOMINATES) και τη συχνότητα κατά την οποία ο Augmented Space κυριαρχείται από τους ανταγωνιστές (στην στήλη IS DOMINATED).

	Augmented Space	
	Dominates	is Dominated
Fast AReS	0/0	0/0
CET	6/9	0/9
GroupCF	7/13	1/13
GLOBE-CE	7/10	0/10
dGLOBE-CE	9/13	1/13
Competitors	29/45	2/45

τουλάχιστον 80% αποτελεσματικότητα. Στον Πίνακα G.2, οι μη πρακτικές λύσεις σημειώνονται με μωβ. Αξιοσημείωτο είναι ότι ο Augmented Space δεν επιστρέφει ποτέ μη πρακτικές λύσεις, ενώ όλες οι λύσεις που παράγονται από τον Fast AReS ταξινομούνται ως μη πρακτικές. Οι υπόλοιπες μέθοδοι παράγουν από 2 έως 5 μη πρακτικές λύσεις η καθεμία. Ο Πίνακας G.4 παρουσιάζει τις Pareto κυριαρχίες, αλλά αυτή τη φορά με συγκρίσεις μόνο μεταξύ πρακτικών λύσεων. Ο αλγόριθμος Augmented Space κυριαρχεί έναντι άλλων μεθόδων σε πολύ υψηλότερο ποσοστό — σε 29 από τις 45 περιπτώσεις, δηλαδή πάνω από 64%. Παραμένει κυριαρχούμενος στις ίδιες δύο περιπτώσεις, με το ποσοστό των περιπτώσεων στις οποίες υστερεί να βρίσκεται σε χαμηλό επίπεδο, μόλις στο 4%.

Ανθεκτικότητα. Η ανθεκτικότητα (robustness) είναι απαραίτητη για τις μεθόδους GCE, καθώς πρέπει να παράγουν *με συνέπεια* αποτελεσματικές, χαμηλού κόστους λύσεις σε διάφορους διαχωρισμούς δεδομένων, διασφαλίζοντας την αξιοπιστία για χρήση σε πραγματικές εφαρμογές. Η έλλειψη ανθεκτικότητας μπορεί να οδηγήσει σε σημαντικές διαφοροποιήσεις στις ενέργειες δικαιοπραξίας, μειώνοντας την εμπιστοσύνη και πιθανώς προκαλώντας άδικα αποτελέσματα, ιδιαίτερα σε κρίσιμους τομείς όπως η υγειονομική περίθαλψη και τα οικονομικά. Για να εκτιμήσουμε την πρακτική καταλληλότητα, είναι σημαντικό να αξιολογούμε τη σταθερότητα των μετρικών αποτελεσματικότητας και κόστους σε διαφορετικές διαιρέσεις. Η τυπική απόκλιση χρησιμεύει ως μέτρο αυτής της σταθερότητας: μια απόκλιση αποτελεσματικότητας άνω του 5% υποδεικνύει ασυνέπεια στην παροχή αξιόπιστης δικαιοπραξίας, ενώ μια απόκλιση κόστους μεγαλύτερη από το μισό του μέσου κόστους υποδηλώνει μη πρακτικές, απρόβλεπτες ενέργειες (τα κριτήρια αυτά έχουν ληφθεί απευθείας από το Kavouras et al. (2024)). Λύσεις με τόσο μεγάλη μεταβλητότητα είναι αναξιόπιστες και τις έχουμε επισημάνει με κόκκινο στον Πίνακα G.2.

Άλλα Βασικά Συμπεράσματα και Συγκρίσεις

Μικρά Σύνολα Δεδομένων Στα μικρότερα σύνολα, όπως COMPAS και German Credit, ο Augmented Space είναι ισοδύναμος ή καλύτερος σε κόστος και αποτελεσματικότητα σε σύγκριση με τις άλλες μεθόδους. Για παράδειγμα, στο German Credit, πετυχαίνει σχεδόν 100% αποτελεσματικότητα σε όλες τις περιπτώσεις, ενώ οι άλλες μέθοδοι, πλην GroupCF, πέφτουν κάτω από τα πρακτικά επίπεδα σε τουλάχιστον ένα μοντέλο.

Μεγάλα Σύνολα Δεδομένων Στα μεγαλύτερα σύνολα, όπως το Default Credit, ο Augmented Space υπερτερεί στις περισσότερες περιπτώσεις. Στο HELOC, εμφανίζει καλή απόδοση με μικρές αποκλίσεις, ενώ διατηρεί την αξιοπιστία του ακόμα και με μεγαλύτερο χρόνο εκτέλεσης.

Συνολική Απόδοση Ο Augmented Space επιτυγχάνει σταθερά και με συνέπεια καλή ισορροπία μεταξύ κόστους και αποτελεσματικότητας, κάνοντάς τον μια ευέλικτη και πρακτική λύση για πραγματικές εφαρμογές καθολικών αντιθετικών εξηγήσεων.

FACTS: Έλεγχος Δικαιοσύνης μέσω Επεξηγησιμότητας

Το Kavouras et al. (2023) έχει διπλή συνεισφορά, τόσο εννοιολογικά όσο και τεχνικά. Ξεκινάει ορίζοντας τυπικά την κατανομή αποτελεσματικότητας-κόστους, καθώς και την μικρο- και τη μακρο-προοπτική. Στη συνέχεια, βασίζεται σε αυτές τις έννοιες και διατυπώνει έναν αριθμό ορισμών για την δικαιοσύνη της δικαιοπραξίας (fairness of recourse), κάθε ένας από τους οποίους έχει σκοπό να αιχμαλωτίσει μία διαφορετική πλευρά της μεροληψίας. Επίσης, όλοι οι ορισμοί είναι άμεσα προσαρμόσιμοι σε μετρικές δικαιοσύνης της δικαιοπραξίας, τους οποίους χρησιμοποιεί ο αλγόριθμος που προτείνει το Kavouras et al. (2023).

Ο αλγόριθμος, λοιπόν, που αποτελεί την πιο τεχνική συνεισφορά του Kavouras et al. (2023), χρησιμοποιεί αυτές τις μετρικές για να κατατάξει τις πληθυσμιακές ομάδες που αναζητά ως προς την μεροληψία που παρουσιάζουν οι αντίστοιχες προστατευόμενες υποομάδες.

Τον αλγόριθμο αυτόν θα τον δούμε περαιτέρω στις επόμενες ενότητες. Για αρχή, θα δώσουμε μία σύντομη περιγραφή των ορισμών δικαιοσύνης που διατυπώνονται στο Kavouras et al. (2023). Για περισσότερες εξηγήσεις πάνω στους ορισμούς αυτούς, συμπεριλαμβανομένων και επεξηγηματικών παραδειγμάτων για τον κάθε ορισμό, παραπέμπουμε τον αναγνώστη στην Ενότητα 4.1.1 του αγγλόφωνου τμήματος της διπλωματικής.

Συσχέτιση Αποτελεσματικότητας-Κόστους

Εξετάζουμε τον τρόπο επίτευξης δικαιοπραξίας για την ομάδα G μέσω ενός συνόλου πιθανών ενεργειών A. Ορίζουμε την **συνολική αποτελεσματικότητα (aggregate effectiveness)** (aeff) του A για την ομάδα G με δύο τρόπους.

Στη *μικρο-οπτική*, τα άτομα της ομάδας θεωρούνται ανεξάρτητα, και το καθένα επιλέγει την ενέργεια που το ωφελεί περισσότερο.

Ορισμός G.1 (Μικρο-Αποτελεσματικότητα). *Ορίζουμε τη* ***μικρο-αποτελεσματικότητα*** *ενός συνόλου ενεργειών A για την ομάδα G ως το ποσοστό των ατόμων στην G που μπορούν να επιτύχουν δικαιοπραξία με* κάποια *ενέργεια στο A:*

$$\mathsf{aeff}_{\mu}(A, G) = \frac{1}{|G|}|\{x \in G | \exists a \in A, \mathsf{eff}(a, x) = 1\}|.$$

Στη *μακρο-οπτική*, η ομάδα εξετάζεται ως σύνολο, και μία ενέργεια εφαρμόζεται συλλογικά σε όλα τα άτομα.

Ορισμός G.2 (Μακρο-Αποτελεσματικότητα). *Ορίζουμε τη* ***μακρο-αποτελεσματικότητα*** *ενός συνόλου ενεργειών A για την ομάδα G ως το μεγαλύτερο ποσοστό ατόμων στην G που*

μπορούν να επιτύχουν δικαιοπραξία μέσω της ίδιας *ενέργειας στο* A*:*

$$\mathsf{aeff}_{\mathsf{M}}(A, G) = \max_{a \in A} \frac{1}{|G|} |\{x \in G | \, \mathsf{eff}(a, x) = 1\}|.$$

Για την ομάδα G, το σύνολο ενεργειών A και ένα προκαθορισμένο όριο κόστους c, ορίζουμε επίσης τις **ενέργειες εντός ορίου** ως το σύνολο των ενεργειών με κόστος που δεν υπερβαίνει το c για κανένα άτομο στην ομάδα G:

$$A_c = \{a \in A | \forall x \in G, \mathsf{cost}(a, x) \leq c\}$$

Χρησιμοποιώντας τους παραπάνω ορισμούς, εισάγουμε την **κατανομή αποτελεσματικότητας-κόστους** (ecd), που αξιολογεί ένα σύνολο ενεργειών σε διάφορα επίπεδα κόστους.

Ορισμός G.3 (Κατανομή Αποτελεσματικότητας-Κόστους). *Ορίζουμε την* ***κατανομή αποτελεσματικότητας-κόστους*** *(*ecd*) ως τη συνάρτηση που για έναν προϋπολογισμό κόστους* c *επιστρέφει την συνολική αποτελεσματικότητα με ενέργειες εντός προϋπολογισμού:*

$$\mathsf{ecd}(c; A, G) = \mathsf{aeff}(A_c, G)$$

Η ecd εκφράζει το ποσοστό ατόμων στην G που μπορούν να επιτύχουν δικαιοπραξία με κόστος το πολύ c, ενώ ecd_{μ} και $\mathsf{ecd}_{\mathsf{M}}$ αναφέρονται στην μικρο- και μακρο-οπτική αντίστοιχα.

Ορισμοί Δικαιοσύνης στη Δικαιοπραξία για Υποομάδες

Ορίζουμε τη δικαιοσύνη της δικαιοπραξίας για έναν ταξινομητή h συγκρίνοντας τις συναρτήσεις ecd των προστατευόμενων υποομάδων. Για απλότητα, υποθέτουμε ότι το προστατευόμενο χαρακτηριστικό λαμβάνει μόνο δύο τιμές, και άρα οι προστατευόμενες υποομάδες μίας ομάδας G είναι δύο, έστω G_0 και G_1.

Ορισμός G.4 (Equal Effectiveness). *Ο ταξινομητής είναι δίκαιος αν το ίδιο ποσοστό ατόμων στις προστατευόμενες υποομάδες μπορεί να επιτύχει δικαιοπραξία:*

$$\mathsf{aeff}(A, G_0) = \mathsf{aeff}(A, G_1)$$

Ορισμός G.5 (Equal Choice for Recourse). *Ο ταξινομητής είναι δίκαιος αν οι προστατευόμενες υποομάδες έχουν τον ίδιο αριθμό επιλογών για δικαιοπραξία με επαρκή αποτελεσματικότητα (δηλαδή τουλάχιστον* $\phi \in [0, 1]$*):*

$$|\{a \in A | \, \mathsf{eff}(a, G_0) \geq \phi\}| = |\{a \in A | \, \mathsf{eff}(a, G_1) \geq \phi\}|.$$

Οι επόμενοι τρεις ορισμοί προϋποθέτουν γνωστό κόστος και έχουν εκδοχές και σε μικρο- και σε μακρο-προοπτική.

Ορισμός G.6 (Equal Cost of Effectiveness). *Ο ταξινομητής είναι δίκαιος αν το ελάχιστο κόστος για επίτευξη αποτελεσματικότητας* ϕ *στις προστατευμένες υποομάδες είναι ίσο:*

$$\mathsf{ecd}^{-1}(\phi; A, G_0) = \mathsf{ecd}^{-1}(\phi; A, G_1)$$

Ορισμός G.7 (Equal Effectiveness within Budget). *Ο ταξινομητής είναι δίκαιος αν το ίδιο ποσοστό ατόμων στις προστατευμένες υποομάδες μπορεί να επιτύχει δικαιοπραξία με κόστος το πολύ* c*:*

$$\mathsf{ecd}(c; A, G_0) = \mathsf{ecd}(c; A, G_1)$$

Ορισμός G.8 (Fair Effectiveness-Cost Trade-Off). *Ο ταξινομητής είναι δίκαιος αν οι προστατευμένες υποομάδες έχουν την ίδια κατανομή αποτελεσματικότητας-κόστους:*

$$\max_c |\,\mathsf{ecd}(c; A, G_0) - \mathsf{ecd}(c; A, G_1)| = 0$$

Στην παραπάνω συνθήκη, το αριστερό μέλος συνιστά την στατιστική Kolmogorov-Smirnov για δύο δείγματα. Με βάση αυτό, διατυπώνεται και μία στατιστική εκδοχή του παραπάνω ορισμού, για στατιστικό έλεγχο της δικαιοσύνης με εμπιστοσύνη α.

Ορισμός G.9 (Στατιστικό Fair Effectiveness-Cost Trade-Off). *Ο ταξινομητής είναι δίκαιος με εμπιστοσύνη α αν ο έλεγχος Kolmogorov-Smirnov για τις κατανομές* ecd *δεν απορρίπτει τη μηδενική υπόθεση:*

$$\max_c |\,\mathsf{ecd}(c; A, G_0) - \mathsf{ecd}(c; A, G_1)| \leq \sqrt{-\ln(\alpha/2)\frac{|G_{p,0}| + |G_{p,1}|}{2|G_{p,0}||G_{p,1}|}}$$

Ο τελευταίος ορισμός βασίζεται στη μικρο-προοπτική και επεκτείνει, ουσιαστικά, την έννοια της *επιβάρυνσης (burden)* (Sharma et al., 2020) που υπάρχει ήδη σε προηγούμενη βιβλιογραφία.

Ορισμός G.10 (Equal (Conditional) Mean Recourse). *Ο ταξινομητής είναι δίκαιος αν το (δεσμευμένο) μέσο κόστος δικαιοπραξίας για τις προστατευόμενες υποομάδες είναι ίσο:*

$$\bar{\mathsf{rc}}^*(A, G_0; A) = \bar{\mathsf{rc}}^*(A, G_1)$$

όπου με $\bar{\mathsf{rc}}^*$ συμβολίζεται ακριβώς το μέσο κόστος δικαιοπραξίας, με δεδομένο ότι επιτυγχάνεται δικαιοπραξία (για ακριβή τυπικό ορισμό, παραπέμπουμε στην Ενότητα 4.1.1).

Όταν η ομάδα G είναι ολόκληρο το σύνολο των επηρεαζόμενων ατόμων, και όλοι μπορούν να επιτύχουν δικαιοπραξία μέσω κάποιας ενέργειας από το A, αυτός ο ορισμός συμπίπτει με τον αντίστοιχο για την επιβάρυνση.

Έλεγχος Δικαιοσύνης στη Δικαιοπραξία μέσω Αντιθετικών Εξηγήσεων σε Ομάδες: FACTS

Εδώ, παρουσιάζουμε το FACTS (Fairness-aware Counterfactuals for Subgroups), που είναι ουσιαστικά ο αλγόριθμος που προτείνεται στο Kavouras et al. (2023) για την ανίχνευση ομάδων πληθυσμού που μπορεί να δέχονται άνιση μεταχείριση από το άγνωστο μοντέλο, με βάση τη δυσκολία επίτευξης δικαιοπραξίας των προστατευόμενων υποομάδων τους. Για να αξιολογήσει και να συγκρίνει ομάδες και να ποσοτικοποιήσει την 'αδικία της δικαιοπραξίας', ο αλγόριθμος εφαρμόζει τη μικρο- και μάκρο-προοπτική και όλους τους ορισμούς δικαιοσύνης που περιγράφηκαν στην προηγούμενη υποενότητα. Το αποτέλεσμα του FACTS είναι πληθυσμιακές ομάδες (οριζόμενες από συνδυασμούς τιμών σε χαρακτηριστικά), οι οποίες συνοδεύονται επιπλέον από (α) μία βαθμολογία αδικίας, που αποτυπώνει τις διαφορές μεταξύ προστατευόμενων υποομάδων και (β) μια εύκολα κατανοητή σύνοψη αντιθετικών εξηγήσεων, που ονομάζεται *Συγκριτικές Αντιθετικές Επεξηγήσεις Υποομάδων* (Comparative Subgroup Counterfactuals - CSC).

Στο Σχήμα G.4 παρουσιάζεται ένα παράδειγμα αποτελέσματος του FACTS, από το σύνολο δεδομένων Adult (Bache and Lichman, 2013). Η συνθήκη "if" περιγράφει την ομάδα G_p, η οποία περιέχει τα επηρεαζόμενα άτομα που ικανοποιούν τη συνθήκη $p = (\textit{hours-per-week = FullTime}) \wedge (\textit{marital-status = Married-civ-spouse}) \wedge$

```
If hours-per-week=FullTime, marital-status=Married-civ-spouse, occupation=Adm-clerical:
    Protected Subgroup = 'Male', 1.87% covered
        Make hours-per-week=Overtime, occupation=Exec-managerial with effectiveness 72.00%
    Protected Subgroup = 'Female', 1.80% covered
        No recourses for this subgroup.
    Bias against 'Female' due to Equal Cost of Effectiveness (threshold=0.7). Unfairness score = inf.
```

Σχήμα G.4: Αναπαράσταση CSC για μία υποομάδα με υψηλό σκορ μεροληψίας με βάση τον ορισμό *Equal Cost of Effectiveness* με $\phi = 0.7$.

(*occupation = Adm-clerical*). Τα στοιχεία κάτω από τη συνθήκη αφορούν τις προστατευόμενες υποομάδες $G_{p,0}$ και $G_{p,1}$, που είναι αντίστοιχα οι γυναίκες και οι άντρες του G_p. Με μπλε επισημαίνεται το ποσοστό κάλυψης, που υποδεικνύει το μέγεθος της προστατευόμενης υποομάδας. Το σημαντικότερο μέρος της αναπαράστασης είναι οι ενέργειες που εφαρμόζονται σε κάθε υποομάδα και αξιολογούνται με βάση την επιλεγμένη μετρική δικαιοσύνης, η οποία στο συγκεκριμένο παράδειγμα είναι το *Equal Cost of Effectiveness* με όριο αποτελεσματικότητας $\phi = 0.7$. Στο παράδειγμα, για το $G_{p,0}$ δεν υπάρχει ενέργεια που να υπερβαίνει το $\phi = 0.7$, ενώ στο $G_{p,1}$ η ενέργεια $a = \{$*hours-per-week* $\rightarrow$ *Overtime*, *occupation* $\rightarrow$ *Exec-managerial*$\}$ έχει αποτελεσματικότητα $0.72 > \phi$, προσφέρει δηλαδή δικαιοπραξία στο 72% των αντρών, ενώ καμία γυναίκα δεν επιτυγχάνει δικαιοπραξία.

Σύνοψη της Μεθόδου Ο αλγόριθμος FACTS περιλαμβάνει τρία βασικά στάδια: (α) *Δημιουργία υποομάδων και χώρου ενεργειών*, όπου παράγονται συχνές υποομάδες και αντίστοιχες ενέργειες από τον ανεπηρέαστο πληθυσμό, (β) *Δημιουργία συνόψεων αντιθετικών εξηγήσεων*, όπου γίνεται αντιστοίχιση υποομάδων και ενεργειών ώστε να δημιουργηθούν έγκυρες αντιθετικές εξηγήσεις για κάθε υποομάδα, και (γ) *Δημιουργία CSC και κατάταξη με βάση τη δικαιοσύνη*, όπου κάθε σύνοψη αποτιμάται με τις μετρικές της προηγούμενης Ενότητας και γίνεται κατάταξη των υποομάδων.

(α) Δημιουργία υποομάδων και χώρου ενεργειών Οι υποομάδες δημιουργούνται με την εκτέλεση του αλγορίθμου fp-growth (Han et al., 2000) στα D_0 και D_1, και στη συνέχεια με την τομή $\mathcal{G} = \mathcal{G}_0 \bigcap \mathcal{G}_1$. Τα στοιχεία αποτελούν συνθήκες μορφής 'χαρακτηριστικό-τελεστής-τιμή', και η τομή διασφαλίζει ότι οι προστατευόμενες υποομάδες θα αξιολογηθούν δίκαια. Το σύνολο όλων των ενεργειών A προκύπτει από τον ανεπηρέαστο πληθυσμό, με στόχο την αποδοτικότητα και τη μείωση της υπολογιστικής πολυπλοκότητας.

(β) Δημιουργία αντιθετικών συνόψεων Για κάθε υποομάδα $G_p \in \mathcal{G}$, βρίσκονται έγκυρες ενέργειες και υπολογίζεται η αποτελεσματικότητα $\mathsf{eff}(a, G_{p,0})$ και $\mathsf{eff}(a, G_{p,1})$ για κάθε έγκυρη ενέργεια a. Αυτή η προσέγγιση παρέχει σημαντικό πλεονέκτημα στο FACTS, καθώς εξασφαλίζει ότι το κόστος ενεργειών ορίζεται εξίσου για όλα τα άτομα της υποομάδας.

(γ) Δημιουργία CSC και κατάταξη δικαιοσύνης Στο τελικό στάδιο, το FACTS αποτιμά όλες τις ομάδες με βάση τις μετρικές της προηγούμενης ενότητας, παράγοντας βαθμολογίες αδικίας για κάθε ομάδα. Το αποτέλεσμα είναι μια ταξινομημένη λίστα CSC, που επιτρέπει στους χρήστες να κατανοήσουν εύκολα την μεροληψία συγκρίνοντας τις διαφορές στις ενέργειες των προστατευόμενων πληθυσμών μέσα σε κάθε ομάδα.

Πειραματική Αξιολόγηση

Σε αυτή την ενότητα παρουσιάζουμε βασικά πειραματικά αποτελέσματα από την εργασία Kavouras et al. (2023), καθώς και επιπλέον αποτελέσματα από τα πειράματα που διεξήγαμε στο πλαίσιο αυτής της διπλωματικής. Εξετάζουμε και αναλύουμε τις Συγκριτικές Αντιθετικές Επεξηγήσεις Υποομάδων (CSCs) για υποομάδες που κατατάσσονται ως οι πλέον άδικες βάσει των ορισμών δικαιοσύνης της πιο πάνω ενότητας. Για αυτή την ενότητα, επιλέχθηκαν συγκεκριμένα σύνολα δεδομένων και μοντέλα ως παραδείγματα.

Για λεπτομερείς πληροφορίες σχετικά με την πειραματική μας διάταξη (σύνολα δεδομένων, μοντέλα κλπ), παραπέμπουμε τον αναγνώστη στην ενότητα 4.3 του αγγλόφωνου μέρους της εργασίας. Επίσης, για περισσότερα, πιο αναλυτικά αποτελέσματα, παραπέμπουμε τον ενδιαφερόμενο αναγνώστη στο Παράρτημα Α.

Πίνακας G.5: Μεροληπτικές υποομάδες που εντοπίστηκαν στο σύνολο δεδομένων Adult.

	Group 1			Group 2			Group 3		
	rank	bias against	unfairness score	rank	bias against	unfairness score	rank	bias against	unfairness score
Equal Effectiveness	2950	Male	0.11	10063	Female	0.0004	275	Female	0.32
Equal Choice for Recourse ($\phi = 0.3$)	Fair	-	0	12	Female	2	Fair	-	0
Equal Choice for Recourse ($\phi = 0.7$)	6	Male	1	1	Female	6	Fair	-	0
Equal Effectiveness within Budget ($c = 5$)	Fair	-	0	2806	Female	0.056	70	Female	0.3
Equal Effectiveness within Budget ($c = 10$)	2350	Male	0.11	8518	Female	0.0004	226	Female	0.3
Equal Effectiveness within Budget ($c = 18$)	2675	Male	0.11	9222	Female	0.0004	272	Female	0.3
Equal Cost of Effectiveness ($\phi = 0.3$)	Fair	-	0	Fair	-	0	1	Female	inf
Equal Cost of Effectiveness ($\phi = 0.7$)	1	Male	inf	12	Female	2	Fair	-	0
Fair Effectiveness-Cost Trade-Off	4065	Male	0.11	3579	Female	0.13	306	Female	0.32
Equal (Conditional) Mean Recourse	Fair	-	0	3145	Female	0.35	Fair	-	0

Αδικημένες υποομάδες Ο Πίνακας G.5 παρουσιάζει τρεις πληθυσμιακές ομάδες που κατατάχθηκαν πρώτες (δηλαδή πιο πιθανές για μεροληψία στη δικαιοπραξία) σύμφωνα με τρεις διαφορετικούς ορισμούς δικαιοσύνης: *Equal Cost of Effectiveness* (με $\phi = 0.7$), *Equal Choice for Recourse* (με $\phi = 0.7$), και *Equal Cost of Effectiveness* (με $\phi = 0.3$). Αυτές οι ομάδες εντοπίστηκαν ως οι πλέον πιθανές για μεροληψία βάσει των αντίστοιχων ορισμών. Για κάθε ομάδα, ο πίνακας δίνει την *κατάταξή* της, την προστατευόμενη υποομάδα εναντίον της οποίας υπάρχει *μεροληψία*, και τη *βαθμολογία αδικίας* της για όλες τις μετρικές δικαιοσύνης. Όταν η βαθμολογία αδικίας είναι 0, η υποομάδα καταγράφεται απλά ως "Fair". Οι CSC αναπαραστάσεις που αντιστοιχούν στους ορισμούς δικαιοσύνης που κατέταξαν πρώτη την καθεμία από τις τρεις αυτές υποομάδες παρουσιάζονται στο Σχήμα 4.8.

Η Υποομάδα 1 κατατάσσεται υψηλότερα ως προς την αδικία με βάση τη μετρική *Equal Cost of Effectiveness* με $\phi = 0.7$, ωστόσο κατατάσσεται πολύ χαμηλότερα ή θεωρείται 'δίκαιη' σύμφωνα με άλλους ορισμούς. Αυτό το μοτίβο παρατηρείται και στην Υποομάδα 2, η οποία κατατάσσεται πρώτη βάσει του *Equal Choice for Recourse* με $\phi = 0.7$, και στην Υποομάδα 3, η οποία είναι πρώτη βάσει του *Equal Cost of Effectiveness* με $\phi = 0.3$, αλλά κατατάσσονται πολύ χαμηλότερα με άλλους ορισμούς. Αυτά τα αποτελέσματα αναδεικνύουν τη χρησιμότητα εφαρμογής και το λόγο ύπαρξης διαφορετικών ορισμών δικαιοσύνης, καθώς κάθε ορισμός αποτυπώνει διαφορετικές πτυχές της δυσκολίας στην επίτευξη δικαιοπραξίας.[5].

Η εφαρμογή αυτών των διαφορετικών πτυχών έχει και πρακτική αξία σε πραγματικές συνθήκες ελέγχου δικαιοσύνης. Για παράδειγμα, το *Equal Cost of Effectiveness* με $\phi = 0.7$ μπορεί να είναι κατάλληλο όταν απαιτείται ευρεία παρέμβαση υποστήριξης για κάποιο

[5] Περαιτέρω παραδείγματα και στατιστικά στοιχεία από ένα μεγαλύτερο δείγμα παρουσιάζονται στο Παράρτημα Α

τμήμα του πληθυσμού, αλλά οι διαθέσιμες ενέργειες είναι περιορισμένες. Σε αυτή την περίπτωση, η μάκρο-προοπτική της CSC της Υποομάδας 1 (το πάνω αποτέλεσμα στο Σχήμα 4.8) είναι ιδιαίτερα χρήσιμη, καθώς βοηθά στον εντοπισμό ομαδικών ενεργειών που μπορούν να προσφέρουν δικαιοπραξία σε σημαντικό ποσοστό της υποομάδας.

Αντίθετα, ο ορισμός *Equal Choice for Recourse* με $\phi = 0.3$, που αγνοεί το κόστος και κατατάσσει τις ομάδες μόνο βάσει της διαφοράς στον αριθμό αποτελεσματικών ενεργειών μεταξύ προστατευόμενων υποομάδων, μπορεί να είναι πιο κατάλληλος όταν η απόδοση κόστους στις ενέργειες είναι προβληματική ή και ηθικά αμφισβητίσιμη. Η μετρική αυτή αποτιμά τη μεροληψία με βάση τη διαθεσιμότητα αποτελεσματικών ενεργειών, όπως φαίνεται στη CSC της Υποομάδας 2 (στο μέσο του Σχήματος 4.8).

Μια ενδιαφέρουσα παρατήρηση αφορά την Υποομάδα 1, όπου εντοπίζεται μεροληψία κατά της υποομάδας *Male*. Αυτό έρχεται σε αντίθεση με προηγούμενες έρευνες που βασίζονται σε πιο στατιστικές προσεγγίσεις (π.χ. Tramer et al., 2017), οι οποίες συνήθως αναφέρουν μεροληψία κατά της υποομάδας *Female*. Το αποτέλεσμα αυτό είναι ίσως σημαντικό, καθώς αναδεικνύει το φαινόμενο του gerrymandering (Kearns et al., 2018), όπου ο χωρισμός σε υποομάδες γίνεται με τέτοιο τρόπο ώστε να καλύπτεται η μεροληψία. Το FACTS δείχνει ανθεκτικότητα σε αυτό το πρόβλημα, επιτρέποντας τον έλεγχο μικρών υποομάδων μέσω της ρύθμισης ενός ελάχιστου ορίου συχνότητας στην υπορουτίνα εύρεσης συχνών υποομάδων (με τον αλγόριθμο frequent itemset mining). Αυτό επιτρέπει τον εντοπισμό προκαταλήψεων που ενδεχομένως παραμένουν κρυφές.

Επίλογος

Σε αυτή τη διπλωματική εργασία μελετήσαμε την πρόκληση του υπολογισμού αποτελεσματικών και κατανοητών καθολικών αντιθετικών εξηγήσεων (GCE) καθώς και τα λεπτά ζητήματα που προκύπτουν για τη χρήση τους για την αξιολόγηση της δικαιοσύνης της δικαιοπραξίας σε συστήματα μαύρου κουτιού.

Αφενός, παρουσιάσαμε το GLANCE, το οποίο προσφέρει μια καινοτόμο προσέγγιση για τον υπολογισμό των GCE, επιτυγχάνοντας ισορροπία μεταξύ αποτελεσματικότητας, επεξηγησιμότητας και πρακτικότητας. Μέσα από προσεκτική συσταδοποίηση και στρατηγική δημιουργία ενεργειών, το GLANCE επιτρέπει στους χρήστες να κατανοούν το όρια απόφασης ενός μοντέλου και να εντοπίζουν πιθανούς τρόπους για την επίτευξη ευνοϊκών αποτελεσμάτων. Η αξιολόγησή του έναντι καθιερωμένων μεθόδων ανέδειξε την ικανότητα του GLANCE να δημιουργεί αξιόπιστες και υψηλής ποιότητας εξηγήσεις για διαφορετικά μοντέλα και σύνολα δεδομένων, με ανταγωνιστικούς δείκτες κυριαρχίας Pareto, που αναδεικνύουν την αποτελεσματικότητα και την αποδοτικότητά του.

Αφετέρου, χρησιμοποιήσαμε το FACTS για να διερευνήσουμε περαιτέρω τη δικαιοσύνη της δικαιοπραξίας μέσω των GCE, εξετάζοντας πόσο προσβάσιμες και δίκαιες είναι οι επιλογές δικαιοπραξίας για άτομα από διαφορετικές δημογραφικές ομάδες. Χάρη στη χρήση του FACTS, αποκτήσαμε βαθύτερη κατανόηση της δικαιοσύνης της δικαιοπραξίας σε συστήματα μαύρου κουτιού. Μέσω αυτής της προσπάθειας, αναδείξαμε περιορισμούς στις υπάρχουσες μετρικές δικαιοσύνης και υπογραμμίσαμε τη σημασία της αξιολόγησης των μεθόδων αντιθετικών εξηγήσεων, όχι μόνο με βάση το κόστος και την αποτελεσματικότητά τους, αλλά και ως προς τον πρακτικό αντίκτυπό τους στη δικαιοσύνη των προστατευόμενων ομάδων.

Μελλοντικές Κατευθύνσεις

Διάφορες κατευθύνσεις παραμένουν ανοιχτές για την περαιτέρω ανάπτυξη τόσο της επεξηγησιμότητας όσο και της δικαιοσύνης με βάση τις καθολικές αντιθετικές εξηγήσεις:

- **Βελτιστοποίηση Επιλογής Ενεργειών στο Τρίτο Στάδιο του AS**: Το τρίτο στάδιο του αλγορίθμου Augmented Space επιλέγει το τελικό σύνολο ενεργειών με μια απλή διαδικασία επιλογής. Μελλοντικά, αυτή η επιλογή θα μπορούσε να βελτιωθεί μέσω πιο εξελιγμένων τεχνικών βελτιστοποίησης, ενισχύοντας την πρακτικότητα και την οικονομία κόστους των ενεργειών που προτείνει ο αλγόριθμος.

- **Ερμηνεύσιμες Τεχνικές Συσταδοποίησης**: Το στάδιο συσταδοποίησης στον Augmented Space αλγόριθμο είναι κρίσιμο για την ταυτοποίηση ομάδων με παρόμοιες ανάγκες για την επίτευξη δικαιοπραξίας. Η ενσωμάτωση ερμηνεύσιμων μεθόδων συσταδοποίησης θα μπορούσε να ενισχύσει ακόμα περισσότερο τη διαφάνεια του AS, επιτρέποντας στους χρήστες και στους ελεγκτές να κατανοούν καλύτερα τα κριτήρια για την ομαδοποίηση, υποστηρίζοντας έτσι περαιτέρω την εμπιστοσύνη στη διαδικασία λήψης αποφάσεων του μοντέλου.

- **Επέκταση των Ορισμών Δικαιοσύνης της Δικαιοπραξίας στο FACTS**: Το FACTS έθεσε θεμελιώδεις ορισμούς για τη δικαιοσύνη στη δικαιοπραξία. Αυτοί οι ορισμοί χρήζουν περαιτέρω ανάπτυξης. Μια βαθύτερη εστίαση στις διαφορές ECD μεταξύ δημογραφικών υποομάδων θα μπορούσε να διασφαλίσει ότι οι μετρικές δικαιοσύνης αντανακλούν ακριβέστερα τις επιπτώσεις στην δίκαιη μεταχείριση εκείνων που έχουν ανάγκη δικαιοπραξίας.

- **Διερεύνηση Εναλλακτικών του Frequent Itemset Mining**: Αν και η εξαγωγή συχνών συνόλων (frequent itemset mining) έχει αποδειχθεί αποτελεσματική για τη δημιουργία ενεργειών στο FACTS, εναλλακτικές μέθοδοι μπορεί να προσφέρουν περισσότερο προσαρμοσμένες ή καλύτερα ενημερωμένες από το πλαίσιο ενέργειες. Η διερεύνηση αυτών των εναλλακτικών θα μπορούσε να οδηγήσει σε πιο λεπτομερείς και σχετικές με το πλαίσιο ενέργειες που ενισχύουν περαιτέρω τη δικαιοπραξία των χρηστών.

- **Ενσωμάτωση Συστηματικών Περιορισμών Εφικτότητας (Feasibility)**: Η διασφάλιση ότι οι αντιθετικές ενέργειες είναι εφικτές είναι ουσιώδης για την εφαρμογή τους στην πραγματική ζωή. Οι μελλοντικές εκδόσεις των GLANCE και FACTS θα μπορούσαν να ενσωματώσουν πιο συστηματικά κριτήρια εφικτότητας, βασισμένα, για παράδειγμα, σε causal constraints ή σε κανόνες ειδικούς για το εκάστοτε πεδίο, διασφαλίζοντας ότι οι προτεινόμενες ενέργειες είναι πρακτικές για εφαρμογή.

Συνοψίζοντας, η διπλωματική αυτή εργασία παρέχει μια ολοκληρωμένη μελέτη του GLANCE ως μιας εύκολα κλιμακούμενης και επεξηγήσιμης μεθόδου παραγωγής καθολικών αντιθετικών εξηγήσεων και μια ανάλυση του FACTS για την αξιολόγηση της δικαιοσύνης στη δικαιοπραξία σε πολύπλοκα μοντέλα μαύρου κουτιού. Μαζί, αυτά τα πλαίσια προσφέρουν πολύτιμες πληροφορίες για τις πρακτικές προκλήσεις που προκύπτουν στην παροχή δίκαιης, εφαρμόσιμης και επεξηγήσιμης δικαιοπραξίας στις αυτοματοποιημένες αποφάσεις. Βελτιώνοντας αυτές τις εργασίες, η μελλοντική έρευνα μπορεί να ενισχύσει τη δικαιοσύνη, την αξιοπιστία και τη διαφάνεια των συστημάτων ΤΝ σε εφαρμογές υψηλού κοινωνικού και οικονομικού ενδιαφέροντος.

Κείμενο στα αγγλικά

Chapter 1

Introduction

In recent years, machine learning (ML) systems have become integral to decision-making processes in critical areas such as finance (Boer et al., 2019), healthcare (Kyrimi et al., 2020), education (Loftus et al., 2018), and employment (Bogen and Rieke, 2018; Cohen et al., 2019). These systems are increasingly responsible for decisions that can have a significant impact on individuals' lives, such as loan approvals, hiring, or medical diagnoses. However, as the deployment of ML models in these sensitive domains continues to grow, so do concerns about the *fairness* and *transparency* of these systems.

Fairness Machine learning models are inherently susceptible to *bias*, as they often rely on historical data that may reflect existing societal prejudices. If left unchecked, these biases can lead to discriminatory outcomes that disproportionately disadvantage individuals or groups based on sensitive attributes such as race, gender, or age (Tashea, 2017; Raji and Buolamwini, 2019), and could potentially harm our society (Osoba and Welser IV, 2017). This makes the study of fairness in machine learning a critical area of research, focusing on ensuring that model outcomes do not systematically disadvantage *protected subgroups*.

Bias towards protected subgroups is most often detected by various notions of *fairness of prediction*, e.g., statistical parity, where all subgroups defined by a protected attribute should have the same probability of being assigned the positive (favorable) predicted class. These definitions capture the *explicit* bias reflected in the model's predictions. Nevertheless, an *implicit* form of bias is the difficulty for, or the *burden* (Sharma et al., 2020; Kuratomi et al., 2023) of, an individual (or a group thereof) to achieve *recourse*, i.e., perform the necessary *actions* to change their features so as to obtain the favorable outcome (Gupta et al., 2019; Ustun et al., 2019; Miller, 2019). Recourse provides explainability (i.e., a counterfactual explanation (Wachter et al., 2017)) and actionability to an affected individual, and is a legal necessity in various domains, e.g., the Equal Credit Opportunity Act mandates that an individual can demand to learn the reasons for a loan denial. *Fairness of recourse* captures the notion that the protected subgroups should bear equal burden (Gupta et al., 2019; Karimi et al., 2022; von Kügelgen et al., 2022; Kuratomi et al., 2023).

Transparency - Explainability At the same time, ML systems are often *black-box models* that provide little insight into their internal decision-making processes. This lack of transparency creates challenges for both end-users and stakeholders who need to understand and trust these decisions. This growing demand for model *explainability* (Burkart and Huber, 2021) has led to the emergence of *explainable AI (XAI)*, which aims to provide clear and interpretable explanations of how ML models arrive at their decisions.

In particular, in many applications of high impact (e.g., loan approvals) it is increasingly required for users to be able to understand how their features influence the outcomes of

the automated decision systems and how they might change them to achieve favorable outcomes. This is the same concept we discussed previously, namely the concept otherwise known as *recourse* (Miller, 2019). Counterfactual explanations have gathered extensive attention for their suitability in achieving algorithmic recourse (Karimi et al., 2020), their interpretability (Wachter et al., 2017), actionability (Ustun et al., 2019), utility in fairness audits (Sharma et al., 2020; Kavouras et al., 2023), etc. A counterfactual action, or simply an *action*, defines the specific feature changes needed to convert an unfavorable decision into a favorable one.

The dual challenges of fairness and explainability are deeply interconnected. In many real-world contexts, ensuring fairness also requires providing affected individuals with recourse—the ability to understand why a decision was made and what specific actions they can take to change an unfavorable outcome. This is not only a desirable quality but also a legal obligation in various domains (e.g. Equal Credit Opportunity Act).

1.1 Subjects of Study

This thesis explores these twin challenges of fairness and explainability through a detailed study of two key frameworks: *FACTS* (Kavouras et al., 2023) and *GLANCE* (Kavouras et al., 2024). Each framework addresses one of these core challenges, but they also offer complementary perspectives on using *counterfactual explanations* on *groups of instances* to evaluate and improve machine learning systems.

- *GLANCE* (Global Actions in a Nutshell for Counterfactual Explainability) addresses the challenge of *explainability* by generating *global counterfactual explanations*. It formulates the problem of finding explainable, actionable, and effective global counterfactuals that work across large portions of the input space. GLANCE builds on the idea that *similar instances can be explained by similar actions* and uses clustering techniques to produce concise sets of counterfactual actions. By segmenting the input space and generating representative actions for each segment, GLANCE enhances both the interpretability and practicality of the generated explanations.

- *FACTS* (Fairness-Aware Counterfactuals for Subgroups) reiterates, as in prior work, that fairness of recourse is an important and distinct notion of algorithmic fairness. However, it also demonstrates that this concept is more nuanced than previously explored in the literature. FACTS distinguishes between two conceptual viewpoints: the *micro viewpoint*, which examines individual-level costs when achieving recourse (the exclusive focus of prior research, as far as we can tell), and the novel *macro viewpoint*, which assesses how actions collectively impact groups of individuals. FACTS systematically explores the feature space to detect subspaces where recourse bias exists and quantifies this bias using *effectiveness-cost distributions* and a set of fairness definitions, each designed to capture a specific aspect of recourse bias. By ranking these subspaces according to the calculated bias, it highlights disparities in the costs and effectiveness of recourse actions across protected subgroups. The framework outputs *Comparative Subgroup Counterfactuals (CSCs)* to provide interpretable summaries of these findings.

Below, we will give a more extended introduction to how each framework approaches the corresponding problem and we will try to intuitively justify that approach.

1.1.1 Soft introduction to GLANCE / Augmented Space Algorithm

Counterfactual explanations have traditionally focused on providing *local explainability*, which tailors insights to specific negatively affected instances. However, in many practical applications, *global counterfactual explanations* are more valuable, as they offer broader explanations applicable on the entire affected population. While it is possible to cover all individuals by generating multiple local counterfactuals, this approach significantly reduces interpretability, a key aspect of global explainability.

Building on the work of studies like Rawal and Lakkaraju (2020); Kanamori et al. (2022), Kavouras et al. (2024) defines Global Counterfactual Explanations (GCE) as a small set of *global actions* designed to provide effective recourse to the affected population. A successful global counterfactual solution must satisfy three key objectives highlighted in prior research:

1. consist of a small number of actions to preserve interpretability (size),
2. minimize the cost associated with implementing these actions (cost),
3. provide recourse to as many affected individuals as possible (effectiveness).

As pointed out by Branke (2008), the relationships between multiple optimization objectives can be intricate, and reducing them to a single objective, as commonly done in practice (Rawal and Lakkaraju, 2020), may be problematic since these objectives are often non-commensurable. By framing GCEs as a multi-objective optimization problem, Kavouras et al. (2024) is able to explore the inherent *trade-offs* between effectiveness and cost, particularly when constrained by the number of actions in the solution set.

Global counterfactual explainability presents a unique optimization challenge compared to its local counterpart. Even if the optimal local actions are identified for each individual and a subset is chosen as global actions, the resulting Global Counterfactual Explanations (or GCEs) may still be suboptimal. For instance, the optimal global action may not be the best local action for any individual instance, but rather a compromise between all best local actions. Therefore, generating effective GCEs requires a more nuanced exploration of the action space.

The proposition of GLANCE The GLANCE framework (Kavouras et al., 2024; Emiris et al., 2024) introduces several approaches to address this complex problem, all of which share a key commonality: they rely on *clustering individuals*, utilizing a semi-standard procedure that accounts for both the feature space and the *action space*.

In this work, we focus on one specific approach within the GLANCE framework — the *Augmented Space* algorithm — for two primary reasons. First, it was the primary focus of the author's efforts during the development of GLANCE. Second, this particular algorithm was not further explored in the second GLANCE study (Kavouras et al., 2024), making it a promising candidate for further investigation.

Like other approaches within the GLANCE framework, the Augmented Space algorithm begins by performing a fine-grained clustering of individuals in the feature space. For each cluster, it identifies a set of diverse actions that can provide recourse to the cluster centroid. The underlying assumption is that *highly similar individuals* are likely to share a similar view of the decision boundary and, therefore, can achieve recourse through *similar, cost-efficient actions*. Figure 1.1a illustrates an example where four fine-grained clusters are

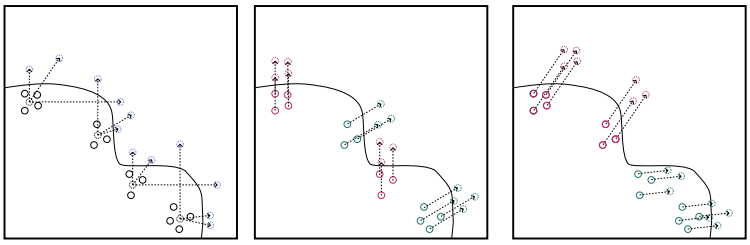


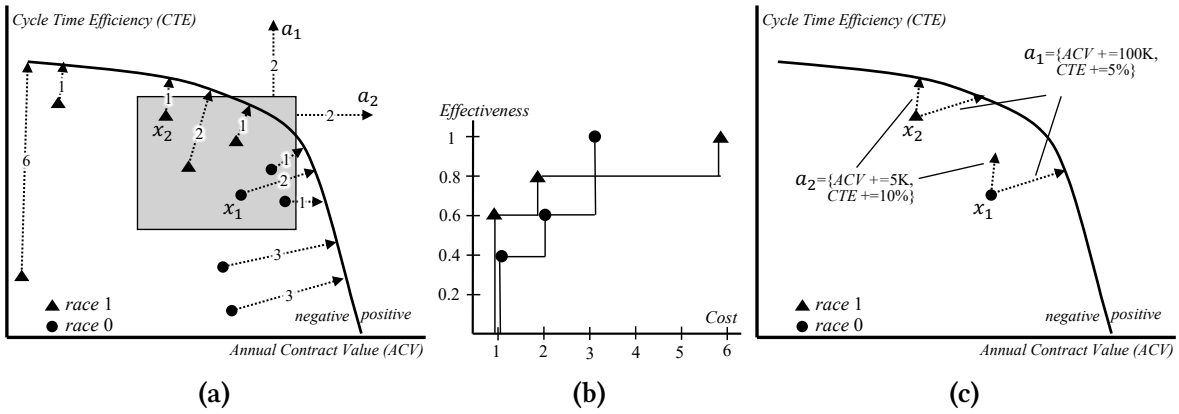
Figure 1.1: Illustrating of the intuition behind clustering methods (taken from Kavouras et al. (2024)). (a) GLANCE begins by generating diverse candidate actions from the centroids of clusters formed in the feature space. (b) Final clusters are formed based on similarity in both the feature space and the actions, grouping instances that may be distant in feature space but share similar actions. (c) In contrast, clustering solely by feature similarity can result in suboptimal global actions, either achieving high effectiveness at a high cost or low cost with minimal effectiveness.

each associated with three diverse actions derived from their centroids. These actions offer different pathways for providing recourse to the cluster members, giving GLANCE the flexibility to later choose the most appropriate action.

At the end of this initial phase, GLANCE associates each cluster of individuals with a set of potential actions. At this point, the various approaches within GLANCE diverge, and from here on, we will focus solely on the Augmented Space algorithm. In the second phase, the algorithm pairs each instance with the best possible action (in a simple, local-counterfactual-like sense) from its respective cluster's action pool. This effectively embeds the instances into a higher-dimensional space — the *augmented space* — which encodes both the features and the associated actions of the instances.

Next, the algorithm performs clustering on these augmented instances, forming t new clusters in the augmented space. From each cluster, one final action is selected based on a straightforward greedy procedure. The key innovation of the Augmented Space algorithm lies in this clustering within the augmented space of features and actions. The rationale for incorporating feature similarity is that it groups together similar individuals, in line with the logic used during the candidate generation phase. The motivation for incorporating actions, however, is that if similar actions provide recourse, it becomes easier to identify a shared low-cost, highly effective action, even when the individuals differ in terms of their features. Figure 1.1b shows an example where individuals are grouped based on both feature and action similarity.

As highlighted in Kanamori et al. (2022), relying solely on coarse-grained clustering in the feature space and assigning a single action to each cluster can result in suboptimal global actions, often compromising either cost or effectiveness. For example, in Figure 1.1c, individuals are clustered based on feature similarity, and a single action is chosen for the entire cluster. In the blue cluster, the actions are quite diverse, so selecting just one forces a trade-off between effectiveness (as illustrated) and cost. This underscores the value of clustering that takes into account both features and actions.



(a) (b) (c)

Figure 1.2: (taken from Kavouras et al. (2023))
(a) A visualization of affected individuals, the decision boundary, actions, and one subpopulation (highlighted in the shaded region) within the feature space
(b) The cumulative distribution of recourse costs for the individuals in (a)
(c) A comparison of two actions enabling recourse for two individuals.

1.1.2 Soft introduction to FACTS

As an introductory example of fairness of recourse (which we will see again in more detail in Section 2.4), consider a company's AI system used for promotion decisions, based on two features: CTE (cycle time efficiency) and ACV (annual contract value). Figure 1.2a shows employees in the negative predicted class, and additionally displays a supposed race attribute as circles and triangles. Arrows indicate the shortest actions for each employee to achieve recourse (in other words, local counterfactuals), and the numbers on the arrows represent their length, considered as the cost of the respective action. For example, x_1 may improve primarily by acquiring high-value projects (ACV), and the specific action displayed is assigned a cost equal to 2 (the arrow length).

The mean cost differs by race: 2 for race 0 and 2.2 for race 1, indicating recourse unfairness against race 1 (notice that prediction fairness in terms of statistical parity is satisfied). The authors of Kavouras et al. (2023), however, argue that fairness of recourse is more nuanced than the simple mean cost of recourse explored in previous literature (e.g. Gupta et al., 2019; Sharma et al., 2020; von Kügelgen et al., 2022; Kuratomi et al., 2023).

Firstly, the mean cost provides a limited perspective on individual recourse costs and may misjudge disparities. For example, in Figure 1.2a, race 1 has one outlier with a high cost of 6, skewing the average. To offer a clearer view, the cumulative distribution of costs, termed the *effectiveness-cost distribution* (ecd) in Kavouras et al. (2023), illustrates the trade-off between recourse and cost (Figure 1.2b). Based on the ecd, new definitions / metrics of recourse fairness are additionally formulated, that go beyond the mean cost. For example, if in a certain application actions over a certain cost, e.g., 2, are considered impractical, only 60% of race 0 can achieve recourse, compared to 80% of race 1, reversing the perceived fairness. This is referred to as *equal effectiveness within budget* in Kavouras et al. (2023) (we will explore this further in Section 4.1.1).

Moreover, even providing individual recourse prior to aggregation (as in Figure 1.2a) may not be meaningful in some practical scenarios. The *macro viewpoint* proposed by Kavouras et al. (2023) treats a group of individuals collectively, assessing actions that affect all members, such as societal or organizational interventions (e.g., tax relief or skill training). This is demonstrated in Figure 1.2a by the individuals in the shaded region, who

can be assigned a horizontal action such as a_1 or a_2. The action a_1 additionally demonstrates the merit of this viewpoint: while both race subgroups would have similar costs in the micro viewpoint, action a_1 disproportionately benefits race 0 over race 1. This perspective highlights systemic biases that individual-level assessments might overlook.

Finally, traditional recourse fairness depends on a cost function, which introduces practical limitations. Defining which attributes are actionable or ethical to change complicates the analysis (Rawal and Lakkaraju, 2020; Ustun et al., 2019). To address this, Kavouras et al. (2023) proposes *cost-oblivious* fairness definitions, such as *equal choice for recourse*, which considers the number of available actions rather than the cost alone. For instance, x_1 has only one option for recourse, while x_2 has two, indicating unfairness even if both have equal costs under action a_1.

Based on the above observations, Kavouras et al. (2023) also proposes FACTS (Fairness-Aware Counterfactuals for Subgroups), a flexible, model-agnostic framework designed to efficiently audit fairness in recourse. FACTS computes the aforementioned effectiveness-cost distribution, which captures the balance between cost and recourse across both micro and macro perspectives. It then efficiently audits recourse fairness by exploring the feature space and identifying subgroups with recourse bias, based on one of the several fairness definitions Kavouras et al. (2023) formulates (computed with the help of the ecd). FACTS ranks biased subspaces and provides interpretable summaries, offering a systematic way to examine fairness in both micro and macro perspectives.

In summary, Kavouras et al. (2023) makes significant contributions both conceptually and technically. On the conceptual side, it differentiates between the *micro viewpoint*, which focuses on individual costs, and the *macro viewpoint*, which assesses group-level impacts. New fairness of recourse notions, that extend beyond the limitations of mean cost, are also introduced. On the technical side, it offers the FACTS framework, which efficiently audits recourse fairness by exploring the feature space and identifying subgroups with recourse bias.

1.2 Key Contributions of the Thesis

This thesis explores these twin challenges of fairness and explainability through a detailed study of two key frameworks: *FACTS* (Kavouras et al., 2023) and *GLANCE* (Kavouras et al., 2024). Each framework addresses one of these core challenges, but they also offer complementary perspectives on using counterfactual explanations to evaluate and improve machine learning systems.

While each framework has a distinct focus — FACTS on fairness auditing and GLANCE on global counterfactuals — this thesis will also attempt to explore the interconnections between these two domains. By reviewing and replicating the methodologies of FACTS and GLANCE, conducting additional experiments, and examining the results in different contexts, this work seeks to highlight how fairness and explainability can complement each other in practical applications.

1.3 Structure of the Thesis

The remainder of this thesis is organized as follows:

- **Chapter 2: Counterfactual Explanations: Definitions, Methods, and Applications** provides an overview of counterfactual explanations, including key definitions and

methods for generating them. It introduces *local* and *global counterfactuals*, discusses prominent approaches such as DiCE (Mothilal et al., 2019), AReS (Rawal and Lakkaraju, 2020), and GLOBE-CE (Ley et al., 2023), and highlights their applications in fairness and explainability.

- **Chapter 3: Clustering Approach for Global Counterfactual Generation** explores the GLANCE framework in detail. It describes the algorithm's clustering-based approach to generating global counterfactuals, outlines its evaluation criteria, and presents the experimental results to demonstrate its effectiveness and interpretability.
- **Chapter 4: Auditing Fairness through Explainability Methods** explores the FACTS framework, which introduces a nuanced approach to auditing fairness of recourse by distinguishing between micro and macro viewpoints. This chapter explains how FACTS systematically identifies and quantifies disparities in recourse across subgroups using effectiveness-cost distributions and a range of fairness definitions. It also presents the results of experiments to validate the framework's effectiveness in detecting and explaining recourse bias through its output of Comparative Subgroup Counterfactuals (CSCs).
- **Chapter 5**: Conclusions summarizes the key findings of the thesis, discusses the broader implications for fairness and explainability in machine learning, and outlines potential directions for future research.

Chapter 2

Counterfactual Explanations: Definitions, Methods and Applications

2.1 Supervised Classification: Concepts, Notation, and Key Considerations

In this section, we introduce the key concepts, terminology, and notations related to *supervised classification* that will be used throughout this work. While the primary focus of this thesis is not on supervised classification itself, establishing a clear foundation in these concepts is essential for developing the appropriate intuition and ensuring consistency in notation.

Supervised classification is a core problem in machine learning, where the objective is to learn a function that maps input features to predefined labels using a dataset containing labeled examples. This involves training a model on a set of input-output pairs, allowing it to predict the labels of new, unseen data. For the purposes of this work, we formalize this setting as follows:

Let $\mathcal{X} = \prod_{i=1}^{d} \mathcal{X}_i$ denote a Cartesian product of $d \in \mathbb{Z}_+$ one-dimensional sets. Each set $\mathcal{X}_i$ typically represents a finite set of values or a subset of $\mathbb{R}$. Additionally, let $\mathcal{Y}$ denote a set of possible values, typically a finite set, representing the labels.

To introduce some nomenclature, each point $x \in \mathcal{X}$ is called an *instance* (or *individual* when referring to a person), and the i-th component of x is referred to as its i-th *feature*. This is also sometimes used for $\mathcal{X}$ as a whole, i.e. "the i-th feature" to refer to the i-th dimension of $\mathcal{X}$ (namely, $\mathcal{X}_i$). We also refer to the entire space $\mathcal{X}$ as the *feature space* and $\mathcal{Y}$ as the *label space*.

We assume that there exists an unknown function $f\colon \mathcal{X} \to \mathcal{Y}$, which represents a real-world process that assigns labels from $\mathcal{Y}$ to instances in $\mathcal{X}$. A machine learning practitioner is provided with a finite set of labeled instances, known as a *dataset*: $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $y_i = f(x_i)$. The task is to find a function $h\colon \mathcal{X} \to \mathcal{Y}$ that approximates the unknown function f as accurately as possible.

Example 2.1. *Consider a problem where the objective is to predict whether a loan applicant will default on their payment based on their income and age. Here, we have two* features*: income and age. The feature space can be defined as $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, where both $\mathcal{X}_1$ and $\mathcal{X}_2$ represent positive integers (assuming age and income are discrete values). The label space is given by $\mathcal{Y} = \{-1, 1\}$, where -1 signifies a default and 1 signifies no default.*

In this context, the machine learning task involves learning a function based on a dataset that includes the income and age values of a sample of individuals, along with their actual default status. The goal is to train a model capable of predicting whether an individual will default based on these features.

Key Considerations in Supervised Classification The above formulation captures the basic structure of supervised classification. Though not directly relevant to the subject of the current thesis, there are additional key considerations that are central to building effective models:

- *Classification Algorithms*: There are various algorithms designed for supervised classification, each with its own underlying principles and trade-offs. Some common approaches include logistic regression, decision trees, support vector machines (SVM), k-nearest neighbors (KNN), and neural networks. Each algorithm offers unique strengths and limitations, making it essential to choose the most suitable one based on the problem's characteristics.
- *Model Evaluation Metrics*: To evaluate a model's performance, several metrics are commonly used, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). These metrics help assess how well the model differentiates between classes and its robustness in handling imbalanced datasets.
- *Overfitting and Underfitting*: A key challenge in supervised classification is achieving a balance between model complexity and generalization. *Overfitting* occurs when a model learns irrelevant patterns from the training data, leading to poor performance on new data. In contrast, *underfitting* arises when a model is too simplistic to capture the underlying relationships within the data.

The concepts and considerations presented here provide the foundation for understanding how classification models work and how they will be used and evaluated in the subsequent sections of this thesis. By establishing this background, we ensure consistency and clarity in the discussions that follow.

2.2 Local Counterfactual Explanations

Counterfactual explanations play a crucial role in understanding causal relationships by answering 'what if' questions; 'If X had not occurred, Y would not have occurred' (Molnar, 2023). For example, 'If I hadn't slept in today, I would not have been late to work'. What this phrase is trying to convey is the causal relationship between X = 'I slept in' and Y = 'I was late to work'; X is the direct cause of Y, because if X had not occurred, given everything else is the same, then Y would not have occurred. Counterfactual thinking involves envisioning a different reality that contradicts observed facts, making it a powerful tool for human reasoning and decision-making.

In the context of machine learning, such counterfactual explanations can provide insight into individual predictions, by helping us understand what changes to the input features would alter the model's prediction. Conceptually, this approach treats the model's input features as causes and the prediction as the effect. For instance, if a model denies a loan application, a counterfactual explanation might reveal that an increase in annual income or a reduction in the number of credit cards could change the decision to approval.

Creating counterfactuals thus involves changing the input features of an instance to see how the prediction changes. The goal is to identify the smallest modifications that lead to a significant change in the prediction, such as flipping a classification result or reaching a specific probability threshold.

There are, however, serious challenges to be taken into consideration to produce counterfactual explanations. Good counterfactual explanations are characterized by several key criteria, some more straightforward than others. Firstly, as expected, they should achieve a relevant change in prediction as closely as possible to the predefined target. Secondly, the counterfactual should be minimally different from the original instance, both in terms of the number of changed features and the magnitude of those changes. Additionally, providing multiple diverse counterfactuals can offer decision-makers various actionable paths to achieve the desired outcome. Lastly, to be suitable for practical applicability, the counterfactual instances should be realistic and likely within the context of the data's joint distribution, and the change from the factual to the counterfactual instance should also be actionable (e.g. cannot involve changing your year of birth).

As a final note, we shall see in the following sections that there are also counterfactual explanations which consider many individuals at once (these are the core subjects of study for this work). To distinguish between the two notions, the counterfactual explanations we have talked about thus far will also be called *local counterfactuals* from now on.

2.2.1 Definition

Based on the above discussion, the end goal would be to develop algorithms that find good counterfactuals. The first step towards that goal is to formally define the thing that we are looking for, i.e. what is a local counterfactual.

As we said, counterfactuals concern the question: given an instance x, what is the closest instance one can find for which the model's prediction changes? Now, depending on the details of the setting, the exact definitions of "closest instance" and "prediction changes" vary, and may not be straightforward, as we discussed.

Nonetheless, in the simplest case, e.g. for classification where we want to see when the result is flipped, it is easy to give a straightforward formal definition. Consider the following formulation.

Definition 2.1 (Local Counterfactual - Simple). *Assume a model h (as described earlier) that maps instances $x \in \mathcal{X}$ to labels $y \in \mathcal{Y}$. Given a point $x_0 \in \mathcal{X}$ for which $h(x_0) = y_0$ and a desired label $y_T \in \mathcal{Y}$, a counterfactual explanation for x_0 with respect to h is any solution to the following optimization problem:*

$$\begin{aligned} &\textit{Minimize} && d(x_0, x) \\ &\textit{subject to} && h(x) = y_T \\ & && x \in \mathcal{X} \end{aligned} \tag{2.1}$$

This definition conveys the core idea of local counterfactuals: *we want the minimal changes that make the model h give the desired output.* As important as the other complications we have discussed are indeed, we believe it is also important to isolate the central idea behind the concept of local counterfactuals and formulate it as Definition 2.1 does; then we can add more complexity on top of it.

To do that, there is a plethora of approaches. As an example, the seminal work of Wachter et al. (2017) posits that the constraint $h(x) = y_T$ can be relaxed for regression problems or for classification where we get the predicted probabilities; it suffices that $h(x)$ and y_T are as close as possible. As such, they formulate the search for local counterfactuals

as:

$$\begin{aligned} \text{Minimize} \quad & L(x_0, x, y_T, \lambda) = \lambda(h(x) - y_t)^2 + d(x_0, x) \\ \text{subject to} \quad & x \in \mathcal{X} \end{aligned} \tag{2.2}$$

(where d is a normalized variation of the l_1 norm)

To elaborate, they now have 2 objectives (distance of x to original instance and distance of $h(x)$ to the target), which they incorporate in a single objective function using the - relatively standard - linearization approach, i.e. the objective function is the weighted sum of the 2 objectives, with the weight being a tunable hyperparameter (we should mention that Wachter et al. (2017) also suggests a method to tune λ).

As one can perhaps imagine, all sorts of objectives, requirements, or constraints can be incorporated in the formulation of the problem in a manner similar to 2.2. Recall all the additional constraints we discussed briefly in the opening of the section: feasibility, plausibility, actionability, and diversity. All of these are arguably necessary requirements for good counterfactuals; perhaps there are even more, depending on who you ask.

In the most general case, what we want to do is a 'good' compromise between achieving the desired label, remaining close to the original point x, and satisfying any other constraints that concern the interpretability, actionability, etc. of the explanation. Drawing inspiration from the respective definition in Dandl et al. (2020), we give the following general definition.

Definition 2.2 (Local Counterfactual - Complex). *Assume a model h (as described earlier) that maps instances $x \in \mathcal{X}$ to labels $y \in \mathcal{Y}$. Given a point $x_0 \in \mathcal{X}$ for which $h(x_0) = y_0$ and a desired label $y_T \in \mathcal{Y}$, a counterfactual explanation for x_0 with respect to h is any solution to the following multi-objective optimization problem:*

$$\begin{aligned} \textit{Minimize} \quad & (L(y_T, y), d(x_0, x), c_1(x), \ldots, c_k(x)) \\ \textit{subject to} \quad & x \in \mathcal{X} \\ & h(x) = y \end{aligned} \tag{2.3}$$

where:

- the function L should quantify the discrepancy between the desired label and the label of the candidate counterfactual (could be as simple as $\mathbb{1}[y_T, y]$).
- The function d should quantify the distance between the original instance and the counterfactual (could be as simple as an l_r-norm distance).
- $c_1, \ldots, c_k$ signify all extra requirements one could think of imposing on the resulting counterfactual explanations. Note that they don't need to depend only on x, they could involve any one of the other parameters of the problem, e.g. the original instance x_0. We just omitted this for simplicity. Some examples are:
  - l_0 norm: $c(x) = \|x_0 - x\|_0$ (used in Dandl et al. (2020)). For when we consider it important for the resulting counterfactual not only to have small changes in the features, but also for the number of features that get changed to be small. Imagine, for example, that a counterfactual for a hiring model/algorithm suggests that a user change all information on their LinkedIn profile. No matter how small each individual change is, something like this certainly makes it more difficult for the user to act on it.

– $c(x) =$ distance of x to the nearest neighbor from a given dataset of points (used in Dandl et al. (2020)). This can be thought of as a means to make the resulting counterfactual more likely, according to the distribution of the real data.

In contrast to Problem 2.1 (which is by no means easy), the above problem is even more challenging: after all, in Problem 2.3 we have actually done nothing more than write down all demands we would have from a 'perfect' counterfactual. The objectives we have incorporated will be contradicting for most cases (e.g. the lower the cost, the greater we would expect L to be) and the formulation gives no hint as to resolving this ambiguity. So, it is expected that the problem (2.3) will not admit a clear solution for most cases.

One immediate challenge is how we are supposed to decide between 'equally satisfying' solutions (i.e. pareto optimal[1]). Do we want to come up with only one counterfactual in the end? Or would we prefer to give a selection of different but equally 'effective' counterfactuals to the user, and thus also satisfy the *diversity* requirement? After all, these are concerns that arise for every multi-objective optimization problem.

Additionally, there are inherent challenges associated with the problem of finding 'optimal' counterfactuals. One key issue is the nature of the assumptions we make about the model. In this work, as in much of the related literature (e.g., see the method roundup in the survey by Karimi et al. (2022)), we focus on a model-agnostic setting, where we have no direct access to the inner workings of the model h. Instead, we only have "oracle access" — we can query the model with any point $x \in \mathcal{X}$ (whether from the data or artificially generated) and receive the corresponding output $h(x)$. This constraint makes heuristic search through the feature space unavoidable. However, heuristic searches cannot guarantee an optimal solution to the problem (2.3), particularly when we make no assumptions about the model's regularity. Given that machine learning models can be arbitrarily complex, with highly non-linear and intricate decision boundaries, it becomes clear that finding or even adequately approximating an optimal solution to the problem is a difficult and computationally intensive task.

Another significant challenge arises when choosing an appropriate distance function d. Ideally, this function should not simply capture Euclidean distance but should reflect the practical difficulty an individual would encounter when changing their feature values from x_1 to x_2. In other words, it should represent a *cost* function rather than a simple distance metric. However, defining such a cost function is far from straightforward, as the complexity involved is significant due to numerous factors. The difficulty of making changes is highly specific to the dataset and the domain, and can be extremely non-linear or irregular in many cases. For example, relocating might be significantly easier for a person in their 20s than for someone in their 80s. Similarly, a change in age could be entirely infeasible in certain contexts (e.g., qualifying for a one-time scholarship), or feasible only within certain constraints (e.g., a scholarship limited to a specific age range). These examples highlight how constructing a meaningful cost function that aligns with real-world scenarios is such a complex task that it could easily form an entire research focus on its own.

Finally, from a slightly higher point of view, even the somewhat general form of the problem 2.2 may fail to model the problem of finding good local counterfactuals for some cases. To include everything as an extra objective function may not be enough for practitioners of certain applications, who would prefer to include some extra constraints instead.

We still believe it serves as a satisfactory general definition of local counterfactuals. What we want to illustrate here is the difficulty of formulating local counterfactuals, and

[1] the notion of pareto optimality is well-known in the multiobjective optimization literature, see for example Branke (2008); Emmerich and Deutz (2018)

that this is not so much due to technical reasons as due to the complexity of the requirements needed to be satisfied by a good counterfactual.

2.2.2 Overview of Approaches

In recent years, a wealth of methods, frameworks, and algorithms have emerged for generating local counterfactuals (e.g. Wachter et al., 2017; Mothilal et al., 2019; Poyiadzi et al., 2020; Lucic et al., 2021; Laugel et al., 2017), each differing in approach and focus.

A common strategy for generating local counterfactuals is to frame the task as a single-objective optimization problem, aiming to minimize the distance between a query point and its counterfactual. This approach was notably advanced by Wachter et al. (2017), a seminal work that set a foundation for the field. Specifically, they propose minimizing the following loss function:

$$L(x_0, x, y_T, \lambda) = \lambda(h(x) - y_T)^2 + d(x_0, x) \tag{2.4}$$

where y_T is the target label for the counterfactual in a regression setting, and $d(x_0, x)$ represents the distance between the query point x_0 and the counterfactual x. To define d, they use the Manhattan distance, weighted by the inverse median absolute deviation (MAD) of each feature:

$$d(x_0, x) = \sum_{i=1}^{d} \frac{|x_{0i} - x_i|}{MAD_i} \tag{2.5}$$

The MAD value serves as a robust metric similar to standard deviation but based on the median and absolute deviations, making it less sensitive to outliers. For further details on MAD, refer to (Molnar, 2023, Section 9.3).

In Wachter et al. (2017), an additional parameter, λ, balances the conflicting goals of minimizing both the error relative to the target label and the distance between the factual and counterfactual instances. These two objectives can conflict, as minimizing the label error often requires greater changes in feature values, thereby increasing the distance from the original instance. The λ value, which reflects the relative importance of each objective, can be specified by the user or tuned through an outer loop that iteratively adjusts λ. Instead of fixing λ, this loop allows the user to specify a tolerance for the predicted label's proximity to the true label, and reruns the optimization across multiple λ values until this tolerance is met. While this approach adds flexibility, it also increases computational demands due to repeated optimization runs and still requires the user to determine an appropriate tolerance level.

The authors effectively address the ambiguity inherent in multi-objective problems like (2.3) by converting the two conflicting goals into a single objective via a weighted sum, a standard approach in multi-objective optimization (Emmerich and Deutz, 2018; Branke, 2008). This formulation allows the optimization to prioritize one objective over the other through λ, balancing actionability with recourse efficacy.

2.2.3 DiCE

We focus particularly on the DiCE method (Mothilal et al., 2019), as it is our chosen tool for generating local counterfactual explanations. DiCE emphasizes two essential properties for effective counterfactual explanations: *feasibility* and *diversity*. Feasibility ensures that counterfactuals are actionable and contextually plausible for individuals, while diversity provides multiple distinct options for achieving the desired outcome, giving users a

range of choices to consider. To formalize these properties, DiCE constructs an optimization framework that not only minimizes the distance to the counterfactual (feasibility) but also maximizes diversity among the generated examples.

The primary contribution of Mothilal et al. (2019) lies in an algorithm that leverages *determinantal point processes* (DPPs) to generate diverse counterfactuals efficiently. The use of DPPs allows DiCE to produce sets of counterfactuals that are highly distinct from one another, avoiding the common pitfall where generated examples cluster too closely in the feature space, thereby limiting their practical usefulness.

In our work, we leverage DiCE's *random* generation method, which operates by exploring the counterfactual action space through a process of targeted random sampling. This approach systematically minimizes the distance between the original instance and each counterfactual while ensuring a diverse set of counterfactuals. Unlike a purely stochastic search, DiCE's random approach incorporates specific constraints that prioritize proximity to the original instance, balancing the need for both feasibility and minimal feature change.

The random method within DiCE, while simpler than other options such as those based on determinantal point processes (DPPs), is effective for applications where interpretability is central, and the computational cost of generating counterfactuals must remain manageable. This balance makes it a suitable choice for our implementation, where the focus is on providing actionable, locally optimal counterfactuals while maintaining computational efficiency.

2.2.4 Other Methods

Although we have focused on DiCE for our work, we also provide a brief overview of some other prominent approaches, for completeness.

FACE The FACE method (Poyiadzi et al., 2020) addresses limitations in typical counterfactual explanations, especially regarding feasibility and data coherence. FACE argues that traditional methods often yield counterfactuals that may not be achievable or realistic, as they may propose changes that are either contextually inappropriate or difficult for individuals to act upon (e.g., suggesting an applicant double their salary without increasing skills). FACE seeks to ensure that counterfactuals not only belong to dense regions of the data distribution (indicating plausible outcomes) but also provide feasible paths of change. It accomplishes this by utilizing a density-weighted distance metric, which allows counterfactuals to be both feasible and representative of achievable transformations.

FOCUS FOCUS (Lucic et al., 2021) introduces an approach for generating counterfactual explanations that extends to non-differentiable models, particularly tree-based ensembles. Several counterfactual methods assume model differentiability, but FOCUS overcomes this by approximating tree models using smooth probabilistic functions, enabling gradient-based optimization for generating counterfactuals. This method is particularly advantageous for scenarios where non-differentiable models are prevalent, as it allows FOCUS to generate counterfactuals closer to the original instance than existing heuristic methods, while supporting a broader range of model types and maintaining actionability.

Growing Spheres Growing Spheres (Laugel et al., 2017) offers a unique approach to counterfactual generation, particularly suited to settings where no prior knowledge of the model or data is available. Growing Spheres is a data- and model-agnostic method that locates the

decision boundary by incrementally expanding a "sphere" around a given instance until it reaches an area classified differently. This approach minimizes the perturbations required to achieve the desired classification change, ensuring that the counterfactual lies as close as possible to the original instance. Growing Spheres is thus effective for generating actionable counterfactuals even when the model and data are completely opaque to the user.

2.3 Global Counterfactuals

A more recent direction of research focuses on "global counterfactuals", which extend the concept of traditional counterfactual explanations to cover multiple instances simultaneously. Instead of explaining a single instance, global counterfactuals aim to provide explanations that apply to entire groups of instances, specific subregions of the feature space, or even the whole population, offering broader insights into the model's behavior.

For example, in the case of an automatic loan approval system, local counterfactual generators can suggest to a single applicant what change they could perform to get approved. A global counterfactual would be, 'If one has worked for less than 10 years, raising their salary by 50000 will get them approved, otherwise raising their salary by 30000 will get them approved'. This provides both actionable recourse and an insight into how the model makes its decisions.

2.3.1 Preliminaries

We begin with the definition of some fundamental concepts, not only to our work but already in the existing literature on *global counterfactual explanations* for *binary classification problems* (e.g. Rawal and Lakkaraju, 2020; Ley et al., 2023; Kanamori et al., 2022)[2].

We consider again a standard binary classifier $h : \mathcal{X} \to \{-1, 1\}$, where the positive outcome is the favorable one. For any given h, and always assuming a dataset $\mathcal{D}$, we focus on $\mathcal{D}_{aff} = \{x_i : (x_i, y_i) \in \mathcal{D} \text{ and } h(x_i) = -1\}$, i.e. the **adversely affected individuals**; those who receive the unfavorable outcome.

We denote as $\mathcal{A}$ the set of all possible actions, where an **action** $a \in \mathcal{A}$ is a set of changes to feature values, e.g., $a = \{\textit{country} \to \textit{US}, \textit{education-num} \to +2\}$, which, when applied to an instance x, results in a **counterfactual** instance $x' = a(x)$. For the previous example, if $x = (\textit{Canada}, 7)$, then $a(x) = (\textit{US}, 9)$.

Every action a has a cost, denoted as $\mathsf{cost}(a, x)$, and is effective for an instance x if $h(a(x)) = h(x') = 1$. The recourse cost $\mathsf{rc}(A, x)$ of an instance x is the minimum cost incurred from an effective action in $A \subseteq \mathcal{A}$:

$$\mathsf{rc}(A, x) = \min\{\mathsf{cost}(a, x) | a \in A : h(a(x)) = 1\}$$

Let $X = \{x \in \mathcal{D}_{aff} | h(a(x)) = 1, a \in A\}$ be the set of instances that flip their prediction by using one of the actions in A, then the effectiveness of the actions for the affected instances is defined as:

$$\mathsf{eff}(A, \mathcal{D}_{aff}) = \frac{|X|}{|\mathcal{D}_{aff}|},$$

[2] The concepts defined in prior works may have different names or slightly different technicalities in their definitions, but the core logic and intuition is the same.

and their cost as:

$$\mathsf{avc}(A, \mathcal{D}_{aff}) = \frac{\sum_{x \in X} \mathsf{rc}(A, x)}{|X|}.$$

Finally, we use $\mathsf{size}(A)$ to refer to the cardinality of a set A.

2.3.2 AReS

One of the earliest works on global counterfactuals and seminal in the field was the work of Rawal and Lakkaraju (2020). Their proposed approach for finding interpretable global counterfactual explanations, namely the Actionable Recourse Summaries (AReS) framework, was to formulate an optimization problem based on the requirements they want their counterfactuals to satisfy, and then use known algorithms to solve this problem.

A key innovation in AReS is its objective function, which balances three critical aspects: the accuracy of recourse, interpretability of explanations, and minimizing recourse costs for the population. The framework optimizes these elements by learning a small set of interpretable rule sets that define recourse actions for specific subgroups, allowing decision-makers to answer important questions such as how recourse differs across demographic groups.

Problem Formulation In slightly more concrete terms, Rawal and Lakkaraju (2020) seeks to identify a set of actionable rules of the form "if $f_1 \lessgtr v_1$ and $f_2 \lessgtr v_2$, then change f_1 to v_1' and f_2 to v_2'," where f_1, f_2 are features, and v_1, v_2, v_1', v_2' are feature values. The specific *problem* addressed is to find a set of such rules that maximizes the following objective function:

$$\lambda_1 \cdot \text{Cov} + \lambda_2 \cdot \text{Eff} - \lambda_3 \cdot \text{Cost}_1 - \lambda_4 \cdot \text{Cost}_2,$$

where Cov and Eff represent the coverage of the rule set (i.e., the proportion of individuals for whom there is a rule) and the effectiveness of those rules in providing successful recourse, respectively. The terms Cost_1 and Cost_2 quantify different aspects of the costs associated with applying the rules for an individual, or group thereof (exact details are out of scope for our work, but the interested reader can of course find out more in Rawal and Lakkaraju (2020)).

The optimization problem is also subject to certain *constraints* that limit the maximum width of individual rules (i.e., the number of conditions within a rule) and the total number of rules in the set. This set of constraints aims to capture the *explainability* requirements, since having a very large rule set as output would not be interpretable by a human being.

The formulation uses a linear combination of these factors due to the inherent trade-offs between coverage, effectiveness, and cost, which makes maximizing these quantities simultaneously infeasible. The weights $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ in the objective function, as well as the constraints on the maximum width of each rule and the total number of rules, are user-provided parameters. It should also be noted that the appropriate values for these quantities are very likely to vary based on the dataset and domain in question.

In Figure 2.1 we can see an example of the method's output. The metrics (effectiveness, cost) are not displayed, but we can understand the general form of the output and why the authors claim that it is interpretable. Indeed, the way the population has been split into groups based on values of the features and for each cohort a straightforward change is suggested could most probably be understood by anyone.

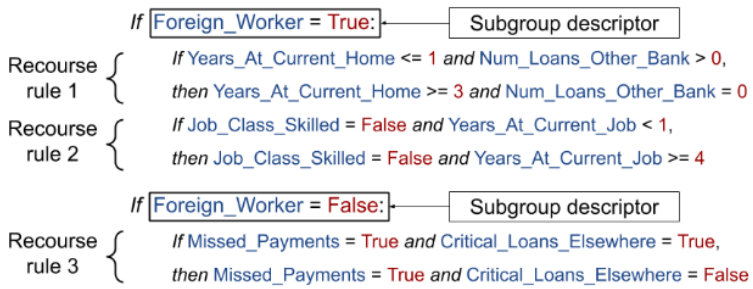


Figure 2.1: Example output of the AReS framework. Taken from Rawal and Lakkaraju (2020).

	Algorithms	Datasets					
		COMPAS		Credit		Bail	
		Recourse Acc	Mean FCost	Recourse Acc	Mean FCost	Recourse Acc	Mean FCost
DNN	AR-LIME	**99.40%**	3.42	10.26 %	5.08	92.23 %	2.90
	AR-KMeans	57.76 %	6.39	48.72 %	2.50	87.98 %	7.50
	FACE	73.71%	4.48	50.32%	2.48	91.37%	8.43
	AReS	99.37%	**2.83**	**78.23%**	**2.23**	**96.81%**	**2.45**
RF	AR-LIME	65.88 %	6.34	26.33 %	3.32	90.46 %	2.87
	AR-KMeans	60.48 %	5.31	16.00 %	4.02	92.03 %	7.02
	FACE	62.38%	4.48	31.32%	1.35	92.37%	9.31
	AReS	**72.43%**	**2.52**	**39.87%**	**1.09**	**97.11%**	**1.78**
LR	AR	**100%**	5.41	**100%**	1.69	**100%**	8.07
	FACE	98.22%	6.12	95.31%	2.08	94.37%	7.35
	AReS	99.53%	**4.02**	99.61%	**1.28**	**100%**	**6.45**

Figure 2.2: Results of AReS on the COMPAS, Credit and Bail datasets. Taken from Rawal and Lakkaraju (2020). Effectiveness (called recourse correctness or accuracy in Rawal and Lakkaraju (2020)) and mean feature cost are shown. AR-LIME, AR-KMeans and FACE are local counterfactual methods with which Rawal and Lakkaraju (2020) compared itself due to lack of more relevant research at that time. They are out of scope for our work.

Additionally, in Figure 2.2 we show one part of the results presented in Rawal and Lakkaraju (2020). The cost is calculated in a different way than in the works of Kavouras et al. (2023) and Kavouras et al. (2024), so conclusions are difficult to draw from these values. Nonetheless, from the effectiveness values we can make some comments. The results are very satisfactory in 6 out of the 9 cases, with AReS demonstrating an effectiveness of 96% and above. However,in the other three cases it achieves only 72.43%, 78.23%, and 39.87%. As we shall see later in this work, these are substantially low values compared to the more recent methods such as Ley et al. (2023) and Kavouras et al. (2024).

Incidentally, we should mention that Rawal and Lakkaraju (2020) did not have publicly available source code for their experiments. For this reason, we built our own implementation from scratch. At around the same time, Ley et al. (2022), having developed a faster version of the AReS framework named *Fast AReS*, also made their implementation publicly available at `https://github.com/danwley/GLOBE-CE/`. We used both implementations for the evaluation of the framework.

Overall, the AReS method demonstrates several notable strengths. It is grounded in solid theoretical principles that justify its approach, particularly in terms of balancing effectiveness, cost, and interpretability. The rules it generates are highly interpretable, which is crucial for transparency in decision-making processes. Additionally, the method produces results that are generally effective, making it a useful tool in actionable recourse.

However, there are also some notable limitations. First, the optimization procedure described in Rawal and Lakkaraju (2020) tends to be quite slow. This was evident both in our

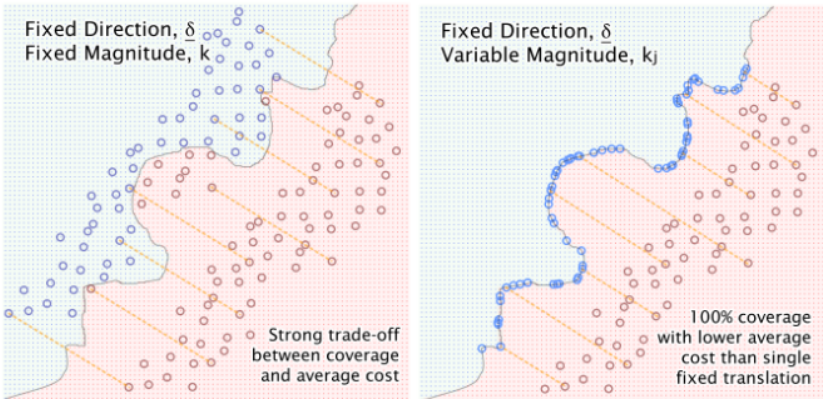


Figure 2.3: Illustration of actions on the same direction but with different magnitudes; taken from Ley et al. (2023).

own implementation and in the one provided by Ley et al. (2022). The slow runtime can pose significant challenges, especially when dealing with larger datasets or more complex models. Furthermore, the results displayed considerable variability, as we can see from Figure 2.2. While the effectiveness is very high—sometimes near 100%—in many cases, it drops to much lower levels in others. Given that effectiveness is a crucial measure for counterfactuals, this inconsistency can undermine their meaningfulness if a reasonably high level of success is not achieved.

Finally, another important limitation is the lack itself of publicly available source code for the experiments conducted in Rawal and Lakkaraju (2020). This absence made it more difficult to accurately reproduce their results, hindering efforts to thoroughly study, extend, or improve upon their work.

2.3.3 GLOBE-CE

Another significant contribution comes from Ley et al. (2023), who introduced GLOBE-CE, a novel framework to generate reliable and scalable global counterfactual explanations (GCEs). GLOBE-CE provides global directions along which groups of input data can be perturbed to alter their predictions. Unlike prior methods, GLOBE-CE relaxes the traditional approach by allowing counterfactual translations to have fixed directions but varying magnitudes across different inputs. This flexibility helps balance the trade-off between effectiveness (the percentage of inputs that can be altered) and cost (the difficulty of implementing the changes), a significant limitation in prior global counterfactual methods. An illustration of this concept is shown in Figure 2.3.

One of the key innovations of GLOBE-CE is its ability to handle categorical features by mathematically formulating translations for one-hot encoded data using interpretable If/Then rules. This ensures that GLOBE-CE can work effectively across datasets with both continuous and categorical features, providing interpretable and actionable counterfactuals. Additionally, GLOBE-CE is computationally efficient, outperforming existing methods in terms of speed, while also achieving better effectiveness and lower recourse costs.

Problem Formulation Ley et al. (2023) does not explicitly define a formal optimization problem, but the framework aims to achieve a balance between maximizing effectiveness (referred to as *coverage* in Ley et al. (2023)) and minimizing the average cost of recourse for individuals. Based on the description in the original work, the core task can be understood

Models	Algorithms	Datasets											
		COMPAS			German Credit			Default Credit			HELOC		
		Cov.	*Cost*	*Time*	*Cov.*	*Cost*	*Time*	*Cov.*	*Cost*	*Time*	*Cov.*	*Cost*	*Time*
DNN	AReS	51%	2.31	101s	73%	1.6	2712s	7.22%	1.0	7984s	5.4%	1.0	9999s
	Fast AReS	64%	**1.45**	32.0s	72%	1.43	12.8s	99.8%	4.2	37.3s	52%	5.5	109.1s
	GLOBE-CE	66%	1.53	**7.08s**	85%	1.2	**2.28s**	98.5%	1.3	**3.6s**	93%	4.3	**4.66s**
	dGLOBE-CE	**70%**	1.46	9.15s	**90%**	**1.1**	2.63s	**100%**	**1.1**	7.86s	**95%**	3.8	5.46s
XGB	AReS	45%	1.9	205s	61%	1.5	2092s	11%	1.0	9999s	1.7%	1.0	9999s
	Fast AReS	83%	1.9	47.6s	65%	1.75	34.33s	93%	2.3	29.97s	28%	**2.1**	93.58s
	GLOBE-CE	78%	1.8	**9.61s**	**95%**	**1.02**	**5.04s**	96%	1.1	**2.94s**	58%	2.4	**4.7s**
	dGLOBE-CE	**91%**	**1.4**	12.4s	83%	1.03	5.95s	**100%**	**0.7**	6.35s	**80%**	2.4	5.6s
LR	AReS	79%	1.5	506s	85%	1.3	3566s	31%	1.2	9999s	4.8%	1.0	9999s
	Fast AReS	82%	1.7	43.0s	85%	1.3	9.3s	99%	2.1	17.82s	92%	1.6	127.3s
	GLOBE-CE	83%	1.20	**8.43s**	82%	**1.2**	**3.39s**	**100%**	**1.0**	**3.42s**	**100%**	**0.5**	**3.11s**
	dGLOBE-CE	**84%**	**1.18**	11.7s	**91%**	1.3	3.87s	**100%**	**1.0**	7.21s	**100%**	**0.5**	3.85s

Figure 2.4: Results of GLOBE-CE on the COMPAS, German Credit, Default Credit and HELOC datasets. Taken from Ley et al. (2023). Effectiveness (called coverage in Ley et al. (2023)) and mean cost are shown. Fast AReS is an improved implementation of AReS by Ley et al. (2022) (out of scope for our work). Highlighted in red are cases with effectiveness less than 10% or more than 3 hours runtime. Best metrics are shown in **bold**.

as identifying a single direction in the feature space that not only provides recourse for the largest possible number of individuals but also minimizes the average cost associated with moving them along this direction. What distinguishes GLOBE-CE from other approaches is that it relaxes the constraint of having a fixed action for all individuals by allowing each person to be assigned a different magnitude along the identified direction.

This approach offers computational efficiency and typically achieves high effectiveness at relatively low costs. However, the flexibility in magnitudes can pose challenges in terms of interpretability. In cases where the number of distinct magnitudes assigned to individuals becomes too large, the resulting global counterfactual explanations may resemble a collection of local counterfactuals rather than presenting a unified global action applicable across the population.

Some indicative results from Ley et al. (2023) are shown in Figure 2.4. GLOBE-CE compares itself to AReS. In this comparison, GLOBE-CE is clearly superior, since it always achieves greater effectiveness (coverage) and lower cost; with a single exception in the COMPAS dataset and DNN model, where the cost is slightly higher than AReS, but the effectiveness is also higher.

Other than this comparison, it can be observed that GLOBE-CE in general has quite satisfactory performance. The interpretation of costs, of course, depends highly on the choice of cost function, so no immediate conclusions can be drawn from that other than comparisons (especially since the cost function used by Ley et al. (2023) is a weighted version of the l_1 norm with weights depending on the features' ranges). Effectiveness, however, is more a more objective measure. As we can see, GLOBE-CE achieves relatively high values, ranging from 58% to 100%, with many cases having a perfect effectiveness of 100%.

On the other hand, GLOBE-CE also has some drawbacks. Most notably, the fact that only the direction is fixed but not the magnitude has some implications for the globality and / or interpretability of the method. In many cases, as described in Kavouras et al. (2024), the number of discrete actions needed by GLOBE-CE can even be several hundreds. These action sets are not suitable for inspection and interpretation by humans. Interestingly, one could also argue that in such cases the actions GLOBE-CE discovers have a slight resemblance to local counterfactuals, since each instance may be assigned a slightly different action; albeit in the same direction.

Other minor drawbacks include the necessity to perform one-hot encoding for categor-

ical features, which sometimes requires a lot of memory, and the lack of a comprehensive interpretable output. There exist, of course, the If/Then rules we mentioned, but these refer only to categorical features.

2.3.4 CET

Counterfactual Explanation Tree (CET), proposed by Kanamori et al. (2022), uses decision trees to assign consistent and transparent counterfactual actions to groups of individuals. CET claims two core advantages: (1) Transparency — it provides an interpretable tree structure to summarize the reasons behind assigned actions, and (2) Consistency — it ensures that similar instances receive consistent actions without conflicts. This framework is especially useful in organizational settings where decisions (such as promotions or salary adjustments) affect multiple individuals, and consistent, understandable justifications are necessary to avoid perceptions of unfair treatment.

Problem Formulation In Kanamori et al. (2022), the goal is to maximize effectiveness while minimizing costs, all while maintaining interpretability — essentially the same objective pursued by other methods like AReS and GLOBE-CE. The formulation here is conceptually similar to AReS in that it defines an objective function that combines these three competing goals — effectiveness, cost, and interpretability — into a single weighted sum. The details of how these quantities are defined differ slightly from AReS; the reader is referred to Kanamori et al. (2022) for specifics. The central entity in this framework is the *Counterfactual Explanation Tree* (CET), which operates similarly to rule sets in AReS but differs in that it employs a decision tree structure. The tree partitions the feature space into groups (represented by its leaves), and each group is assigned an action to achieve recourse.

The problem formulation in CET can then be expressed as minimizing the following objective function:

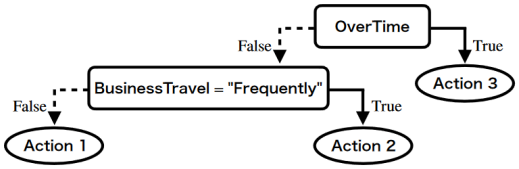
$$\text{Cost} - \gamma\text{Eff} + \lambda(\#\text{Leaves}),$$

where:

- **Cost**: Represents the mean recourse cost for the individuals.
- **Eff**: Stands for effectiveness, i.e. the proportion of individuals who achieve recourse by following their assigned actions.
- **# Leaves**: Denotes the number of leaves in the tree, which reflects the complexity of the model and is used as a proxy for interpretability.

By combining these components, the framework tries to balance the competing objectives of cost, effectiveness, and interpretability within a single optimization problem. The coefficients γ and λ are basically user-defined, allowing but also requiring the practitioner to prioritize these aspects differently and appropriately depending on the specific application and dataset. Although Kanamori et al. (2022) provides some theoretical insights linking these coefficients to the maximum number of leaves, they are generally chosen by the user based on the domain requirements.

Method The algorithm CET framework learns the actions and the structure of the tree in two steps: (i) it identifies a single effective action for multiple instances, balancing the trade-offs between cost and the success rate of flipping predictions, and (ii) it constructs a decision



	HowToChange	Effectiveness Cost	Effectiveness Flip rate
Action 1	MonthlyIncome : + 1282$	0.17	83 %
Action 2	BusinessTravel : "Frequently" → "Rarely"	0.19	80 %
Action 3	OverTime : True → False	0.27	86 %

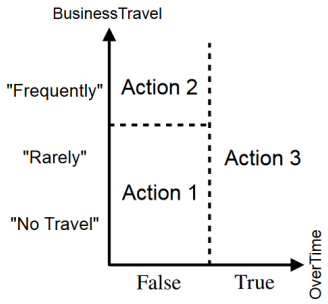


Figure 2.5: Example output from the CET framework. Taken from Kanamori et al. (2022). The dataset used is the IBM HR Analytics Employee Attrition dataset (Kaggle, 2017). The constructed tree assigns to each instance on of three different actions, based on the features 'OverTime' and 'BuisinessTravel'.

Dataset	Method	Train			Test		
		Cost	Loss	Invalidity	Cost	Loss	Invalidity
Attrition	Clustering	0.055 ± 0.04	0.951 ± 0.03	1.0 ± 0.01	0.047 ± 0.05	0.958 ± 0.06	1.01 ± 0.02
	AReS	0.436 ± 0.06	0.435 ± 0.07	0.871 ± 0.04	0.45 ± 0.08	0.298 ± 0.09	0.748 ± 0.09
	CET	0.349 ± 0.1	0.4 ± 0.11	0.749 ± 0.05	0.383 ± 0.12	0.318 ± 0.19	0.701 ± 0.12
German	Clustering	0.034 ± 0.02	0.915 ± 0.05	0.949 ± 0.04	0.039 ± 0.02	0.917 ± 0.05	0.956 ± 0.04
	AReS	0.452 ± 0.09	0.232 ± 0.05	0.683 ± 0.11	0.467 ± 0.12	0.265 ± 0.08	0.732 ± 0.14
	CET	0.103 ± 0.01	0.301 ± 0.07	0.404 ± 0.07	0.107 ± 0.02	0.276 ± 0.11	0.384 ± 0.1

(a) LightGBM classifier

Dataset	Method	Train			Test		
		Cost	Loss	Invalidity	Cost	Loss	Invalidity
Attrition	Clustering	0.509 ± 0.16	0.487 ± 0.24	0.996 ± 0.23	0.515 ± 0.16	0.513 ± 0.23	1.03 ± 0.23
	AReS	0.161 ± 0.1	0.235 ± 0.1	0.396 ± 0.1	0.164 ± 0.1	0.281 ± 0.16	0.445 ± 0.18
	CET	0.312 ± 0.12	0.198 ± 0.14	0.51 ± 0.19	0.296 ± 0.13	0.258 ± 0.21	0.554 ± 0.23
German	Clustering	0.01 ± 0.01	0.981 ± 0.02	0.99 ± 0.01	0.009 ± 0.01	0.99 ± 0.01	0.999 ± 0.01
	AReS	0.18 ± 0.08	0.405 ± 0.14	0.585 ± 0.2	0.188 ± 0.09	0.427 ± 0.17	0.614 ± 0.22
	CET	0.121 ± 0.08	0.242 ± 0.21	0.362 ± 0.19	0.12 ± 0.08	0.234 ± 0.21	0.355 ± 0.19

(b) TabNet classifier

Figure 2.6: Quantitative results of CET on the Attrition and German datasets. Taken from Kanamori et al. (2022). For each experiment, Cost, Loss and Invalidity are shown. 'Cost' is the cost of the actions found, 'Loss' is essentially $1-$ effectiveness, and 'Invalidity' is a weighted sum of the two which CET uses as objective function. `Clustering` is a simple method developed in-house by CET due to lack of other competitors except AReS at the time.

tree to partition the input space and assign actions in an interpretable and efficient way. As we discussed, the decision tree structure allows CET to maintain interpretability while ensuring each instance is assigned a unique action. This contrasts with other CE methods that lack transparency when applied to large groups, leading to potentially conflicting or unclear explanations.

An example of the output produced by CET is shown in Figure 2.5. One can observe that the actions, as well as the splitting of the feature space based on feature values, are simple, small and understandable. They achieve satisfactory effectiveness, while the cost is again based on a different definition, so immediate conclusions cannot be drawn. Nonetheless, the output could be described, we believe, as straightforward and seemingly effective.

A portion of the results presented in Kanamori et al. (2022) is shown in Figure 2.6. We are mainly interested in Loss, since the costs are defined again in a different way in Kanamori et al. (2022) and invalidity is simply a combination of the two. CET performs reasonably well, with a maximum loss of 0.4, i.e. minimum effectiveness 60%. This value for effective-

Classifier	Dataset	Clustering	AReS	CET
LightGBM	Attrition	24.023 ± 3.77	1268.14 ± 107.41	8021.79 ± 2600.78
	German	19.865 ± 1.32	18767.0 ± 538.51	5782.37 ± 901.46
TabNet	Attrition	2036.75 ± 595.11	4118.18 ± 375.96	27009.4 ± 5891.93
	German	4258.59 ± 1012.93	36242.8 ± 2844.69	9969.62 ± 2463.63
Average		1584.81	15099.0	12695.8

Figure 2.7: Average computational time of CET on the Attrition and German datasets. Taken from Kanamori et al. (2022).

ness is fairly low, compared for example with GLOBE-CE which we saw earlier, but it may still be acceptable in certain scenarios.

On the other hand, the CET framework also has several drawbacks. Firstly, it is very computationally demanding. Even in the runtimes reported by CET itself, and shown verbatim in Figure 2.7, the computations take several thousand seconds, i.e. several hours. Additional experiments we conducted exhibited somewhat longer runtimes. Not by orders of magnitude for the specific cases of the 'Attrition' and 'German' datasets, which are relatively small, so one could argue that this may be attributable to differences in hardware.

However, when applied to larger datasets, CET's computational demands became more apparent. Notably, when running CET on the 'Adult' dataset — which contains around 30,000 instances and 100 dimensions after one-hot encoding, compared to the 'German' dataset's 1,000 instances and 71 dimensions — it initially failed to produce a solution within the timeout limits of the MIP solver we were using (GUROBI). While GUROBI is a well-known, efficient solver comparable to CPLEX (which the authors originally used but we did not have access to), raising the timeout limit still resulted in CET not terminating even after approximately 20 hours. This issue stems from the inherent complexity of the optimization problem Kanamori et al. (2022) tries to solve, which requires solving an integer programming subproblem at each step.

Additionally, it should be noted that, as the authors mention themselves, the method and its implementation do take advantage of model internals for specific cases (e.g. linear classifiers). There are specific types of models for which the authors have formulated a suitable mixed integer programming problem for the discovery of counterfactuals at each node of the tree. The procedure does fall back to LIME approximation when the model is not recognized to be one of the known ones, so all cases of models can be handled by the framework. Nonetheless, it is worth noting that the method may perform better with models it can directly optimize, and we believe it would be valuable to conduct further experiments using black-box models to better assess CET's performance in these cases.

2.3.5 Problem Formulation

To approach the problem of global counterfactuals in a principled way, Kavouras et al. (2024) begins by specifying the requirements one would expect from a global counterfactual.

First of all, a global counterfactual explanation should be interpretable. For example, it is clear we cannot be allowed to simply provide a local counterfactual for each point and call it a day. We want the number of actions we propose to be relatively small.

Additionally, no action which does not flip any individual would have any merit. As such, we want the actions we choose to flip as many individuals as possible.

Finally, actions that are too large are implausible, so we want them to be as small as possible.

The above requirements are quantified by Kavouras et al. (2024) with the following definition, which is the one we also adopt for the current work.

Definition 2.3 (Global Counterfactuals). *A global counterfactual for $\mathcal{D}_{aff}$ is a set $S \subseteq \mathcal{A}, S \neq \emptyset$ that represents a solution to the following multi-objective optimization problem:*

$$\begin{aligned} &\underset{S \subseteq \mathcal{A}}{\text{minimize}}\ \mathsf{size}(S) \\ &\underset{S \subseteq \mathcal{A}}{\text{minimize}}\ \mathsf{avc}(S, \mathcal{D}_{aff}) \\ &\underset{S \subseteq \mathcal{A}}{\text{maximize}}\ \mathsf{eff}(S, \mathcal{D}_{aff}) \end{aligned} \tag{2.6}$$

In summary, to choose a set of actions as a global counterfactual explanation, we want it to have low cost, high effectiveness and be as small as possible.

It can be seen that these three objectives are somewhat contradicting. For example, the zero action (which changes nothing) has the minimum possible cost of 0, but also terrible effectiveness, also 0. Intuitively (albeit a bit over simplified), we would expect the cost to rise as the effectiveness rises, and vice versa. As for the number of actions, we already mentioned that one can choose a set of actions comprising only of local counterfactuals, and this will minimize the cost and maximize the effectiveness; however, the number of actions will too large, extinguishing explainability altogether.

In addressing Problem 2.3, we employ an agglomerative approach (one of those proposed in our team's work Kavouras et al. (2024)). The main idea is to start by forming multiple small clusters of individuals and determine representative actions for each cluster. These initial clusters will be merged using the proposed procedure until the desired number of clusters is reached. Ultimately, a single optimal action will be extracted from each cluster, applicable across all instances. The underlying principle is that *"similar individuals are expected to achieve recourse with similar counterfactual actions"*, utilizing the concept of proximity in both the original feature space and the action space.

Chapter 3 is devoted to the exposition of this approach and the analysis of corresponding experimental results.

2.4 Fairness of Recourse: A Critical Application of Counterfactual Explanations

As we have discussed throughout this chapter, counterfactual explanations provide insights into how individuals can modify their attributes to achieve favorable outcomes, such as loan approvals or job promotions. These explanations offer transparency and a practical way for individuals to understand and alter the decisions made by machine learning models. However, beyond explainability and transparency, counterfactuals may also play a crucial role in ensuring *fairness of recourse* — an increasingly important area in fairness and machine learning.

2.4.1 Fairness in Machine Learning: A Broader Context

Traditionally, fairness in machine learning has focused on *fairness of prediction*, aiming to ensure that model outcomes do not disproportionately disadvantage individuals or groups based on protected attributes such as race, gender, or age. These fairness definitions — such as statistical parity, equal opportunity, and disparate impact (Mehrabi et al., 2022) —

evaluate bias in terms of model predictions. For example, statistical parity requires that all subpopulations defined by a protected attribute have equal probabilities of receiving a favorable prediction. These notions of fairness are well-studied and have led to various pre-processing, in-processing, and post-processing methods designed to detect or mitigate prediction bias (Mehrabi et al., 2022).

However, this focus on fairness of prediction captures only one dimension of bias: the *explicit* bias present in model outcomes. It does not address a more *implicit* form of bias — namely, the *burden* (Sharma et al., 2020; Kuratomi et al., 2023) or difficulty an individual (or group thereof) faces in achieving *recourse*.

2.4.2 Fairness of Recourse: Beyond Predictions

Recourse refers to an individual or group performing the necessary *actions* to change their features so as to obtain the favorable outcome (Gupta et al., 2019; Ustun et al., 2019). This provides explainability (in the form of counterfactual explanations, as discussed earlier) and actionability to an affected individual. This concept is particularly important in sensitive domains such as finance (Boer et al., 2019), healthcare (Kyrimi et al., 2020), and employment (Bogen and Rieke, 2018; Cohen et al., 2019), where individuals must be able to take action to improve their standing. In certain domains, offering explanations for adverse decisions is not just a good practice but a legal obligation. For instance, the **Equal Credit Opportunity Act** gives individuals the right to demand the reasons for a loan denial.

Fairness of recourse extends the concept of fairness to the realm of *recourse* and *counterfactual explanations*. It seeks to ensure that individuals from different protected subgroups (e.g., based on race or gender) bear an equal burden when seeking recourse (Gupta et al., 2019; Karimi et al., 2022; von Kügelgen et al., 2022; Kuratomi et al., 2023). In other words, fairness of recourse captures the notion that individuals from different subgroups ought to have equal opportunities to change their attributes and obtain a favorable outcome.

In this context, fairness of recourse measures whether the burden of achieving recourse is distributed equally across protected groups. Recourse is typically evaluated based on the *cost* of the actions needed to cross the decision boundary. For example, in a credit scoring model, recourse might involve increasing income or paying off debts to improve one's credit score. If members of one protected subgroup consistently face higher costs than members of another group to achieve the same favorable outcome, the system is considered unfair with respect to recourse.

2.4.3 Connecting Recourse to Statistical Fairness

The intuition behind fairness of recourse shares similarities with statistical fairness concepts like **statistical parity**. When we aim to ensure that recourse is equally achievable by different subgroups, we are essentially requiring a form of **parity** — not of predictions, but of the ability to achieve recourse. This can be thought of as "parity of recourse", where the goal is to provide protected subgroups with the same opportunities to change their outcomes through actions. Just as statistical parity requires equal access to favorable predictions, fairness of recourse requires equal access to effective actions for changing those predictions.

However, fairness of recourse goes further, as it not only concerns the outcome but also the path an individual must take to alter the outcome. This adds a more nuanced layer to fairness, as the *distribution of costs* and the availability of *actionable changes* across

subgroups need to be considered.

2.4.4 Measuring Fairness of Recourse: The Role of Counterfactual Explanations

To measure fairness of recourse, we rely on counterfactual explanations, which provide a set of actions an individual or group of individuals can take to flip a model's decision. For example, in a loan application setting, a counterfactual explanation might suggest increasing income or reducing debt to move from a denial to an approval. These actions incur a cost for the individual or individuals, which typically measures the effort required to change certain features, such as moving income from one level to another.

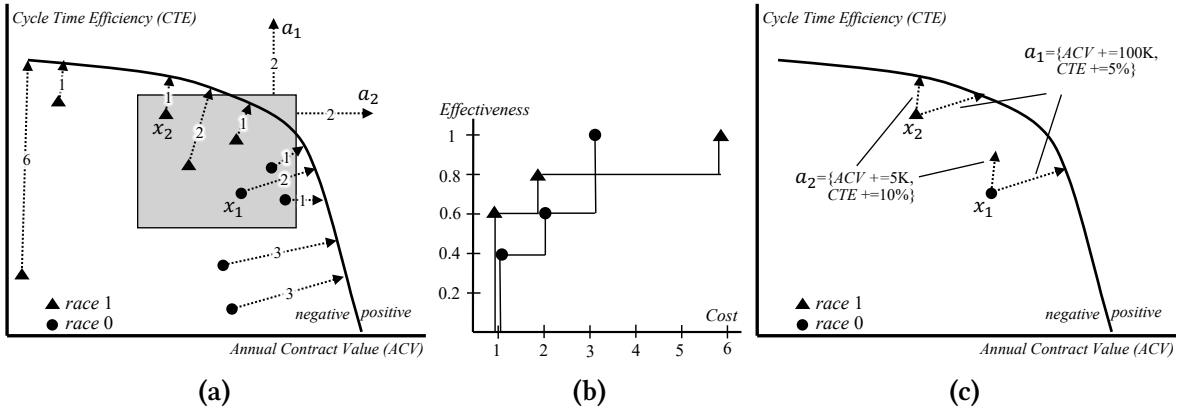


Figure 2.8: (a) An example of affected individuals, the decision boundary, actions, and a subpopulation (in the shaded region), depicted in the feature space; (b) Cumulative distribution of cost of recourse for the individuals in (a); (c) Comparison of two actions to achieve recourse for two individuals.

To illustrate these concepts, consider a company using an AI system to support promotion decisions by classifying employees as good candidates for promotion (the favorable positive class) or bad candidates for promotion (the negative class) based on various performance metrics; for illustration purposes, let's assume these are just their **cycle time efficiency** (CTE) and the **annual contract value** (ACV) of the projects they lead. In Figure 2.8a, we can see a two-dimensional plot where the x-axis represents ACV and the y-axis represents CTE. The classifier's decision boundary — the line that separates the favorable and unfavorable classes — is shown as a curved sold line. Each point on the plot represents an employee classified by the model as not ready for promotion (the negative class). The protected attribute in this example is **race**, with two subgroups visually distinguished by the shape of the points: circles represent employees from one racial group, and triangles represent employees from another.

For each employee, we have identified an action — represented as an arrow — that indicates the shortest path from the employee's current position to the decision boundary. This action reflects the smallest change that the employee must make in either CTE or ACV to become a strong candidate for promotion. The *cost* of an action is calculated as the length of the arrow, which measures the distance from the employee's current position to the decision boundary (Gupta et al., 2019).

For example, in the case of employee x_1, the easiest way to become a strong candidate is to increase their ACV. On the other hand, for employee x_2, the easiest path involves increas-

ing CTE. Each arrow is therefore a visual representation of a counterfactual explanation, providing clear and actionable guidance for each employee.

We are now interested in assessing whether there is bias in the cost of achieving recourse between the two racial subgroups. To do this, the measure called *burden* is typically used. Burden is defined as the average cost required for all individuals within a subgroup to achieve recourse (Sharma et al., 2020). In our example, one can calculate that the mean cost for employees with race 0 (represented by circles) is 2, while the mean cost for employees with race 1 (represented by triangles) is 2.2. This suggests that employees with race 1 face a higher burden to achieve recourse compared to those with race 0. Notice that despite this disparity in recourse costs, the model might still satisfy traditional fairness metrics like statistical parity, which focuses solely on the rate of positive outcomes and does not consider the difficulty of achieving those outcomes.

2.4.5 Beyond Mean Cost: The Effectiveness - Cost Distribution

The concept of using a single average cost (or burden) to measure fairness has limitations. It may fail to capture the distribution of costs within a subgroup and can potentially mask, introduce or even invert important differences within subgroups. In our example, observing again Figure 2.8a, the mean cost for race 1 is higher primarily because one employee in that subgroup has a particularly high cost of 6, while the rest face costs of at most 2.

Figure 2.8b illustrates this through the cumulative distribution curves of the cost for each subgroup, which is called the *effectiveness-cost distribution* (ecd) by Kavouras et al. (2023) (see also Chapter 4 for more details). The ecds provide a more comprehensive picture of the costs faced by individuals within each subgroup and allow the fairness auditor to better understand the *trade-offs* between recourse cost and effectiveness.

For example, if recourse with a cost higher than 2 is deemed unrealistic in a given context (e.g. due to time or feasibility constraints), fairness of recourse could be assessed by examining the proportion of individuals from each subgroup that can achieve recourse within this budget — this leads to a definition termed *equal effectiveness within budget* by Kavouras et al. (2023) (see also Section 4.1.1). In this case, only 60% of subgroup 0 can achieve recourse within the budget, compared to 80% of subgroup 1, indicating unfairness against subgroup 0 under this specific budget constraint.

2.4.6 A Macro View: Collective Actions and Systemic Bias

The authors of Kavouras et al. (2023) argue that the individual-level (or *micro viewpoint* as they call it) evaluation of actions and burden employed in earlier literature (Gupta et al., 2019; Sharma et al., 2020; von Kügelgen et al., 2022; Kuratomi et al., 2023) might not be meaningful for many real-world systems. For this reason, they introduce a *micro viewpoint* from which fairness of recourse can also be assessed. This perspective involves considering subgroups as a *collective* and assessing the impact of a *single action* on the entire subgroup, as opposed to individual-level actions. An action from the macro viewpoint represents a more systemic intervention, such as a training program to improve productivity or a policy to reduce tax or loan interest rates.

In this perspective, the cost of recourse does not burden the individuals directly but is borne by an external entity, such as society or an organization. This viewpoint offers a more intuitive way to *audit* a system for fairness of recourse, as it aims to uncover systemic biases that affect large groups of individuals collectively.

For example, in Figure 2.8a, consider the shaded region that encompasses a specific group of employees. From a micro viewpoint, each employee seeks recourse individually, and their actions are assessed independently. However, when we consider a macro action — depicted as a_1 on the boundary of the shaded region — we discover that this action leads to favorable outcomes for two-thirds of race 1 employees but none of race 0. This highlights a systemic bias in this population group (shaded region) where race 0 employees are disproportionately affected by the intervention, even though the cost of achieving recourse is balanced on an individual level.

2.4.7 Moving Beyond Cost: Cost-Oblivious Fairness of Recourse

A final challenge in fairness of recourse lies in defining the *cost function* itself. Current notions of fairness of recourse rely on a cost function that reflects an individual's ability to modify their attributes. Defining such a function can be complex and might involve a learning process, as seen in Rawal and Lakkaraju (2020), or require practitioners to adapt standard cost functions, as suggested in Ustun et al. (2019). Moreover, determining which attributes are actionable is not straightforward. For instance, certain attributes cannot be changed (e.g. age), while others may raise ethical concerns if suggested as actionable (e.g. changing marital status from married to divorced) (Venkatasubramanian and Alfano, 2020).

At the same time, the definition of cost can significantly impact the assessment of fairness. Suppose we have two actions, a_1 and a_2, available to employees x_1 and x_2, as illustrated in Figure 2.8c. It's challenging to determine which action is less costly because we must compare changes both within and across very different attributes. For example, if the cost function indicates that action a_1 is cheaper, both x_1 and x_2 achieve recourse with the same cost. However, if action a_2 is considered cheaper, only x_2 achieves recourse. The question of whether the classifier is fair hinges on how we define and compare the costs of these actions.

To overcome this limitation, the authors in Kavouras et al. (2023) propose fairness definitions that are independent of the cost function, or in other words, *cost-oblivious*. These evaluate fairness based on the availability of effective actions for each protected subgroup, regardless of their cost. For instance, in Figure 2.8c, since employee x_2 has two available actions to achieve recourse (both a_1 and a_2), while employee x_1 has only one ($a1$), we might consider the classifier unfair against x_1. This leads to a fairness notion which Kavouras et al. (2023) calls *equal choice for recourse*, where fairness is evaluated based on the number of available options rather than their associated costs (further explained in Section 4.1.1).

2.4.8 Conclusion: Fairness of Recourse as a Key Application of Counterfactuals

In conclusion, *fairness of recourse* is a crucial application of counterfactual explanations in machine learning. While fairness of prediction addresses explicit bias in model outcomes, fairness of recourse ensures that protected subgroups are not disproportionately burdened when seeking to change those outcomes. Through the careful analysis of recourse costs, distributions, and collective actions, fairness of recourse provides a deeper understanding of systemic biases and offers pathways for more equitable machine learning systems.

Chapter 3

Clustering Approach for Global Counterfactual Generation

3.1 Algorithm Description

The algorithm developed for solving the global counterfactual problem 2.3 is called **Augmented Space**, and its pseudocode is presented in Algorithm 1. Below, we offer a textual description.

From a high-level perspective, Augmented Space is a clustering-based algorithm that leverages an embedding from the feature space into a *higher* dimensional space. This embedding allows the algorithm to consider both *instances* and *actions* simultaneously, ensuring that both are factored into the decision-making process.

The algorithm operates in three main phases. First, it creates multiple small clusters of instances and generates diverse actions for each cluster. The instances from each cluster are then paired with the most suitable action from the corresponding cluster's generated actions. These pairs are combined in a new *augmented space*, where each instance is represented by both its original feature values and the associated action. This allows us to exploit the notion of proximity in both the original feature space and the action space. We do this in the third step, where a second clustering step is performed in this augmented space. Finally, a single action is extracted from each of these final clusters, although with universal applicability across *all* instances.

Phase 1: Generation of Initial Clusters and Actions. The feature space is partitioned into k clusters, by a clustering algorithm (e.g., k-means, line 3, Algorithm 1). We compute the *centroid* of each cluster and generate m *diverse* counterfactual actions for each centroid (line 4, Algorithm 1), employing any local counterfactual generation method (in the case of our experimental evaluation, this was Dice (Mothilal et al., 2019)). This yields a total of km actions. We aim to efficiently and effectively explore the action space by generating actions from widely dispersed points within the feature space and guiding them in diverse directions that cross the decision boundary.

Phase 2: Creation of Augmented Space. For each instance in a cluster, the algorithm selects the action from its cluster that converts its label to positive with the smallest possible cost[1] (lines 6–7, Algorithm 1). This action is then combined with the instance's feature vector to create a *new augmented space*, where the original instance is now represented by both its feature values and the best cost-effective action (line 8, Algorithm 1).

[1] The action's feature vector contains d entries, where each entry indicates the change required for the corresponding feature. For example, $(0, +1, 0)$ means increasing the second feature by one unit. If no action in the cluster changes the instance's label, the action's feature vector will consist of zeros.

Phase 3: Reclustering in the Augmented Space. In the final phase, the algorithm clusters the augmented instances, producing t new clusters in the augmented space (line 9, Algorithm 1). From each cluster, it extracts the most effective action that applies to the entire cluster (line 10). This ensures that the final set of actions is representative of the clusters and cost-effective.

This procedure creates final clusters that account for both the instances and their corresponding most cost-efficient actions. The reclustering ensures that the final set of t actions achieves a good balance between effectiveness and cost when applied to the affected population. Without this reclustering step, finding the best subset of t actions from the initially generated actions would require an exhaustive search, which is computationally prohibitive, as it would involve evaluating $\binom{|\mathcal{A}|}{t}$ possible action subsets.

It is important to note that the *Augmented Space* algorithm inherently prioritizes minimizing average cost slightly. Each original instance is associated with the lowest-cost action from its initial cluster, which results in a more cost-efficient subset of $l < km$ actions being selected. The final action for each cluster is chosen from this filtered subset of lower-cost actions.

Algorithm 1 Augmented Space

1: **Input**: $\mathcal{D}_{aff}$, number of local counterfactuals m, number of initial clusters k, number of final clusters t
2: **Output**: t global counterfactuals
3: Cluster $\mathcal{D}_{aff}$ into k clusters.
4: For each cluster, generate m counterfactuals from the cluster centroid.
5: For each instance x_i:
6: Find the set A_i of cluster actions that flip its prediction.
7: Find minimum cost action a_i from A_i.
8: Concatenate x_i and a_i.
9: Cluster the augmented data into t clusters.
10: Return the single optimal action from each of the t clusters.

Approach strengths The Augmented Space algorithm supports various clustering methods, cost metrics, and action generation techniques, making it adaptable to different applications. Additionally, it demonstrates speed, robustness, and near-optimal solution quality across datasets, as shown in the experimental evaluation section 3.3.

3.1.1 Example Output

Figure 3.1 presents a counterfactual summary consisting of three global actions derived by GLANCE for a Logistic Regression classifier trained on the HELOC dataset. Table 3.1 quantitatively compares the action sets derived by GLANCE with those from local counterfactual explainability (Local CE) and GLOBE-CE. Note that Local CE provides an action for each instance, achieving perfect effectiveness and good cost. GLOBE-CE reduces the cost slightly, with a small drop in effectiveness, but still produces a lengthy and less interpretable set of actions. In contrast, GLANCE offers a concise, interpretable action set of three actions (Figure 3.1) that achieves optimal effectiveness with a negligible increase in cost.

```
ACTIONS:
 (1) NumInqLast6Mexcl7days + 9.1
 (2) NumInqLast6Mexcl7days + 13.87 & NumSatisfactoryTrades - 0.75
 (3) NumInqLast6Mexcl7days + 10.27
```

Figure 3.1: Suggested actions (Algorithm 1) for HELOC dataset-LR model.

Table 3.1: Cardinality, cost, and effectiveness of the action sets generated by different methods for the HELOC dataset. "LOCAL CE" refers to local counterfactuals generated from DiCE (Mothilal et al., 2019).

Method	Actions	Cost	Effectiveness
Algorithm 1	3	1.38	100.00%
Globe-CE	588	0.41	99.90%
Local CE	1045	1.16	100.00 %

3.2 Experimental Setting

Building on the theoretical foundations discussed above, Emiris et al. (2024) proceeds to validate the effectiveness of Algorithm 1 in generating satisfactory global counterfactuals through a series of experiments on well-known datasets. We do the same for our work, albeit following the updated methodology of Kavouras et al. (2024). The remaining of this chapter is dedicated to presenting these results and making further commentary.

To begin with, in this section, we outline the experimental setup in detail. The section is formatted as follows.

1. In 3.2.1 we discuss the *Datasets* used in our experimental evaluation and the *Preprocessing* performed in each of them.

2. In 3.2.2 we discuss the *Models* used in our experimental evaluation.

3. In 3.2.3 we discuss all the *Hyperparameters* used in our experimental evaluation, both for the state-of-the-art global counterfactual methods we aim to evaluate as well as for some algorithms utilized ubiquitously in those works.

For the presentation and analysis of the results obtained for the global counterfactual explainability problem 2.3 via the Augmented Space algorithm (and the other methods), the interested reader can skip directly to Section 3.3.

3.2.1 Datasets & Preprocessing

We use four publicly available and well-known *benchmark datasets*; the same ones that Kavouras et al. (2024) uses. Their choice is based on the datasets' established use in previous works. A short description follows. Table 3.2 summarizes the datasets' information, including the number of instances, the number of categorical and continuous features, the input dimensions (i.e., the number of continuous features plus the number of categorical after a preprocessing step using one-hot-encoding), and the number of instances used for train and test after the 80:20 split.

Incidentally, we should mention that the preprocessing of the datasets, the models, and their hyperparameters are all taken from Ley et al. (2023) and Kavouras et al. (2024) (which replicates Ley et al. (2023)). Nonetheless, we provide here a brief description of each dataset, for completeness.

COMPAS The COMPAS dataset (Correctional Offender Management Profiling for Alternative Sanctions) (Angwin et al., 2016) is available at `https://github.com/propublica/compas-analysis/blob/master/compas-scores-two-years.csv`. Detailed description and information on the dataset can be found at `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`. It categorizes recidivism risk based on several factors, including race.

For the preprocessing of this dataset, we drop the "days_b_screening_arrest" feature, as it contains missing values. We also turn jail-in and jail-out dates to durations and turn negative durations to 0. Some additional filters are taken from the COMPAS analysis by ProPublica. Finally, the target variable's values are transformed into the canonical 0 for the negative class and 1 for the positive class.

Adult The Adult dataset, also known as the "Census Income" dataset, is used for predicting whether an individual's income exceeds \$50K per year based on various demographic and employment attributes. The dataset and additional details are available at `https://archive.ics.uci.edu/dataset/2/adult`.

For preprocessing, we begin by dropping the "education-num" feature, as it is redundant with the "education" feature (represented as string categories). We then remove rows with missing values and map the class labels '$\leq$50K' and '$>$50K' to 0 and 1, respectively. Minor additional transformations, mainly concerning data types, are performed and are documented in our source code.

German Credit The German Credit dataset (Dua and Graff, 2019) classifies people as good or bad credit risks, based on a set of attributes. A detailed description and the dataset can be found in `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`.

The only preprocessing step we performed for this dataset was the transformation of the target variable's values into 0 - 1.

Default Credit The Default Credit dataset (Yeh and Lien, 2009) is designed to classify the risk of default on customer payments based, aiming to support the development and assessment of models for predicting creditworthiness and the likelihood of loan default. It can be obtained at `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`.

To properly work with this dataset, we needed to drop the "ID" feature, since it holds no useful information, and transform the target labels into the canonical 0 - 1 values.

HELOC The HELOC (Home Equity Line of Credit) dataset (Brown et al., 2018) contains anonymized information about home equity line of credit applications made by real homeowners, classifying credit risk. It is available at `https://community.fico.com/s/explainable-machine-learning-challenge`. All the features on this dataset are numeric.

A substantial percentage of these features' values are missing, so the main preprocessing step we performed here was to remove any rows where all values are missing, and then replace all remaining missing values with the median of the respective feature. Other than that, we only needed to transform target labels to 0-1.

Table 3.2: Summary of the datasets used in our experiments. Specifically, we list the number of instances, input dimensions (i.e., the number of continuous features plus the number of categorical after a preprocessing step using one-hot-encoding), the number of categorical and continuous features, and the number of instances used for train and test after the 80:20 split.

Dataset	No. Instances	Input Dim.	Categorical	Continuous	Train	Test
COMPAS	6172	15	4	2	4937	1235
German Credit	1000	71	17	3	800	200
Default Credit	30000	91	9	14	24000	6000
HELOC	9871	23	0	23	7896	1975

3.2.2 Models

We use three distinct types of models: gradient boosted decision trees from XGBoost (XGB), Logistic Regression (LR), and Deep Neural Networks (DNNs). We train these models with the same 80:20 train-test split as done by both Ley et al. (2023) and Kavouras et al. (2024); the hyperparameters for each model were also obtained from these two works, facilitating a standardized basis for the comparative analysis of the results.

XGBoost (XGB) Implementation from the common `xgboost`[2] library. Hyperparameter values for each dataset and the model's accuracy on the test set are shown in 3.3.

Table 3.3: XGBoost Hyperparameter Configurations.

Dataset	Depth	Estimators	γ, α, λ	Test Accuracy
Compas	4	100	1,0,1	68%
German Credit	6	500	0,0,1	74%
Default Credit	10	200	2,4,1	83%
Heloc	6	100	4,4,1	74%

Logistic Regression (LR) Implementation from the common `scikit-learn`[3] library. Hyperparameter values for each dataset and the model's accuracy on the test set are shown in 3.4.

[2] `https://xgboost.readthedocs.io/en/stable/`

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

Table 3.4: Logistic Regression Hyperparameter Configurations.

Dataset	Max Iter.	Class Weights(0:1)	Test Accuracy
Compas	1000	1:1	65%
German Credit	1000	1:1	76%
Default Credit	2000	0.65:0.35	83%
Heloc	2000	1:1	75%

Deep Neural Network (DNN) Implemented using the`pytorch`[4] library. Hyperparameter values for each dataset and the model's accuracy on the test set are shown in 3.5.

Table 3.5: Deep Neural Network Hyperparameter Configurations.

Dataset	Width	Depth	Dropout	Test Accuracy
Compas	30	5	0.4	65%
German Credit	50	10	0.3	78%
Default Credit	80	5	0.3	81%
Heloc	50	5	0.3	74%

3.2.3 Hyperparameters & Implementation Details

In this section, we discuss the hyperparameters for the most prominent algorithms developed or used by Kavouras et al. (2024), Ley et al. (2023), and Kanamori et al. (2022). We explain the purpose of each parameter and present the values used in our experiments. We begin by describing the parameters for the developed algorithms — Augmented Space, GLOBE-CE, and CET — followed by the algorithms used predominantly in Kavouras et al. (2024), such as k-means and DiCE. We close with a brief explanation of the cost function that we use, which is the same as in Kavouras et al. (2024) and Ley et al. (2023).

Augmented Space

The Augmented Space algorithm starts by clustering individuals in the feature space and generating representative counterfactuals for each cluster. These counterfactuals are then translated into actions and applied to all individuals within the cluster. The first two important parameters are n_initial_clusters, which defines the number of initial clusters, and n_cluster_samples, which specifies how many counterfactual explanations to generate per cluster. Counterfactuals can be generated using two sampling methods — `sampling` or `centroid` — which is controlled by the Sampling_method parameter.

In the second phase, the algorithm concatenates each instance with its most cost-effective action, followed by re-clustering these augmented instances. The n_final_clusters parameter defines the number of final clusters. The algorithm's output includes these clusters, along with the most effective action for each cluster, reported alongside its effectiveness

[4] `https://pytorch.org/docs/stable/index.html`

and cost. Additionally, the DIST_FUNC parameter defines the cost function for measuring the distance between instances. To ensure reproducibility, we use the RANDOM_STATE parameter.

In our experiments, we set N_INITIAL_CLUSTERS to 100 and N_CLUSTER_SAMPLES to 10 for most datasets, except for the German Credit dataset, where N_INITIAL_CLUSTERS was set to 20 and N_CLUSTER_SAMPLES to 50, maintaining a constant total number of generated counterfactuals. We consistently used the "centroid" option for SAMPLING_METHOD and set N_FINAL_CLUSTERS to 3 across all experiments. The DIST_FUNC parameter was defined as the cost function detailed below, in Section 3.2.3, and RANDOM_STATE was set to 13.

GLOBE-CE

The state-of-the-art GLOBE-CE algorithm (Ley et al., 2023) requires several key parameters, including N_SAMPLE, which defines the number of directions sampled, and MAGNITUDE, which specifies the magnitude of these direction vectors. The N_FEATURES parameter determines how many features are altered by the generated actions.

A critical parameter is N_DIV, which sets the number of best directions selected from the initial sample. For standard GLOBE-CE, this is set to 1, meaning only the direction with the highest effectiveness is chosen. For dGLOBE-CE, multiple ($d > 1$) directions are selected. After selecting the directions, the algorithm explores each by scaling them with different scalar values, the number of which is controlled by the N_SCALARS parameter.

In our experimental setup, we standardized N_SAMPLE and N_SCALARS to 1000. We also experimented with varying values for SPARSITY_POWER (1 and 5), MAGNITUDE (1 and 2), and N_FEATURES (5 and 2), consistent with parameter choices from GLOBE-CE's source code. These values were chosen to achieve the best cost-effectiveness trade-off across our experiments. For dGLOBE-CE, N_DIV was set to 3.

GroupCF

The GroupCF method, introduced by Warren et al. (2023), was implemented using custom code. While we initially attempted to use the original GitHub repository (`https://github.com/e-delaney/group_cfe`), certain ambiguities in the setup and implementation details made it challenging to reproduce the results as described in the paper. Therefore, we developed a custom implementation based on our interpretation of the algorithm to ensure consistency with our experimental setup.

In terms of hyperparameters, GroupCF relies heavily on parameters governing the similarity and distance metrics, as these directly influence the formation of instance groups and the recourse actions they receive. Additionally, we set the maximum group size and recourse coverage threshold based on validation performance, adjusting these to balance interpretability (via fewer groups) with coverage (ensuring a high percentage of individuals receive actionable recourse).

CET

The tree-based CET algorithm (Kanamori et al., 2022) requires parameters such as MAX_ITERATION, which limits the number of iterations of the iterative local search algorithm, MAX_LEAF_SIZE, which controls the maximum size of tree leaves, and MAX_CHANGE_NUM, which defines the maximum number of features that can be altered.

Additionally, two scalar parameters, GAMMA and LAMBDA, are used in the objective function of the optimization problem that CET solves at each tree node. We should mention that tweaking these parameters appears to be essential for achieving satisfactory results.

In our experiments, we set MAX_ITERATION to 100, MAX_CHANGE_NUM to 4, GAMMA to 1, and LAMBDA to 0.02. We ran each experiment twice: once with MAX_LEAF_SIZE set to 4 and once with it set to 8.

AReS

The AReS method, first introduced by Rawal and Lakkaraju (2020), was implemented using the enhanced Fast AReS version provided by Ley et al. (2022). We utilized the implementation from the same GitHub repository as GLOBE-CE (`https://github.com/danwley/GLOBE-CE`).

AReS relies on several hyperparameters that directly impact the trade-offs between coverage, cost, and interpretability of the generated rule sets. These parameters include the weights assigned to coverage and cost in the optimization function, as well as constraints on the maximum number of rules and rule width. The selection of these parameters was guided by some initial tests, where we aimed to maintain interpretability by limiting the rule complexity while ensuring sufficient coverage across affected subgroups.

k-means

As part of the Augmented Space algorithm, we used the k-means clustering algorithm (MacQueen et al., 1967). The k-means algorithm requires three parameters: N_CLUSTERS, which defines the number of clusters or centroids to form; N_INIT, which sets the number of initializations with different centroid seeds; and RANDOM_SEED, ensuring reproducibility. We used the `scikit-learn` implementation of k-means, for which details can be found in `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`, setting N_CLUSTERS according to the final cluster counts in the Augmented Space algorithm. We set N_INIT to 10 and RANDOM_SEED to 13.

DiCE

Finally, we employed the DiCE algorithm (Mothilal et al., 2019) for generating local counterfactuals as part of the Augmented Space method. DiCE requires several key parameters, such as DATASET (the dataset being used), MODEL (the trained ML model), and CONTINUOUS_FEATURES (a list of continuous features in the dataset). The target variable is defined via the OUTCOME_NAME parameter, and the model backend is specified using the BACKEND parameter. Lastly, the METHOD parameter defines the approach for generating counterfactuals.

We used the `dice-ml` Python library, available at `https://interpret.ml/DiCE/`. In our experiments, BACKEND was set to "sklearn" to match our use of `scikit-learn` models. We used the "random" setting for METHOD to enable Randomized Sampling. The DATASET and MODEL were specified according to the dataset and model for each run, and CONTINUOUS_FEATURES and OUTCOME_NAME were defined on a per-dataset basis.

Action Costs

For the computation of an **action's cost**, we follow the paradigm outlined in Ley et al. (2023) and again in Kavouras et al. (2024), by binning continuous features into 10 equal intervals

post-training and scaling the cost of each change proportionally to the bin length. We also set the cost of moving from one category to another for the categorical features to be 1. To ensure the **reproducibility** of our results, a consistent random seed of 13 is applied across all models and algorithms.

3.3 Experimental Evaluation

Following the discussion of the experimental setup in Section 3.2, we now present and analyze the results obtained by the Augmented Space algorithm and compare them with state-of-the-art methods for global counterfactual explainability (GLOBE-CE and its variants, CET, etc.). The experiments described in this section were conducted on all the combinations of the datasets and models we discussed previously, allowing us to assess the performance of these algorithms in real-world scenarios.

We focus on evaluating the results based on the key criteria of effectiveness, cost, number of actions, and runtime, providing a comprehensive comparison with GLOBE-CE and its variants. We summarize our findings across the various dataset-model combinations and highlight important observations drawn from these results.

3.3.1 Results Overview

For the global counterfactual explainability problem 2.3, we assess, just like Kavouras et al. (2024) does, Algorithm 1 (Augmented Space) against multiple configurations state-of-the-art methods in Global Counterfactual Explanations:

- AReS, using the Fast-AReS implementation of Ley et al. (2022)
- CET (Kanamori et al., 2022)
- GroupCF (Warren et al., 2023)
- GLOBE-CE and its variant dGLOBE-CE (Ley et al., 2023)

All algorithms are constrained with a predefined maximum action set size.

Table 3.6 summarizes the results of all competing methods. We evaluate the Augmented Space algorithm against five alternative approaches across five datasets and three models, yielding 75 head-to-head comparisons in total. However, due to CET's inability to solve the underlying optimization problem for the Adult dataset across all models, the final count includes 72 comparisons. All results pertain to Problem 2.6 with a constrained size of $\mathsf{size}(S) = 4$ to ensure a fairer comparison across methods.

Pareto Dominance We assess method performance by evaluating whether one solution dominates another in terms of effectiveness, cost, and size. Specifically, *a solution S to Problem 2.6* Pareto dominates *another solution S' if it matches or exceeds S' in effectiveness and cost, has an equal or smaller size, and is strictly superior in at least one of these objectives.*[5] Table 3.7 indicates that Augmented Space dominates other methods in 34 out of 72 cases (47%), and is dominated only once by GroupCF (in COMPAS-LR, see Table 3.6) and once by dGLOBE-CE (in HELOC-DNN). Notably, in these two instances, the effectiveness of Augmented Space is very close to that of the dominating methods.

[5] Pareto dominance is a well-established concept in multiobjective optimization. For further detail, see Branke (2008).

Table 3.6: Comparison of effectiveness and cost of the Augmented Space algorithm with Fast AReS, CET, Group-CF, GLOBE-CE, and dGLOBE-CE for solving problem 2.6 under the constraint $\mathsf{size}(S) = 4$. Solutions with effectiveness below the 80% applicability threshold are marked in purple. Non-robust GCEs — those with either an effectiveness standard deviation (std) over 5% across folds or a cost std greater than half the mean cost — are marked in red.

Models	Algorithms	Datasets									
		Adult		COMPAS		Default Credit		German Credit		HELOC	
		eff	cost	eff	cost	eff	cost	eff	cost	eff	cost
dnn	Fast AReS	12.39 ± 1.06	1.0 ± 0.0	55.0 ± 0.86	1.21 ± 0.09	18.88 ± 2.16	1.0 ± 0.0	52.39 ± 1.63	1.0 ± 0.0	12.19 ± 0.58	1.03 ± 0.05
	CET	NaN	NaN	63.62 ± 10.35	0.96 ± 0.24	98.87 ± 0.62	6.32 ± 2.28	97.3 ± 2.46	1.58 ± 0.54	86.78 ± 10.62	8.67 ± 3.25
	GroupCF	100.0 ± 0.0	10.08 ± 0.03	100.0 ± 0.0	4.48 ± 2.53	79.6 ± 20.79	1.53 ± 0.62	97.8 ± 4.4	1.85 ± 0.13	80.4 ± 10.17	3.09 ± 0.91
	GLOBE-CE	99.92 ± 0.0	4.24 ± 0.42	100.0 ± 0.0	4.54 ± 3.31	76.94 ± 37.55	5.14 ± 0.35	93.31 ± 3.48	2.0 ± 1.55	42.72 ± 46.97	11.77 ± 15.87
	dGLOBE-CE	99.92 ± 0.0	10.89 ± 1.37	100.0 ± 0.0	7.96 ± 3.91	87.38 ± 18.69	5.96 ± 4.14	97.36 ± 0.82	2.49 ± 0.27	99.96 ± 0.05	11.07 ± 8.6
	Augmented Space	100.0 ± 0.0	5.92 ± 0.66	100.0 ± 0.0	4.24 ± 1.83	99.62 ± 0.29	1.36 ± 0.33	98.11 ± 1.68	1.05 ± 0.04	99.36 ± 0.86	11.96 ± 2.44
lr	Fast AReS	11.74 ± 2.4	1.0 ± 0.0	62.5 ± 1.82	1.24 ± 0.14	10.85 ± 5.45	1.07 ± 0.13	75.27 ± 2.96	1.0 ± 0.0	9.23 ± 1.24	1.12 ± 0.1
	CET	NaN	NaN	73.18 ± 4.34	1.24 ± 0.15	100.0 ± 0.0	3.79 ± 1.31	96.5 ± 2.85	2.42 ± 0.24	100.0 ± 0.0	3.57 ± 1.48
	GroupCF	100.0 ± 0.0	1.71 ± 0.39	100.0 ± 0.0	3.97 ± 2.38	95.4 ± 9.2	1.94 ± 1.2	97.6 ± 2.94	9.34 ± 3.85	90.6 ± 3.93	2.4 ± 1.38
	GLOBE-CE	99.92 ± 0.0	2.68 ± 0.17	95.74 ± 8.52	5.14 ± 3.77	99.94 ± 0.07	3.42 ± 1.99	57.09 ± 20.03	0.75 ± 1.04	99.9 ± 0.0	0.6 ± 0.54
	dGLOBE-CE	99.92 ± 0.0	5.91 ± 0.93	100.0 ± 0.0	6.71 ± 0.23	99.94 ± 0.07	10.38 ± 7.76	69.89 ± 15.35	2.47 ± 0.23	99.9 ± 0.0	1.63 ± 0.35
	Augmented Space	100.0 ± 0.0	1.47 ± 0.44	100.0 ± 0.0	5.49 ± 0.92	100.0 ± 0.0	1.22 ± 0.26	99.55 ± 0.91	1.64 ± 0.32	100.0 ± 0.0	2.37 ± 1.03
xgb	Fast AReS	6.13 ± 0.42	1.0 ± 0.0	59.83 ± 3.12	1.1 ± 0.05	31.86 ± 5.12	1.05 ± 0.04	51.27 ± 1.57	1.0 ± 0.0	8.49 ± 1.32	1.16 ± 0.13
	CET	NaN	NaN	58.4 ± 9.3	1.06 ± 0.24	86.29 ± 9.94	4.5 ± 2.64	100.0 ± 0.0	2.73 ± 0.49	86.78 ± 6.7	12.51 ± 2.75
	GroupCF	96.8 ± 1.72	1.41 ± 0.54	100.0 ± 0.0	4.06 ± 2.1	95.2 ± 1.6	1.41 ± 0.64	100.0 ± 0.0	5.78 ± 4.11	78.4 ± 5.82	5.63 ± 1.93
	GLOBE-CE	82.87 ± 12.14	30.1 ± 10.39	87.13 ± 11.14	9.75 ± 7.2	82.7 ± 7.26	20.82 ± 1.73	77.05 ± 11.26	1.14 ± 1.24	27.66 ± 5.06	12.52 ± 32.48
	dGLOBE-CE	93.76 ± 1.98	64.76 ± 1.29	99.84 ± 0.31	12.46 ± 3.42	97.47 ± 0.82	42.58 ± 3.57	86.96 ± 9.79	2.66 ± 0.77	77.64 ± 11.51	128.0 ± 0.0
	Augmented Space	99.78 ± 0.14	6.15 ± 3.05	98.77 ± 1.17	3.54 ± 1.04	95.33 ± 1.44	2.52 ± 0.55	99.43 ± 1.14	1.06 ± 0.08	98.33 ± 1.42	23.94 ± 1.66

Table 3.7: Evaluation of Pareto dominance among solutions for Problem 2.6 with the constraint $\mathsf{size}(S) = 4$. The table shows the frequency (number of instances over possible comparisons) at which the Augmented Space algorithm dominates competitors (listed in the dominates column) and the frequency at which it is dominated by competitors (listed in the is dominated column).

	Augmented Space	
	Dominates	is Dominated
Fast AReS	0/15	0/15
CET	6/12	0/12
GroupCF	8/15	1/15
GLOBE-CE	9/15	0/15
dGLOBE-CE	11/15	1/15
competitors	34/72	2/72

Solution Practicality In the Pareto dominance evaluation (Table 3.7), we compare solutions with optimal or near-optimal effectiveness to many with notably lower effectiveness. These lower-performing solutions are impractical for the role of global counterfactual explanations, as the aim is to provide recourse to a substantial portion of the population. Solutions with low effectiveness fall short of this objective, diminishing their relevance for real-world applications. If a solution leaves a large segment of individuals without viable options for recourse, it fails the fundamental purpose of GCEs, as highlighted also by Ley et al. (2023). Additionally, achieving low recourse costs is more feasible for smaller subgroups, especially those closer to the decision boundary. We classify a solution as practical if it achieves at least 80% effectiveness. In Table 3.6, impractical solutions are marked in purple. Notably, Augmented Space never returns impractical solutions, while all solutions

Table 3.8: Evaluation of Pareto dominance among solutions for Problem 2.6 with constrained $\mathsf{size}(S) = 4$, applying the $\mathsf{eff} > 80\%$ threshold. The table shows the frequency (number of instances across comparisons) at which Augmented Space dominates other methods (in the DOMINATES column) and the frequency at which Augmented Space is outperformed by competitors (in the IS DOMINATED column).

	Augmented Space	
	Dominates	is Dominated
Fast AReS	0/0	0/0
CET	6/9	0/9
GroupCF	7/13	1/13
GLOBE-CE	7/10	0/10
dGLOBE-CE	9/13	1/13
Competitors	29/45	2/45

generated by Fast AReS are classified as impractical. The other methods produce between 2–5 impractical solutions each. Table 3.8 presents head-to-head comparisons limited only to practical solutions. Augmented Space dominates other methods in a significantly higher proportion — 29 out of 45 cases, or over 64%. It remains dominated in the same two cases, keeping the rate of dominated instances at a low 4%.

Robustness. Robustness is essential for GCE methods, as they should *consistently* produce highly effective, low-cost solutions across various data splits, ensuring reliability for real-world deployment. A lack of robustness can result in significant variations in recourse actions, reducing trust and potentially leading to unfair outcomes, especially in critical fields like healthcare and finance. To assess practical applicability, it is important to evaluate the stability of effectiveness and cost metrics across different folds. Standard deviation serves as a measure of this stability: an effectiveness deviation above 5% signals inconsistency in providing reliable recourse, while a cost deviation greater than half of the average cost suggests impractical, unpredictable actions. Such variability renders a solution unreliable, and we have highlighted these cases in red in Table 3.6.

3.3.2 Other Key Insights and Comparisons

Performance on Smaller Datasets

On smaller datasets like COMPAS and German Credit, where fewer actions are required for recourse, the Augmented Space algorithm performs comparably or better in terms of cost and effectiveness in comparison to the other methods[6]. We go back to the results of Table 3.6 to see this in more detail.

On the German Credit dataset, Augmented Space always achieves nearly 100% effectiveness, while the rest of the methods, except for GroupCF, all perform below practicality levels in at least one model. GroupCF is the only method that always achieves comparable effectiveness, but at a significantly greater cost.

In the COMPAS dataset both AReS and CET always achieve a much lower effectiveness

[6] sole exception is the COMPAS-LR combination where GroupCF has slightly better cost (3.97, versus 5.49 of Augmented Space). However, this low cost comes with a high standard deviation (2.38), so Augmented Space's solution is superior in terms of robustness, and we believe this makes it a better solution, even if only qualitatively.

(below 75%). As for the others, their costs are always significantly higher, except only for GroupCF on the LR model, where, however, the cost's standard deviation is significantly high, which means the solution loses in terms of robustness.

Performance on Larger Datasets

On larger datasets such as Default Credit and HELOC, the scalability of the Augmented Space algorithm becomes more evident (although with slightly longer runtimes). Referring again to Table 3.6, in the Default Credit dataset, Augmented Space is considerably better in either effectiveness or cost in almost all cases.

In the HELOC dataset, there is a case where Augmented Space is dominated by dGLOBE-CE (DNN model) and one case where dGLOBE-CE has only slightly lower effectiveness but considerably lower cost (LR model). However, apart from these two cases, Augmented Space is better by far on all other comparisons. Additionally, in the DNN case, the dominance is borderline. Augmented Space is extremely close in *both* effectiveness and cost, and additionally it has a much smaller standard deviation than dGLOBE-CE in the cost. Essentially, only in the LR model it is clear that the solution of dGLOBE-CE is better than the one of Augmented Space, albeit only slightly.

For the Adult dataset, similar patterns emerge. In one instance, GLOBE-CE demonstrates slightly lower effectiveness with a slightly smaller cost (in the DNN model). The relative cost difference is modest, suggesting that this may simply represent two reasonable trade-off points between effectiveness and cost. Another case which we should perhaps comment on[7] arises with the XGB model, where the GroupCF method produces a solution with 96.8% effectiveness and a cost of 1.41, compared to the Augmented Space solution's 99.78% effectiveness at a higher cost of 6.15. While neither solution dominates the other, one might argue that the cost increase could be considered excessive for a modest 3% gain in effectiveness. Nevertheless, we believe this scenario highlights GLANCE's intended focus: emphasizing that in cases without clear dominance, selecting between solutions may require balancing nuanced trade-offs, and clear selection may even be impossible.

It is important to note that, aside from these two cases, the Augmented Space algorithm decisively outperforms - by far - all other methods on the Adult dataset.

Discussion

The results demonstrate that the Augmented Space algorithm *consistently* provides a favorable balance between cost and effectiveness. This is demonstrated by the numerous cases in which it outperforms other methods.

However, we believe the most distinctive strength of the Augmented Space algorithm — and the GLANCE framework overall — lies in its remarkable *consistency*: it performs well to excellently across *all* of our experiments. This level of reliability is unique to Augmented Space and GLANCE, as other methods exhibited at least 2–3 instances where their solutions significantly underperformed or failed to meet practical thresholds.

[7] We do not actually believe so, but a reader may - quite reasonably nonetheless - argue otherwise. Additionally, we wanted to isolate these two cases in order to be able to claim truthfully that, in all other cases, not only is AS better, but also by a large margin.

Chapter 4

Auditing Fairness through Explainability Methods

4.1 Effectiveness-Cost TradeOff

We want to examine how recourse is achieved for the group G through a set of possible actions A. We define the **aggregate effectiveness** (aeff) of A for G in two distinct ways.

In the *micro viewpoint*, the individuals in the group are considered independently, and each may choose the action that benefits itself the most. Concretely:

Definition 4.1 (Micro Effectiveness). *We define the **micro-effectiveness** of set of actions A for group G as the proportion of individuals in G that can achieve recourse through* some *action in A, i.e.,:*

$$\mathsf{aeff}_{\mu}(A, G) = \frac{1}{|G|}|\{x \in G | \exists a \in A, \mathsf{eff}(a, x) = 1\}|.$$

In contrast, in the *macro viewpoint*, the group is considered as a whole, and an action is applied collectively to all individuals in the group. Concretely:

Definition 4.2 (Macro Effectiveness). *We define the **macro-effectiveness** of set of actions A for group G as the largest proportion of individuals in G that can achieve recourse through* the same *action in A, i.e.,:*

$$\mathsf{aeff}_{\mathsf{M}}(A, G) = \max_{a \in A} \frac{1}{|G|}|\{x \in G | \, \mathsf{eff}(a, x) = 1\}|.$$

Additionally, for a group G, actions A, and a cost budget c, we define the **in-budget actions** as the set of actions whose cost does not exceed c for any individual in G, Formally, this is expressed as

$$A_c = \{a \in A | \forall x \in G, \mathsf{cost}(a, x) \leq c\}$$

Using this definition, we can introduce the effectiveness-cost distribution, which provides a more detailed view of the actions by evaluating them across varying cost budgets, rather than just considering the aggregate effectiveness of the whole set. Specifically, it allows us to assess the effectiveness of actions for every possible cost threshold. Concretely:

Definition 4.3 (Effectiveness-Cost Distribution). *We define the **effectiveness-cost distribution** (*ecd*) as the function that for a cost budget c returns the aggregate effectiveness possible with in-budget actions:*

$$\mathsf{ecd}(c; A, G) = \mathsf{aeff}(A_c, G)$$

We use ecd_{μ}, $\mathsf{ecd}_{\mathsf{M}}$ to refer to the micro and macro viewpoints of aggregate effectiveness, respectively.

```
If capital-gain=0,education-num=10,hours-per-week=FullTime,race=White:
    Protected Subgroup = 'Male', 8.40% covered
        Make hours-per-week=OverTime with effectiveness 8.92% and counterfactual cost=2.0
        Make education-num=11 with effectiveness 8.92% and counterfactual cost=3.0
        Make hours-per-week=BrainDrain with effectiveness 16.98% and counterfactual cost=4.0
        Make education-num=11,hours-per-week=OverTime with effectiveness 23.16% and counterfactual cost=5.0
        Make education-num=12 with effectiveness 23.16% and counterfactual cost=6.0
        Make education-num=12,hours-per-week=OverTime with effectiveness 31.56% and counterfactual cost=8.0
        Aggregated cost of the above recourses = 0.32
    Protected Subgroup = 'Female', 8.32% covered
        Make hours-per-week=OverTime with effectiveness 1.79% and counterfactual cost=2.0
        Make education-num=11 with effectiveness 1.79% and counterfactual cost=3.0
        Make hours-per-week=BrainDrain with effectiveness 3.27% and counterfactual cost=4.0
        Make education-num=11,hours-per-week=OverTime with effectiveness 4.46% and counterfactual cost=5.0
        Make education-num=12 with effectiveness 4.46% and counterfactual cost=6.0
        Make education-num=12,hours-per-week=OverTime with effectiveness 5.95% and counterfactual cost=8.0
        Aggregated cost of the above recourses = 0.06
    Bias against 'Female' due to Equal Effectiveness. Unfairness score = 0.26.
```

Figure 4.1: Example for demonstration of the "Equal Effectiveness" definition. Run for the Adult dataset with sex as the sensitive attribute. Male individuals achieve an aggregate effectiveness of 31.56% while female individuals only 5.95% with the same actions.

The value $\mathsf{ecd}(c; A, G)$ is the proportion of individuals in G that can achieve recourse through actions A with cost at most c. Therefore, the ecd function has an intuitive probabilistic intepretation. Consider the subspace X_p determined by predicate p, and define the random variable C as the cost required by an instance $x \in X_p$ to achieve recourse. The function $\mathsf{ecd}(c; A, G_p)$ is the empirical cumulative distribution function of C using sample G_p.

The **inverse effectiveness-cost distribution** function $\mathsf{ecd}^{-1}(\phi; A, G)$ takes as input a effectiveness level $\phi \in [0, 1]$ and returns the minimum cost required so that $\phi|G|$ individuals achieve recourse.

4.1.1 Definitions of Subgroup Recourse Fairness

We define recourse fairness of classifier h for a group G by comparing the ecd functions of the protected subgroups G_0, G_1 in different ways. The first two definitions are *cost-oblivious*, and simply compare the protected subgroups based on the aggregate effectiveness of a set of (common, as always) actions.

Definition 4.4 (Equal Effectiveness). *This definition has a micro and a macro interpretation, and says that the classifier is fair if the same proportion of individuals in the protected subgroups can achieve recourse:*

$$\mathsf{aeff}(A, G_0) = \mathsf{aeff}(A, G_1)$$

To demonstrate, consider the example subpopulation group and actions depicted in Figure 4.1. For the male persons within this group, the actions achieve an effectiveness percentage of 31.56%, while for the female persons with *the same defining characteristics* (the group's), the *same actions* achieve an effectiveness of just 5.95%.

This is why we deem that, in such a case, recourse is much easier for males than for females, and thus the model in question has potential bias and warrants further inspection.

Definition 4.5 (Equal Choice for Recourse). *This definition has only a macro interpretation, and claims that the classifier is fair if the protected subgroups can choose among the same number of sufficiently effective actions to achieve recourse, where sufficiently effective means the actions should work for at least* $100\phi\%$ *(for* $\phi \in [0, 1]$*) of the subgroup:*

$$|\{a \in A | \mathsf{eff}(a, G_0) \geq \phi\}| = |\{a \in A | \mathsf{eff}(a, G_1) \geq \phi\}|.$$

```
If capital-gain=0,marital-status=Divorced:
    Protected Subgroup = 'Asian-Pac-Islander', 12.95% covered
        Make capital-gain=15024,marital-status=Married-civ-spouse with effectiveness 100.00%.
        Make capital-gain=7688,marital-status=Married-civ-spouse with effectiveness 75.00%.
        Make capital-gain=7298,marital-statu=Married-civ-spouse with effectiveness 75.00%.
    Protected Subgroup = 'Amer-Indian-Eskimo', 20.37% covered
        Make capital-gain=15024,marital-status=Married-civ-spouse with effectiveness 100.00%.
        Make capital-gain=7688,marital-status=Married-civ-spouse with effectiveness 81.82%.
        Make capital-gain=7298,marital-status=Married-civ-spouse with effectiveness 72.73%.
    Protected Subgroup = 'Black', 15.15% covered
        Make capital-gain=15024,marital-status=Married-civ-spouse with effectiveness 98.90%.
    Protected Subgroup = 'Other', 12.97% covered
        Make capital-gain=15024,marital-status=Married-civ-spouse with effectiveness 100.00%.
        Make capital-gain=7688,marital-status=Married-civ-spouse with effectiveness 76.92%.
        Make capital-gain=7298,marital-status=Married-civ-spouse with effectiveness 69.23%.
    Protected Subgroup = 'White', 16.96% covered
        Make capital-gain=15024,marital-status=Married-civ-spouse with effectiveness 99.41%.
        Make capital-gain=7688,marital-status=Married-civ-spouse with effectiveness 73.90%.
        Make capital-gain=7298,marital-status=Married-civ-spouse with effectiveness 70.43%.
    Bias against 'Black' due to Equal Choice for Recourse. Unfairness score = 2.
```

Figure 4.2: Example for demonstration of the "Equal Choice for Recourse" definition. Run for the Adult dataset with race as the sensitive attribute. All subgroups have a choice of 3 actions to achieve recourse for at least $\phi = 70\%$ of their individuals; except for Black individuals, who only have one choice.

```
If PREDICTOR RAT ARRESTS VIOLENT OFFENSES=1,PREDICTOR RAT NARCOTIC ARRESTS=1,PREDICTOR RAT VICTIM BATTERY OR ASSAULT=1:
  Protected Subgroup = 'Black', 1.04% covered
        No recourses for this subgroup.
  Protected Subgroup = 'White', 1.00% covered
        Make PREDICTOR RAT ARRESTS VIOLENT OFFENSES=0,PREDICTOR RAT NARCOTIC ARRESTS=0,PREDICTOR RAT VICTIM BATTERY OR
             ASSAULT=0 with effectiveness 72.73%
        Make PREDICTOR RAT ARRESTS VIOLENT OFFENSES=0,PREDICTOR RAT NARCOTIC ARRESTS=1,PREDICTOR RAT VICTIM BATTERY OR
             ASSAULT=0 with effectiveness 72.73%
        Make PREDICTOR RAT ARRESTS VIOLENT OFFENSES=0,PREDICTOR RAT NARCOTIC ARRESTS=2,PREDICTOR RAT VICTIM BATTERY OR
             ASSAULT=0 with effectiveness 72.73%
  Bias against 'Black' due to Equal Cost of Effectiveness (threshold = 0.7). Unfairness score = inf.
```

Figure 4.3: Example for demonstration of the "Equal Cost of Effectiveness" definition. Run for the SSL dataset with race as the sensitive attribute. White individuals achieve an effectiveness of 70% with the 3 shown actions, the minimum cost of which is 1 according to our heuristic cost definition. On the contrary, Black individuals cannot, *with the same actions*, achieve 70% effectiveness at all.

As an example, observe the population group depicted in Figure 4.2. With a user-selected effectiveness threshold of 70%, all protected subgroups have 3 actions to choose from, except for the 'Black' subgroup which has only one. This is why we consider a violation of this definition as an indication that the model exhibits bias.

The subsequent three definitions assume the cost function is known, and have both a micro and a macro interpretation.

Definition 4.6 (Equal Cost of Effectiveness). *The classifier is fair if the minimum cost to achieve an aggregate effectiveness of $\phi \in [0, 1]$ in the protected subgroups is equal:*

$$\mathsf{ecd}^{-1}(\phi; A, G_0) = \mathsf{ecd}^{-1}(\phi; A, G_1)$$

As an example, consider the group depicted in Figure 4.3. In this example, we have set the effectiveness threshold at a reasonable level of $\phi = 70\%$. From the actions that we have generated, there are 3 that, if applied to White individuals, provide recourse to at least 70% of them, with a minimum cost of 1 (e.g. for the first action, which changes the 'PREDICTOR RAT NARCOTIC ARRESTS' feature from 1 to 0). On the contrary, for Black individuals there are no actions that achieve this level of effectiveness, so we assign to them an aggregate cost of ∞, and classify this specific subpopulation group as having 'infinite' bias score.

```
If juv-other-count=3:
    Protected Subgroup = 'Caucasian', 1.54% covered
        Make juv-other-count=0 with effectiveness 66.67% and counterfactual cost=3.0
        Aggregated cost of the above recourses = 0.67
    Protected Subgroup = 'African-American', 1.16% covered
        Make juv-other-count=0 with effectiveness 0.00% and counterfactual cost=3.0
        Aggregated cost of the above recourses = 0.00
    Bias against 'African-American' due to Equal Effectiveness within Budget(budget=3).Unfairness score=0.667.
```

Figure 4.4: Example for demonstration of the "Equal Effectiveness within Budget" definition. Run for the COMPAS dataset with race as the sensitive attribute. For a budget of e.g. $c = 4$, the 'Caucasian' subgroup attains 66.67% effectiveness, while the 'African-American' subgroup 0%.

Definition 4.7 (Equal Effectiveness within Budget). *The classifier is fair if the same proportion of individuals in the protected subgroups can achieve recourse with a cost at most c:*

$$\mathsf{ecd}(c; A, G_0) = \mathsf{ecd}(c; A, G_1)$$

We provide a corresponding example for this definition in Figure 4.4. There, one action is available for the protected subgroups. Using this action, which has cost 3 for all individuals in the specific population group, the 'Caucasian' subgroup achieves effectiveness 66.67%, while in the 'African-American' subgroup nobody achieves recourse! So, we should consider the specific population group as an indication of unfairness.

The way the definition would be applied in this case is that, for a budget of $c = 4$, for example, the individuals in the 'Caucasian' subgroup can "pay" the cost of 3 and achieve 66.67% effectiveness. On the contrary, 'African-American' individuals are stuck with 0% for any budget value. So, applying the definition 'Equal Effectiveness within Budget' would yield a negative answer, i.e. the classifier is not deemed to be necessarily fair for this subpopulation group.

Definition 4.8 (Fair Effectiveness-Cost Trade-Off). *The classifier is fair if the protected subgroups have the same effectiveness-cost distribution, or equivalently for each cost budget c, their aggregate effectiveness is equal:*

$$\max_c |\,\mathsf{ecd}(c; A, G_0) - \mathsf{ecd}(c; A, G_1)| = 0$$

Writting the definition in this form, we can see that the left-hand side represents the two-sample Kolmogorov-Smirnov statistic for the empirical cumulative distributions (ecd) of the protected subgroups. Based on this and (Knuth, 2014, Section 3.3.1), we formulate a statistical version of the above definition.

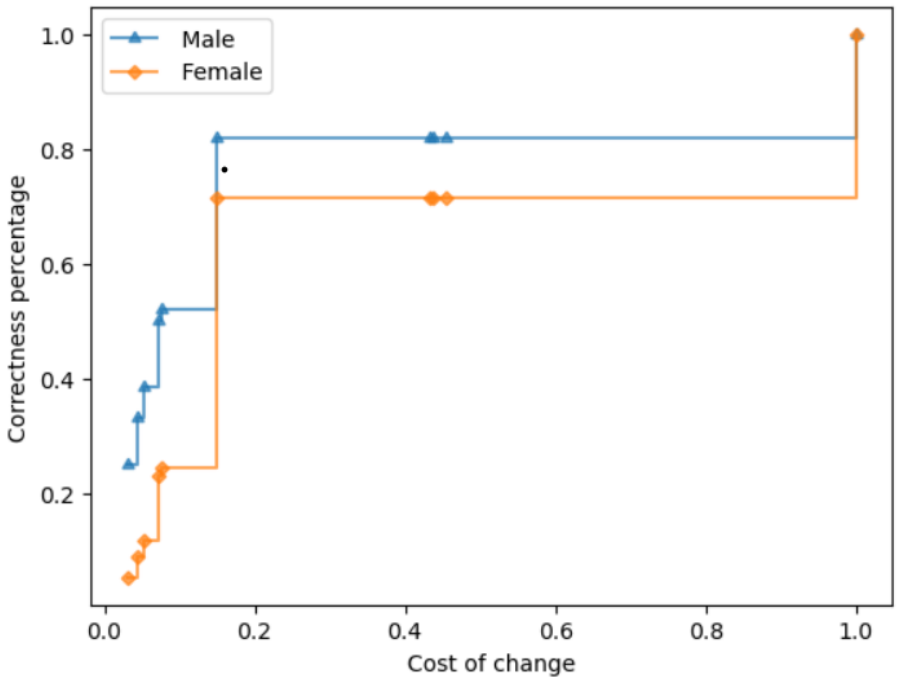
Definition 4.9 (Fair Effectiveness-Cost Trade-Off - Statistical). *We say that the classifier is fair with confidence α if the two-sample Kolmogorov-Smirnov test for the empirical cumulative distributions (ecd) of the protected subgroups does not reject its null hypothesis (that the distributions are different). Formally, when it holds that:*

$$\max_c |\,\mathsf{ecd}(c; A, G_0) - \mathsf{ecd}(c; A, G_1)| \leq \sqrt{-\ln(\alpha/2)\frac{|G_{p,0}| + |G_{p,1}|}{2|G_{p,0}||G_{p,1}|}}$$

In Figure 4.5 we demonstrate a counterexample for this definition. We show both the population group and its actions (in rule-set form) and the respective empirical cumulative distributions, for a better understanding. From the two ecd functions it is clear why we

```
If capital-gain=0,capital-loss=0:
    Protected Subgroup = 'Male', 93.81% covered
        Make capital-gain=3103 with effectiveness 25.06% and counterfactual cost=0.03.
        Make capital-gain=4386 with effectiveness 33.33% and counterfactual cost=0.04.
        Make capital-gain=5178 with effectiveness 38.53% and counterfactual cost=0.05.
        Make capital-gain=7298 with effectiveness 49.99% and counterfactual cost=0.07.
        Make capital-gain=7688 with effectiveness 52.13% and counterfactual cost=0.08.
        Make capital-gain=15024 with effectiveness 82.07% and counterfactual cost=0.15.
        Make capital-loss=1887 with effectiveness 82.07% and counterfactual cost=0.43.
        Make capital-loss=1902 with effectiveness 82.07% and counterfactual cost=0.44.
        Make capital-loss=1977 with effectiveness 82.07% and counterfactual cost=0.45.
        Make capital-gain=99999 with effectiveness 100.00% and counterfactual cost=1.0.
    Protected Subgroup = 'Female', 93.54% covered
        Make capital-gain=3103 with effectiveness 5.32% and counterfactual cost=0.03.
        Make capital-gain=4386 with effectiveness 8.90% and counterfactual cost=0.04.
        Make capital-gain=5178 with effectiveness 11.82% and counterfactual cost=0.05.
        Make capital-gain=7298 with effectiveness 22.92% and counterfactual cost=0.07.
        Make capital-gain=7688 with effectiveness 24.53% and counterfactual cost=0.08.
        Make capital-gain=15024 with effectiveness 71.48% and counterfactual cost=0.15.
        Make capital-loss=1887 with effectiveness 71.48% and counterfactual cost=0.43.
        Make capital-loss=1902 with effectiveness 71.48% and counterfactual cost=0.44.
        Make capital-loss=1977 with effectiveness 71.48% and counterfactual cost=0.45.
        Make capital-gain=99999 with effectiveness 100.00% and counterfactual cost=1.0.
    Bias against 'Female' due to Fair Effectiveness-Cost Trade-off. Unfairness score = 13.49.
```

(a) Comparative Subgroup Counterfactuals (output of the FACTS framework)



(b) Empirical cumulative cost distributions for the 'Male' and 'Female' protected subgroups

Figure 4.5: Example for demonstration of the "Fair Effectiveness-Cost Trade-Off" definition. Run for the Adult dataset with sex as the sensitive attribute. Between the empirical cumulative distributions of the 'Male' and 'Female' subgroup, their maximum absoulute difference is observed for cost 0.08 and is equal to 27.6%. Multiplied with the quantity $\sqrt{\frac{2|G_{p,0}||G_{p,1}|}{|G_{p,0}|+|G_{p,1}|}}$ it yields the 'unfairness score' of 13.49.

```
If ages = 55-64, income = <100K, religion = Unknown:
    Protected Subgroup = 'Male', 1.02% covered
        Make religion=Christianity with effectiveness 0.00%.
        Make religion=Other with effectiveness 0.00%.
    Protected Subgroup = 'Female', 1.08% covered
        Make religion=Christianity with effectiveness 9.86%.
        Make religion=Other with effectiveness 9.86%.
    Bias against Male due to Equal Conditional Mean Recourse. Unfairness score = inf.
```

Figure 4.6: Example for demonstration of the "Equal (Conditional) Mean Recourse" definition. Run for the Ad Campaign dataset with sex as the sensitive attribute. The 'Female' subgroup can achieve effectiveness 9.86% with cost 1, and it has no actions of different cost, thus the mean recourse cost is 1. In the 'Male' subgroup nobody can achieve recourse, so the mean recourse cost is infinite, in a sense.

would posit that the model exhibits bias against the 'Female' subgroup of this population group. The whole ecd of the 'Female' subgroup is lower than that of the 'Male' subgroup. This means that females are less likely to receive recourse when exercising actions with the same cost as the males. Put another way, *for any cost budget*, the females will achieve lower effectiveness.

This is the situation that the above definition tries to capture. The way it would be applied here is that we would identify the maximum absolute difference between the two distributions, which is 27.6%. Then, after determining a desired level of confidence α, we would check whether this maximum difference is less than the quantity $\sqrt{-\ln(\alpha/2)\frac{|G_{p,0}|+|G_{p,1}|}{2|G_{p,0}||G_{p,1}|}}$. If not, we would classify this group as receiving unfair treatment from the model, according to this definition.

The last definition takes a micro viewpoint and extends the notion of *burden* (Sharma et al., 2020) from existing literature, which is the mean recourse cost of an action over all individuals, to the case where not all individuals may achieve recourse.

Specifically, notice that the straightforward **mean recourse cost** of a group G,

$$\bar{\mathsf{rc}}(A, G) = \frac{1}{|G|} \sum_{x \in G} \mathsf{rc}(A, x)$$

considers individuals that cannot achieve recourse through A and have a recourse cost of c_∞. To exclude them, we denote as G^* the set of individuals of G that *can* achieve recourse through an action in A, i.e., $G^* = \{x \in G | \exists a \in A \text{ such that } h(a(x)) = 1\}$. Then, we say that the **conditional mean recourse cost** is the mean recourse cost among those that can achieve recourse:

$$\bar{\mathsf{rc}}^*(A, G) = \frac{1}{|G^*|} \sum_{x \in G^*} \mathsf{rc}(A, x).$$

If $G = G^*$, the definitions coincide with burden.

Definition 4.10 (Equal (Conditional) Mean Recourse). *The classifier is fair if the (conditional) mean recourse cost for the protected subgroups is the same:*

$$\bar{\mathsf{rc}}^*(A, G_0; A) = \bar{\mathsf{rc}}^*(A, G_1)$$

An example showcasing this definition is shown in Figure 4.6. In this specific population group, in contrast to previous examples, the model is indicated to exhibit bias against the 'Male' subgroup. The two actions discovered by our framework result in an effectiveness of

```
If hours-per-week=FullTime, marital-status=Married-civ-spouse, occupation=Adm-clerical:
    Protected Subgroup = 'Male', 1.87% covered
        Make hours-per-week=Overtime, occupation=Exec-managerial with effectiveness 72.00%
    Protected Subgroup = 'Female', 1.80% covered
        No recourses for this subgroup.
    Bias against 'Female' due to Equal Cost of Effectiveness (threshold=0.7). Unfairness score = inf.
```

Figure 4.7: CSC for a highly biased subgroup in terms of *Equal Cost of Effectiveness* with $\phi = 0.7$.

9.86% for the females, while for males 0%, i.e. nobody receives recourse. This is something detected by the equal mean recourse definition, as the mean recourse cost for females is 1 (we assign a cost of 1 to a change in a categorical feature), while for the males, who do not achieve recourse, is, in a sense, infinite.

As a final note, when the group G is the entire dataset of affected individuals, and all individuals can achieve recourse through A, i.e. $G = G^*$ this fairness notion coincides with fairness of burden (Ustun et al., 2019; Sharma et al., 2020).

4.2 Fairness of Recourse Auditing via Group Counterfactuals: FACTS

This section presents FACTS (Fairness-aware Counterfactuals for Subgroups) (Kavouras et al., 2023), a framework whose goal is the detection of population groups that are likely to elicit unfair treatment from the unknown predictive model in terms of difficulty of the group's subgroups to achieve recourse. To assess and compare groups, and to essentially quantify this 'unfairness of recourse', the framework implements both the micro and the macro viewpoint discussed earlier, and all respective fairness definitions provided in Section 4.1.1. The output of FACTS comprises population groups that are assigned (a) an unfairness score that captures the disparity between protected subgroups according to different fairness definitions and allows us to rank groups and (b) a user-intuitive, easily explainable counterfactual summary, which is termed *Comparative Subgroup Counterfactuals* (CSC).

Figure 4.7 presents an example result of FACTS derived from the adult dataset (Bache and Lichman, 2013). The "if clause" represents the subgroup G_p, which contains all the affected individuals that satisfy the predicate $p = (\textit{hours-per-week} = \textit{FullTime}) \wedge (\textit{marital-status} = \textit{Married-civ-spouse}) \wedge (\textit{occupation} = \textit{Adm-clerical})$. The information below the predicate refers to the protected subgroups $G_{p,0}$ and $G_{p,1}$ which are the female and male individuals of G_p respectively. With blue color, we highlight the percentage $\mathsf{cov}(G_{p,i}) = |G_{p,i}|/|D_i|$ which serves as an indicator of the size of the protected subgroup. The most important part of the representation is the actions applied that appear below each protected subgroup and are evaluated in terms of fairness metrics. In this example, the metric is the *Equal Cost of Effectiveness* with effectiveness threshold $\phi = 0.7$. For $G_{p,0}$ there is no action surpassing the threshold $\phi = 0.7$, therefore we display a message accordingly. On the contrary, the action $a = \{\textit{hours-per-week} \rightarrow \textit{Overtime}, \textit{occupation} \rightarrow \textit{Exec-managerial}\}$ has effectiveness $0.72 > \phi$ for $G_{p,1}$, thus allowing a respective 72% of the male individuals of G_p to achieve recourse. The unfairness score is thus "inf", since no recourse is achieved for the female subgroup.

Method overview Deploying FACTS comprises three main steps: (a) *Subgroup and action space generation*, where frequent subgroups defined as frequent sets of feature-operator-value combinations in the affected population are derived and, respectively, frequent ac-

tions are collected in similar way from the unaffected population; (b) *Counterfactual summaries generation*, where proper matching of subgroups and actions is performed to produce valid counterfactual summaries for each subgroup and (c) *CSC construction and fairness ranking*, where each counterfactual summary is scored according to each definition from Section 4.1.1, producing a respective number of CSCs for each subgroup and allowing to produce separate subgroup rankings per definition. Next, we describe these steps in detail.

(a) Subgroup and action space generation Subgroups are generated by executing the fp-growth (Han et al., 2000) frequent itemset mining algorithm on D_0 and on D_1 resulting in the sets of subgroups $\mathcal{G}_0$ and $\mathcal{G}_1$ and then by computing the intersection $\mathcal{G} = \mathcal{G}_0 \bigcap \mathcal{G}_1$. In our setting, an item is a feature-level predicate of the form "feature-operator-value" and, consequently, an itemset is a predicate p defining a subgroup G_p. This step guarantees that the evaluation in terms of fairness will be performed between the common subgroups $\mathcal{G}$ of the protected populations. The set of all actions A is generated by executing fp-growth on the unaffected population to increase the chance of more effective actions and to reduce the computational complexity. The above process is parameterizable w.r.t. the selection of the protected attribute(s) and the minimum frequency threshold for obtaining candidate subgroups.

(b) Counterfactual summaries generation For each subgroup $G_p \in \mathcal{G}$, the following steps are performed: (i) Find valid actions, i.e., the actions in A that contain a subset of the features appearing in p and at least one different value in these features; (ii) For each valid action a compute $\mathsf{eff}(a, G_{p,0})$ and $\mathsf{eff}(a, G_{p,1})$.

The aforementioned process extracts, for each subgroup G_p, a subset V_p of the actions A, with each action having exactly the same cost for all individuals of G_p. Therefore, individuals of $G_{p,0}$ and $G_{p,1}$ are evaluated in terms of subgroup-level actions, with a fixed cost for all individuals of the subgroup, in contrast to methods that rely on aggregating the cost of individual counterfactuals. This approach provides a key advantage to FACTS in cases where the definition of the exact costs for actions is either difficult or ambiguous: a misguided or even completely erroneous attribution of a cost to an action will equally affect all individuals of the subgroup and only to the extent that the respective fairness definition allows it. In the setting of individual counterfactual cost aggregation, changes in the same sets of features could lead to highly varying action costs for different individuals within the same subgroup.

(c) CSC construction and fairness ranking At this stage, FACTS evaluates all definitions of Section 4.1.1 on all subgroups, producing an unfairness score per definition, per subgroup. In particular, each definition of Section 4.1.1 quantifies a different aspect of the difficulty for a protected subgroup to achieve recourse. This quantification directly translates to difficulty scores for the protected subgroups $G_{p,0}$ and $G_{p,1}$ of each subgroup G_p, which are then compared accordingly (computing the absolute difference between them) to arrive at the unfairness score of each G_p based on the particular fairness metric.

The outcome of this process is the generation, for each fairness definition, of a ranked list of CSC representations, in decreasing order of their unfairness score. Apart from unfairness ranking, the CSC representations will allow users to intuitively understand unfairness by directly comparing differences in actions between the protected populations within a subgroup.

4.3 Experimental Setting

In this section, we describe the experimental setup used in the subsequent experiments. Readers interested solely in the experimental results may skip ahead to Section 4.4.

The experimental setup follows the methodology outlined by Kavouras et al. (2023). We organize the presentation as follows:

- The datasets we used are presented in Subsection 4.3.1.
- The models are presented in Subsection 4.3.2.
- Other settings for the experiments, such as train-test splits, cost functions used etc. are presented in Subsection 4.3.3.

4.3.1 Datasets Description

In line with the original work of Kavouras et al. (2023), we use four datasets in our experimental evaluation. For each dataset, we detail the preprocessing steps, specify the types of features used, and provide the cost feature weights applied during the experiments.

Adult

We generated CSCs for the Adult dataset[1] using two different features as protected attributes, i.e., 'sex', and 'race'. The assessment of bias for each protected attribute is done separately. The results for 'sex' as the protected attribute are presented in the main paper. Before we present our results for race as the protected attribute, we briefly discuss the preprocessing procedures and feature weights used for the adult dataset.

We generated Comparative Subgroup Counterfactuals (CSCs) for the Adult dataset[2] using two different protected attributes: 'sex' and 'race'. The bias assessment for each protected attribute was conducted separately. Before presenting our findings, we briefly outline the preprocessing steps and feature weights applied for the Adult dataset.

Preprocessing We removed the 'fnlwgt' and 'education' features, along with any rows containing unknown values. The 'hours-per-week' and 'age' features were discretized into five bins each.

Features All features were treated as categorical, except for 1capital-gain' and 'capital-loss', which were treated as numeric, and 'education-num' and 'hours-per-week', which were treated as ordinal. The feature weights used in the cost function are shown in Table 4.1. These weights are intended as indicative assignments for the purpose of experimentation. They aim to reflect the feasibility or actionability of making changes to specific features.

COMPAS

We generated Comparative Subgroup Counterfactuals (CSCs) for the COMPAS dataset[3], using 'race' as the protected attribute. Before presenting the results, we provide a brief overview of the preprocessing steps and the cost feature weights used for this dataset.

[1] https://raw.githubusercontent.com/columbia/fairtest/master/data/adult/adult.csv

[2] https://raw.githubusercontent.com/columbia/fairtest/master/data/adult/adult.csv

[3] https://aif360.readthedocs.io/en/latest/modules/generated/aif360.sklearn.datasets.fetch_compas.html

Table 4.1: Cost Feature Weights for Adult

feature name	weight value	feature name	weight value
native-country	4	Workclass	2
marital-status	5	hours-per-week	2
relationship	5	capital-gain	1
age	10	capital-loss	1
occupation	4	education-num	3

Preprocessing The features 'age' and 'c_charge_desc' were discarded. The 'priors_count' feature was discretized into five bins: [-0.1, 1), [1, 5), [5, 10), [10, 15), and [15, 38), aiming to keep the frequency of each bin approximately equal. Due to the highly asymmetric distribution of values, this was not fully achievable using direct methods such as `pandas.qcut`[4].

Features The features 'juv_fel_count', 'juv_misd_count', and 'juv_other_count' were treated as numerical, while the rest were considered categorical. The feature weights applied in the cost function are shown in Table 4.2.

Table 4.2: Cost Feature Weights for COMPAS

feature name	weight value
age_cat	10
juv_fel_count	1
juv_fel_count	1
juv_other_count	1
priors_count	1
c_charge_degree	1

SSL

We generated Comparative Subgroup Counterfactuals (CSCs) for the SSL dataset[5], using 'race' as the protected attribute. Before presenting the results, we briefly describe the preprocessing steps and feature weights applied, following the same procedures outlined in the original FACTS paper (Kavouras et al., 2023).

Preprocessing We removed all rows containing missing values ('U' or 'X'). Additionally, the feature 'PREDICTOR RAT TREND IN CRIMINAL ACTIVITY' was discretized into six bins. Since the target labels range between 0 and 500, we followed the approach used in Black et al. (2020), binarizing the labels by treating values above 344 as positively impacted and values below 345 as negatively impacted.

Features All features, except for the protected attribute 'race', were treated as numerical. The feature weights used for the cost function are presented in Table 4.3.

[4] `https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.qcut.html`

[5] `https://raw.githubusercontent.com/samuel-yeom/fliptest/master/exact-ot/chicago-ssl-clean.csv`

Table 4.3: Cost Feature Weights for SSL

feature name	weight value
PREDICTOR RAT AGE AT LATEST ARREST	10
PREDICTOR RAT VICTIM SHOOTING INCIDENTS	1
PREDICTOR RAT VICTIM BATTERY OR ASSAULT	1
PREDICTOR RAT ARRESTS VIOLENT OFFENSES	1
PREDICTOR RAT GANG AFFILIATION	1
PREDICTOR RAT NARCOTIC ARRESTS	1
PREDICTOR RAT TREND IN CRIMINAL ACTIVITY	1
PREDICTOR RAT UUW ARRESTS	1

Ad Campaign

We also generated CSCs for the Ad Campaign dataset[6], using 'gender' as the protected attribute.

Preprocessing Unlike the other datasets, we opted not to remove missing values, as they represent the majority for several features. However, we restricted actions in the CSCs to avoid any that lead to missing values in the final representation.

Features In this dataset, all features, except the protected attribute, were treated as categorical. The feature weights used in the cost function are detailed in Table 4.4.

Table 4.4: Cost Feature Weights for Ad Campaign

feature name	weight value
religion	5
politics	2
parents	3
age	10
income	3
area	2
college_educated	3
homeowner	1

4.3.2 Models

For our experiments, we used a Logistic Regression model, and specifically the standard implementation provided by the Python package `scikit-learn`[7]. This model serves as the black-box classifier that is audited for fairness of recourse in our framework, similar to the approach used in Kavouras et al. (2023).

[6] https://developer.ibm.com/exchanges/data/all/bias-in-advertising/

[7] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

4.3.3 Additional Experimental Specifications

Train-Test Split All datasets were split into training and test sets, with 70% of the data used for training and 30% for testing. The split process included both shuffling and stratification based on the labels to ensure balanced class distribution across the splits. The random seed value used for reproducibility was set to 131313, and the data splitting was done using the `train_test_split`[8] function from `scikit-learn`. Following the methodology of Kavouras et al. (2023), all audits were performed on the test set.

Frequent Itemset Mining Subgroups and actions were generated using the fp-growth algorithm[9] for frequent itemset mining, specifically the implementation provided by the Python package `mlxtend`[10]. We set the support threshold to **1%**, meaning that we required subgroups and actions to appear in at least 1% of the respective populations. Subgroups were derived from the affected populations D_0 and D_1, while actions were derived from the unaffected population, following the same approach as in Kavouras et al. (2023).

Effectiveness and Budgets As outlined in Section 4.1.1, some of the fairness metrics — *Equal Choice for Recourse* and *Equal Cost of Effectiveness* — require the specification of a target effectiveness level, denoted as ϕ. Additionally, the *Equal Effectiveness within Budget* metric necessitates the definition of a target cost level (or budget) c. We replicate the approach of Kavouras et al. (2023) for both cases.

Regarding effectiveness-based metrics, we set two effectiveness levels: a relatively low effectiveness level of $\phi = 30\%$ and a relatively high effectiveness level of $\phi = 70\%$.

For the budget-based metrics, we used the more detailed procedure outlined in Kavouras et al. (2023) to determine cost levels:

1. First, we compute the *Equal Cost of Effectiveness* (micro definition) with a target effectiveness of $\phi = 50\%$ to calculate the minimum cost required to flip at least 50% of both protected subgroups G_0 and G_1, for all population groups G.
2. We gather the minimum costs for all groups into an array.
3. Finally, we selected budget values corresponding to the **30%**, **60%**, and **90%** percentiles of this array.

Cost Functions Our implementation allows users to define custom cost functions based on domain knowledge and specific requirements. For the purposes of evaluation and demonstration, we defined the cost of changing a feature value v to a new value v' as follows:

1. **Numerical features**: $|norm(v) - norm(v')|$, where $norm$ is a normalization function that maps values to the range $[0, 1]$.
2. **Categorical features**: a cost of 1 if $v \neq v'$, and 0 otherwise.
3. **Ordinal features**: $|pos(v) - pos(v')|$, where pos is a function that assigns the order to the values.

[8] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

[9] `https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/fpgrowth/`

[10] `https://github.com/rasbt/mlxtend`

In addition to these basic costs, users of FACTS can specify feature-specific weights to represent the difficulty of changing a particular feature. For each dataset, the total cost of an action is determined by specifying the feature types (numerical, categorical, or ordinal) and the associated weights.

Feasibility To ensure realistic and feasible actions, we incorporated certain constraints in our cost model. For example:

- The age and education features cannot be reduced, ensuring that actions do not suggest unrealistic changes.
- Actions that result in unknown or missing values are disallowed.

These constraints help maintain the validity of the suggested actions.

Compute resources The experiments were run on commodity hardware featuring an AMD Ryzen 5 5600H processor and 8GB of RAM. All computations were performed within an isolated conda environment using Python 3.9.16.

4.4 Experimental Evaluation

This section presents key experimental results from Kavouras et al. (2023), as well as results from additional experimentation we conducted for the purposes of this thesis. We present and discuss Comparative Subgroup Counterfactuals for subgroups ranked as the most unfair according to the various definitions of Section 4.1.1. For this section we have selected specific datasets and models as illustrative examples. For a more exhaustive collection of experimental results, we defer the interested reader to Appendix A.

Table 4.5: Unfair subgroups identified in the Adult dataset.

	Group 1			Group 2			Group 3		
	rank	bias against	unfairness score	rank	bias against	unfairness score	rank	bias against	unfairness score
Equal Effectiveness	2950	Male	0.11	10063	Female	0.0004	275	Female	0.32
Equal Choice for Recourse ($\phi = 0.3$)	Fair	-	0	12	Female	2	Fair	-	0
Equal Choice for Recourse ($\phi = 0.7$)	6	Male	1	1	Female	6	Fair	-	0
Equal Effectiveness within Budget ($c = 5$)	Fair	-	0	2806	Female	0.056	70	Female	0.3
Equal Effectiveness within Budget ($c = 10$)	2350	Male	0.11	8518	Female	0.0004	226	Female	0.3
Equal Effectiveness within Budget ($c = 18$)	2675	Male	0.11	9222	Female	0.0004	272	Female	0.3
Equal Cost of Effectiveness ($\phi = 0.3$)	Fair	-	0	Fair	-	0	1	Female	inf
Equal Cost of Effectiveness ($\phi = 0.7$)	1	Male	inf	12	Female	2	Fair	-	0
Fair Effectiveness-Cost Trade-Off	4065	Male	0.11	3579	Female	0.13	306	Female	0.32
Equal (Conditional) Mean Recourse	Fair	-	0	3145	Female	0.35	Fair	-	0

Unfair subgroups Table 4.5 presents three subpopulation groups that were ranked first according to three different fairness definitions: *Equal Cost of Effectiveness (ϕ = 0.7)*, *Equal Choice for Recourse (ϕ = 0.7)* and *Equal Cost of Effectiveness (ϕ = 0.3)*. These groups are identified as having the highest levels of unfairness based on the respective definitions. For each group, the table provides its *rank*, the subgroup it is *biased against*, and its *unfairness score* across all the fairness definitions listed in the leftmost column. When the unfairness score is 0, we label the subgroup as "Fair" in the rank column. Note that groups with identical scores for a particular definition receive the same rank. The Comparative Subgroup Counterfactual (CSC) representations for the fairness definitions that ranked these three subgroups at the top are displayed in Figure 4.8.

Subgroup 1 ranks highest in unfairness based on the *Equal Cost of Effectiveness with* $\phi = 0.7$, yet is ranked much lower or even considered "fair" under most of the other definitions (as seen in the "rank" column for Subgroup 1). This pattern is similarly observed for Subgroup 2, which ranks first under *Equal Choice for Recourse with* $\phi = 0.7$, and Subgroup 3, which tops the ranking for *Equal Cost of Effectiveness with* $\phi = 0.3$, but both rank much lower under other definitions. These findings highlight the utility of applying different fairness definitions, as each captures distinct aspects of the difficulty in achieving recourse.[11].

The relevance of these different aspects is further motivated by their real-world applicability in auditing contexts. For instance, the *Equal Cost of Effectiveness* with $\phi = 0.7$ might be appropriate when a broad intervention to support a subpopulation is needed, but the number of actions that can be taken is limited. In this case, the macro-level perspective provided by Subgroup 1's CSC (the top result in Figure 4.8) is particularly useful, as it helps identify group-level actions that can achieve recourse for a significant portion of the disadvantaged subpopulation.

Conversely, *Equal Choice for Recourse* with $\phi = 0.3$, which ignores cost and ranks groups based solely on the difference in the number of effective actions between protected subgroups, may be more suitable in scenarios where assigning costs to actions is problematic or potentially unethical. This definition measures bias based on the availability of effective actions rather than their cost, as shown in Subgroup 2's CSC (middle result in Figure 4.8).

Another noteworthy observation comes from Subgroup 1, where bias is detected against the *Male* subgroup. This finding contrasts with prior work employing more statistical-oriented approaches (e.g. Tramer et al., 2017), which primarily reported bias against the *Female* subgroup. This finding is significant, as it highlights the phenomenon of gerrymandering (Kearns et al., 2018), where subgroup partitions are manipulated to mask bias. The FACTS framework demonstrates robustness to this issue by enabling the examination of sufficiently small subgroups through the minimum frequency threshold, as discussed in Section 4.2. This allows the detection of biases that might otherwise remain hidden.

[11] Further examples and statistics supporting this observation, based on a larger sample, are provided in Appendix A

Subgroup 1

```
If age=(41.0, 50.0], marital-status=Never-married, race=White, relationship=Not-in-family:
    Protected Subgroup = 'Male', 1.34% covered
        No recourses for this subgroup.
    Protected Subgroup = 'Female', 1.47% covered
        Make marital-status=Married-civ-spouse, relationship=Married with effectiveness 70.49%
    Bias against 'Male' due to Equal Cost of Effectiveness (threshold = 0.7). Unfairness score = inf.
```

Subgroup 2

```
If worklass=Private, hours-per-week=FullTime, marital-status=Married-civ-spouse, occupation=Adm-clerical, race=White:
  Protected Subgroup = 'Male', 1.04% covered
    Make hours-per-week=OverTime, occupation=Exec-managerial with effectiveness 70.00%
    Make hours-per-week=OverTime, occupation=Prof-specialty with effectiveness 70.00%
    Make hours-per-week=BrainDrain, occupation=Exec-managerial with effectiveness 70.00%
    Make hours-per-week=BrainDrain, occupation=Prof-specialty with effectiveness 70.00%
    Make Workclass=Self-emp-in, occupation=Exec-managerial with effectiveness 70.00%
    Make Workclass=Self-emp-in, hours-per-week=OverTime, occupation=Exec-managerial with effectiveness 80.00%
    Make Workclass=Self-emp-in, hours-per-week=OverTime, occupation=Sales with effectiveness 70.00%
    Make Workclass=Self-emp-in, hours-per-week=BrainDrain, occupation=Exec-managerial with effectiveness 70.00%
  Protected Subgroup = 'Female', 3.51% covered
    Make Workclass=Self-emp-in, hours-per-week=OverTime, occupation=Exec-managerial with effectiveness 74.51%
    Make Workclass=Self-emp-in, hours-per-week=BrainDrain, occupation=Exec-managerial with effectiveness 74.51%
  Bias against 'Female' due to Equal Choice for Recourse (threshold = 0.7). Unfairness score = 6
```

Subgroup 3

```
If age=(41.0, 50.0], occupation=Sales:
    Protected Subgroup 'Male', 1.18% covered
        Make occupation Craft-repair with effectiveness 0.00%
        Make occupation=Adm-clerical with effectiveness 0.00%
        Make occupation=Tech-support with effectiveness 19.23%
        Make occupation=Prof-specialty with effectiveness 28.21%
        Make occupation=Exec-managerial with effectiveness 39.74%
    Protected Subgroup 'Female', 1.56% covered
        Make occupation=Craft-repair with effectiveness 0.00%
        Make occupation=Adm-clerical with effectiveness 0.00%
        Make occupation=Tech-support with effectiveness 0.00%
        Make occupation=Exec-managerial with effectiveness 6.94%
        Make occupation=Prof-specialty with effectiveness 6.94%
        Make age=(50.0,90.0] with effectiveness 6.94%
        Make age=(50.0,90.0], occupation=Prof-specialty with effectiveness 6.94%
        Make age=(50.0,90.0], occupation=Craft-repair with effectiveness 6.94%
        Make age=(50.0,90.0], occupation=Adm-clerical with effectiveness 6.94%
        Make age=(50.0,90.0], occupation=Exec-managerial with effectiveness 6.94%
    Bias against 'Female' due to Equal Cost of Effectiveness (threshold = 0.3). Unfairness score = inf.
```

Figure 4.8: Comparative Subgroup Counterfactuals for the subgroups of Table 4.5.

Chapter 5

Conclusions

In this thesis we have studied both the challenge of efficiently calculating effective and interpretable global counterfactual explanations (GCE) and the nuances of using global counterfactual explanations to assess the recourse fairness of black-box systems.

On one hand, we presented the GLANCE framework, which offers an innovative approach to calculating GCEs, balancing effectiveness, interpretability, and practicality. Through careful clustering and strategic action generation, GLANCE enables users to interpret a model's decision boundaries and understand potential pathways to favorable outcomes. Evaluations against benchmark methods demonstrated GLANCE's capacity to generate high-quality, reliable explanations across different models and datasets, with competitive Pareto dominance rates that underscore its effectiveness and efficiency.

On the other hand, we used the FACTS framework to further investigate recourse fairness in GCEs, addressing how accessible and equitable recourse options are for individuals in distinct demographic groups. By employing FACTS, we achieved a deeper understanding of recourse fairness within black-box systems, examining how well different counterfactual approaches uphold fairness metrics. Through these efforts, we highlighted limitations in current fairness metrics and underscored the importance of evaluating counterfactual methods not only by their cost and effectiveness but also by their practical fairness impact across protected groups.

Future Directions

Several directions remain open for advancing both frontiers of explainability and fairness based on global counterfactual explanations:

- **Enhanced Action Selection at AS's Third Stage**: The third stage of the Augmented Space algorithm, which finalizes the set of actions, currently follows a straightforward selection process. Future work could refine this selection through more sophisticated optimization techniques, potentially improving both the practicality and the cost-efficiency of actions suggested by AS.

- **Interpretable Clustering Techniques**: The clustering stage in AS is critical for identifying groups with similar recourse needs. Incorporating interpretable clustering methods could enhance the transparency of AS, allowing end-users and auditors to better understand the basis for clustering, thus supporting more trust in the model's decision-making process.

- **Extending Fairness of Recourse Definitions in FACTS**: The FACTS framework laid out foundational definitions for fairness in recourse. These definitions definitely warrant further development. A deeper focus on ECD differences across demographic

subgroups could ensure that fairness metrics more accurately reflect the fairness impacts on those requiring recourse.

- **Exploring Alternatives to Frequent Itemset Mining**: While frequent itemset mining has been effective for generating actions for FACTS, alternative methods may offer more tailored or context-sensitive actions. Investigating these alternatives could lead to more granular, contextually relevant actions that further support user recourse.
- **Incorporating Systematic Feasibility Constraints**: Ensuring that counterfactual actions are feasible is essential for real-world applicability. Future versions of both GLANCE and FACTS could integrate systematic feasibility criteria, informed by causal constraints or domain-specific rules, to ensure that the recommended actions are practical for users to implement.

In summary, this thesis provides a comprehensive study of GLANCE as a scalable and interpretable GCE generation method and an analysis of FACTS to evaluate recourse fairness in complex, black-box models. Together, these frameworks offer valuable insights into the practical challenges of providing fair, actionable, and interpretable recourse in automated decision-making. By further refining these frameworks, future research can enhance the fairness, reliability, and transparency of AI-driven systems in high-stakes applications.

Bibliography

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *Ethics of Data and Analytics*, pages 254–264, 5 2016. doi: 10.1201/9781003278290-37.

K. Bache and M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

E. Black, S. Yeom, and M. Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

N. Boer, D. Deutch, N. Frost, and T. Milo. Just in time: Personal temporal insights for altering model decisions. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1988–1991. IEEE, 2019.

M. Bogen and A. Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. *Analysis & Policy Observatory*, 2018.

J. Branke. *Multiobjective optimization: Interactive and evolutionary approaches*, volume 5252. Springer Science & Business Media, 2008.

K. Brown, D. Doran, R. Kramer, and B. Reynolds. HELOC applicant risk performance evaluation by topological hierarchical decomposition. *CoRR*, abs/1811.10658, 2018. URL `http://arxiv.org/abs/1811.10658`.

N. Burkart and M. F. Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

L. Cohen, Z. C. Lipton, and Y. Mansour. Efficient candidate screening under multiple tests and implications for fairness. *arXiv preprint arXiv:1905.11361*, 2019.

S. Dandl, C. Molnar, M. Binder, and B. Bischl. *Multi-Objective Counterfactual Explanations*, page 448–469. Springer International Publishing, 2020. ISBN 9783030581121. doi: 10.1007/978-3-030-58112-1_31. URL `http://dx.doi.org/10.1007/978-3-030-58112-1_31`.

D. Dua and C. Graff. Uci machine learning repository. university of california, school of information and computer science, irvine, ca (2019), 2019.

I. Emiris, D. Fotakis, G. Giannopoulos, D. Gunopulos, L. Kavouras, K. Markou, E. Psaroudaki, D. Rontogiannis, D. Sacharidis, N. Theologitis, D. Tomaras, and K. Tsopelas. Glance: Global actions in a nutshell for counterfactual explainability, 2024. URL `https://arxiv.org/abs/2405.18921v1`.

M. T. M. Emmerich and A. H. Deutz. A tutorial on multiobjective optimization: fundamentals and evolutionary methods. *Natural Computing*, 17(3):585–609, 2018. ISSN 1572-9796. doi: 10.1007/s11047-018-9685-y. URL https://doi.org/10.1007/s11047-018-9685-y.

V. Gupta, P. Nokhiz, C. D. Roy, and S. Venkatasubramanian. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.

J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In W. Chen, J. F. Naughton, and P. A. Bernstein, editors, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, pages 1–12. ACM, 2000. doi: 10.1145/342009.335372. URL https://doi.org/10.1145/342009.335372.

K. Kanamori, T. Takagi, K. Kobayashi, and Y. Ike. Counterfactual explanation trees: Transparent and consistent actionable recourse with decision trees. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 1846–1870. PMLR, 2022. URL https://proceedings.mlr.press/v151/kanamori22a.html. ISSN: 2640-3498.

A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020.

A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

L. Kavouras, K. Tsopelas, G. Giannopoulos, D. Sacharidis, E. Psaroudaki, N. Theologitis, D. Rontogiannis, D. Fotakis, and I. Emiris. Fairness aware counterfactuals for subgroups. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=38dQv3OwN3.

L. Kavouras, E. Psaroudaki, K. Tsopelas, D. Rontogiannis, N. Theologitis, D. Sacharidis, G. Giannopoulos, D. Tomaras, K. Markou, D. Gunopulos, D. Fotakis, and I. Emiris. Glance: Global actions in a nutshell for counterfactual explainability, 2024. URL https://arxiv.org/abs/2405.18921.

M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.

D. Knuth. *The Art of Computer Programming: Seminumerical Algorithms, Volume 2*. Pearson Education, 2014. ISBN 9780321635761. URL https://books.google.gr/books?id=Zu-HAwAAQBAJ.

A. Kuratomi, E. Pitoura, P. Papapetrou, T. Lindgren, and P. Tsaparas. Measuring the burden of (un) fairness using counterfactuals. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part I*, pages 402–417. Springer, 2023.

E. Kyrimi, M. R. Neves, S. McLachlan, M. Neil, W. Marsh, and N. Fenton. Medical idioms for clinical bayesian network development. *Journal of Biomedical Informatics*, 108:103495, 2020.

T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki. Inverse classification for comparison-based interpretability in machine learning, 2017.

D. Ley, S. Mishra, and D. Magazzeni. Global counterfactual explanations: Investigations, implementations and improvements. In *ICLR Workshop on Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022.

D. Ley, S. Mishra, and D. Magazzeni. GLOBE-CE: A translation based approach for global counterfactual explanations. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19315–19342. PMLR, 2023. URL `https://proceedings.mlr.press/v202/ley23a.html`. ISSN: 2640-3498.

J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

A. Lucic, H. Oosterhuis, H. Haned, and M. de Rijke. Focus: Flexible optimizable counterfactual explanations for tree ensembles, 2021.

J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning, 2022.

T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

C. Molnar. *Interpretable Machine Learning*. Molnar, Christoph, 2023. URL `https://christophm.github.io/interpretable-ml-book/`.

R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2019. doi: 10.1145/3351095.3372850. URL `http://arxiv.org/abs/1905.07697`.

O. A. Osoba and W. Welser IV. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation, 2017.

R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20. ACM, Feb. 2020. doi: 10.1145/3375627.3375850. URL `http://dx.doi.org/10.1145/3375627.3375850`.

I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

K. Rawal and H. Lakkaraju. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems*, 33:12187–12198, 2020.

S. Sharma, J. Henderson, and J. Ghosh. CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *AIES*, pages 166–172. ACM, 2020.

J. Tashea. Courts are using ai to sentence criminals. that must stop now, Apr 2017. URL `https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/`.

F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert, A. Juels, and H. Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE, 2017.

B. Ustun, A. Spangher, and Y. Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

S. Venkatasubramanian and M. Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 284–293, 2020.

J. von Kügelgen, A. Karimi, U. Bhatt, I. Valera, A. Weller, and B. Schölkopf. On the fairness of causal algorithmic recourse. In *AAAI*, pages 9584–9594. AAAI Press, 2022.

S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

G. Warren, M. T. Keane, C. Gueret, and E. Delaney. Explaining groups of instances counterfactually for xai: a use case, algorithm and user study for group-counterfactuals. *arXiv preprint arXiv:2303.09297*, 2023.

I. Yeh and C.-H. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36:2473–2480, 03 2009. doi: 10.1016/j.eswa.2007.12.020.

Appendix A

FACTS: Additional Experiments

In this appendix, following the methodology of Kavouras et al. (2023), we repeat the experiment described in Section 4.4, which involved the Adult dataset with 'gender' as the protected attribute, and apply it for three additional cases.

Specifically, we present three subgroups that were ranked first in terms of unfairness according to one of the fairness metrics, explain why these subgroups were identified as unfair by the FACTS framework, and summarize their unfairness scores across the remaining metrics.

A.1 Results for the Adult Dataset with Race as the Protected Attribute

We showcase three prevalent subgroups for which the rankings assigned by different fairness definitions truly yield different kinds of information. This is showcased in Table A.1. We once again note that the results presented here are for 'race' as a protected attribute, while the corresponding results for 'gender' are presented in Section 4 of the main paper.

Here we showcase three prominent groups from the Adult dataset, using 'race' as the protected attribute. The rankings provided by different fairness definitions reveal different types of information about subgroup unfairness, as illustrated in Table A.1. For comparison, we remind that the corresponding results for 'gender' as the protected attribute are discussed in Section 4.4.

Table A.1: Example of three unfair groups in Adult (protected attribute race)

	Group 1			Group 2			Group 3		
	rank	bias against	unfairness score	rank	bias against	unfairness score	rank	bias against	unfairness score
Equal Effectiveness	Fair	Fair	0.0	3047.0	Non-White	0.115	1682.0	Non-White	0.162
Equal Choice for Recourse (ϕ = 0.3)	1	Non-White	10.0	10.0	Non-White	1.0	Fair	Fair	0.0
Equal Choice for Recourse (ϕ = 0.7)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 1.15)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 10.0)	303.0	Non-White	0.242	2201.0	Non-White	0.115	4035.0	Non-White	0.071
Equal Effectiveness within Budget (c = 21.0)	Fair	Fair	0.0	2978.0	Non-White	0.115	1663.0	Non-White	0.162
Equal Cost of Effectiveness (ϕ = 0.3)	18.0	Non-White	0.15	1	Non-White	inf	Fair	Fair	0.0
Equal Cost of Effectiveness (ϕ = 0.7)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Fair Effectiveness-Cost Trade-Off	909.0	Non-White	0.242	4597.0	Non-White	0.115	2644.0	Non-White	0.162
Equal (Conditional) Mean Recourse	5897.0	White	0.021	5309.0	White	0.047	1	Non-White	inf

In Figure A.1 we present the Comparative Subgroup Counterfactual representation for the subgroups of Table A.1 that corresponds to the fairness metric for which each subgroup presents the minimum rank.

In Figure A.1, we present the Comparative Subgroup Counterfactuals (CSC) representation for the groups in Table A.1. Each group is represented according to the fairness metric for which it was ranked highest (i.e., most unfair).

These results align with the findings from Section 4.4 on the same dataset (Adult) but with a different protected attribute (race instead of gender). Groups ranked first (most unfair) for one fairness definition are often ranked lower or considered fair by other definitions. This underscores the utility of using different fairness definitions, which is further supported by the diversity of the CSC representations corresponding to Table A.1. For instance, Subgroup 1, ranked first by *Equal Choice for Recourse* ($\phi = 0.3$), demonstrates unfairness by showing a significantly larger number of actions available for the "White" subgroup compared to the "Non-White" subgroup. In contrast, Subgroup 2, ranked first by *Equal Cost of Effectiveness* ($\phi = 0.3$), illustrates a situation where no recourses are found for the required percentage of the "Non-White" protected subgroup, while sufficient recourses exist for the "White" subgroup.

```
Subgroup 1
If workclass=Private, age=(34.0, 41.0], capital-gain=0, capital-loss=0, marital-status=Never-married, native-country=
    United-States, relationship=Not-in-family:
    Protected Subgroup = 'Non-White', 1.09% covered
        Make Workclas=Federal-gov, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married with
            effectiveness 36.84%.
        Make capital-gain=15024, marital-status=Married-civ-spouse, relationship=Married with effectiveness 100.00%.
        Make age=(41.0, 50.0], capital-gain=15024, marital-status=Married-civ-spouse, relationship=Married with
            effectiveness 100.00%.
    Protected Subgroup = 'White', 1.94% covered
        Make age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married with effectiveness 45.14%.
        Make marital-status=Married-civ-spouse, relationship=Married with effectiveness 40.00%.
        Make age=(50.0, 90.0], marital-status= Married-civ-spouse, relationship=Married with effectiveness 42.86%.
        Make Workclass=Local-gov, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married with
            effectiveness 44.00%.
        Make Workclass=Local-gov, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married with
            effectiveness 44.00%.
        Make Workclas=Self-emp-inc, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married with
            effectiveness 52.57%.
        Make Workclass=Self-emp-inc, age=(50.0, 90.0], marital-status=Married-civ-spouse, relationship=Married with
            effectiveness 49.71%.
        Make Workclass=Local-gov, marital-status=Married-civ-spouse, relationship=Married with effectiveness 36.00%.
        Make Workclass=Self-emp-inc, marital-status=Married-civ-spouse, relationship=Married with effectiveness 45.14%.
        Make Workclass= Federal-gov, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married with
            effectiveness 62.29%.
        Make Workclass=State-gov, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married with
            effectiveness 40.57%.
        Make capital-gain=15024, marital-status=Married-civ-spouse, relationship=Married with effectiveness 99.43%.
        Make Workclass=Local-gov, age=(50.0, 90.0], marital-status=Married-civ-spouse, relationship=Married with
            effectiveness 40.00%.
        Make age=(41.0, 50.0], capital-gain=15024, marital-status=Married-civ-spouse, relationship=Married with
            effectiveness 100.00%.
    Bias against 'Non-White' due to Equal Choice for Recourse (threshold = 0.3). Unfairness score = 10.
```

```
Subgroup 2
If hours-per-week = FullTime, native-country = United-States, occupation = Adm-clerical, relationship = Married:
    Protected Subgroup = 'Non-White', 1.66% covered
        No recourses for this subgroup.
    Protected Subgroup = 'White', 1.66% covered
        Make hours-per-week = BrainDrain, occupation = Exec-managerial with effectiveness 70.00%.
    Bias against 'Non-White' due to Equal Cost of Effectiveness (threshold = 0.3). Unfairness score = inf.
```

```
Subgroup 3
If hours-per-week = PartTime, marital-status = Divorced, native-country = United-States:
    Protected Subgroup = 'Non-White', 1.15% covered
        Make marital-status=Married-civ-spouse with effectiveness 0.00%.
        Make hours-per-week=MidTime, marital-statu=Married-civ-spouse with effectiveness 0.00%.
        Make hours-per-week=FullTime, marital-status=Married-civ-spouse with effectiveness 0.00%.
        Make hours-per-week=OverTime, marital-status=Married-civ-spouse with effectiveness 0.00%.
        Make hours-per-week=OverTime, marital-status=Never-married with effectiveness 0.00%.
        Make hours-per-week=BrainDrain, marital-status=Married-civ-spouse with effectiveness 0.00%.
    Protected Subgroup = 'White', 1.66% covered
        Make marital-status=Married-civ-spouse with effectiveness 1.01%.
        Make hours-per-week=MidTime, marital-statu=Married-civ-spouse with effectiveness 1.01%.
        Make hours-per-week=FullTime, marital-status=Married-civ-spouse with effectiveness 7.07%.
        Make hours-per-week=OverTime, marital-status=Married-civ-spouse with effectiveness 7.07%.
        Make hours-per-week=OverTime, marital-status=Never-married with effectiveness 15.15%.
        Make hours-per-week=BrainDrain, marital-status=Married-civ-spouse with effectiveness 16.16%.
    Bias against 'Non-White' due to Equal Conditional Mean Recourse. Unfairness score = inf.
```

Figure A.1: Example of three Comparative Subgroup Counterfactuals in Adult (protected attribute race); ref. Table A.1

A.2 Results for the COMPAS Dataset

In Table A.2, we present ranking statistics for three notable subgroups identified using the COMPAS dataset with 'race' as the protected attribute. As with the Adult dataset, we observe distinct patterns depending on the fairness metric used, highlighting how each definition captures different aspects of subgroup unfairness. The Comparative Subgroup Counterfactuals for these subgroups are displayed in Figure A.2.

Table A.2: Example of three unfair subgroups in COMPAS

	Subgroup 1			Subgroup 2			Subgroup 3		
	rank	bias against	unfairness score	rank	bias against	unfairness score	rank	bias against	unfairness score
Equal Effectiveness	Fair	Fair	0.0	116.0	African-American	0.151	209.0	African-American	0.071
Equal Choice for Recourse (ϕ = 0.3)	Fair	Fair	0.0	3.0	African-American	1.0	Fair	Fair	0.0
Equal Choice for Recourse (ϕ = 0.7)	1	African-American	3.0	Fair	Fair	0.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 1)	66.0	African-American	0.167	79.0	African-American	0.151	185.0	African-American	0.071
Equal Effectiveness within Budget (c = 10)	84.0	African-American	0.167	108.0	African-American	0.151	220.0	African-American	0.071
Equal Cost of Effectiveness (ϕ = 0.3)	Fair	Fair	0.0	1	African-American	inf	Fair	Fair	0.0
Equal Cost of Effectiveness (ϕ = 0.7)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Fair Effectiveness-Cost Trade-Off	3.0	African-American	0.5	214.0	African-American	0.151	376.0	African-American	0.071
Equal (Conditional) Mean Recourse	59.0	African-American	1.667	Fair	Fair	0.0	1	African-American	inf

Subgroup 1

```
If age_cat = 25 - 45, c_charge_degree = M, juv_misd_count = 0, priors_count = (10.0, 15.0]:
    Protected Subgroup = 'Caucasian', 1.03% covered
        Make c_charge_degree=F, priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make priors_count=(1.0, 5.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, c_charge_degree=F, priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, c_charge_degree=F, priors_count=(1.0, 5.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, priors_count=(1.0, 5.0] with effectiveness 100.00%.
    Protected Subgroup = 'African-American', 1.16% covered
        Make priors_count=(-0.1, 1.0] with effectiveness 83.33%.
        Make age_cat=Greater than 45, priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, c_charge_degree=F, priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, priors_count=(1.0, 5.0] with effectiveness 83.33%.
    Bias against African-American due to Equal Choice for Recourse (threshold = 0.7). Unfairness score = 3.
```

Subgroup 2

```
If c_charge_degree = M, juv_other_count = 1:
    Protected Subgroup = 'Caucasian', 3.59% covered
        Make juv_other_count = 0 with effectiveness 42.86%.
    Protected Subgroup = 'African-American', 3.48% covered
        No recourses for this subgroup.
    Bias against African-American due to Equal Cost of Effectiveness (threshold = 0.3). Unfairness score = inf.
```

Subgroup 3

```
If age_cat = Greater than 45, c_charge_degree = F, juv_fel_count = 0, juv_misd_count = 0, juv_other_count = 0, sex = Male:
    Protected Subgroup = 'Caucasian', 7.18% covered
        Make c_charge_degree=M with effectiveness 7.14%.
    Protected Subgroup = 'African-American', 6.00% covered
        Make c_charge_degree=M with effectiveness 0.00%.
    Bias against African-American due to Equal Conditional Mean Recourse. Unfairness score = inf.
```

Figure A.2: Example of three Comparative Subgroup Counterfactuals in COMPAS; ref. Table A.2

A.3 Results for the SSL Dataset

The results for the SSL dataset, using 'race' as the protected attribute, are summarized in Table A.3. Ranking statistics for three key subgroups are provided, and their corresponding Comparative Subgroup Counterfactuals are shown in Figure A.3. As with previous datasets, the results demonstrate how different fairness definitions produce varying rankings, capturing distinct biases in the recourse provided to protected subgroups.

Table A.3: Example of three unfair subgroups in SSL

	Subgroup 1			Subgroup 2			Subgroup 3		
	rank	bias against	unfairness score	rank	bias against	unfairness score	rank	bias against	unfairness score
Equal Effectiveness	1630.0	Black	0.076	70.0	Black	0.663	979.0	Black	0.151
Equal Choice for Recourse (ϕ = 0.3)	Fair	Fair	0.0	12.0	Black	1.0	12.0	Black	1.0
Equal Choice for Recourse (ϕ = 0.7)	13.0	Black	3.0	Fair	Fair	0.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 1)	Fair	Fair	0.0	195.0	Black	0.663	1692.0	White	0.138
Equal Effectiveness within Budget (c = 2)	2427.0	Black	0.111	126.0	Black	0.663	3686.0	White	0.043
Equal Effectiveness within Budget (c = 10)	2557.0	Black	0.076	73.0	Black	0.663	1496.0	Black	0.151
Equal Cost of Effectiveness (ϕ = 0.3)	Fair	Fair	0.0	1	Black	inf	1	Black	inf
Equal Cost of Effectiveness (ϕ = 0.7)	1	Black	inf	Fair	Fair	0.0	Fair	Fair	0.0
Fair Effectiveness-Cost Trade-Off	3393.0	Black	0.111	443.0	Black	0.663	2685.0	Black	0.151
Equal (Conditional) Mean Recourse	3486.0	Black	0.053	1	Black	inf	1374.0	White	0.95

A.4 Results for the Ad Campaign Dataset

For the Ad Campaign dataset, using 'gender' as the protected attribute, we present ranking results for three subgroups in Table A.4. The corresponding Comparative Subgroup Counterfactuals are illustrated in Figure A.4, further supporting the observation that different fairness metrics highlight different types of unfairness.

Table A.4: Example of three unfair subgroups in Ad Campaign

	Subgroup 1			Subgroup 2			Subgroup 3		
	rank	bias against	unfairness score	rank	bias against	unfairness score	rank	bias against	unfairness score
Equal Effectiveness	319.0	Female	0.286	Fair	Fair	0.0	467.0	Male	0.099
Equal Choice for Recourse (ϕ = 0.3)	5.0	Female	1.0	2.0	Female	4.0	Fair	Fair	0.0
Equal Choice for Recourse (ϕ = 0.7)	Fair	Fair	0.0	1	Female	4.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 1)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 5)	Fair	Fair	0.0	Fair	Fair	0.0	345.0	Male	0.099
Equal Cost of Effectiveness (ϕ = 0.3)	1	Female	inf	Fair	Fair	0.0	Fair	Fair	0.0
Equal Cost of Effectiveness (ϕ = 0.7)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Fair Effectiveness-Cost Trade-Off	331.0	Female	0.286	Fair	Male	0.0	547.0	Male	0.099
Equal (Conditional) Mean Recourse	Fair	Fair	0.0	Fair	Fair	0.0	1	Male	inf

```
Subgroup 1
If PREDICTOR RAT ARRESTS VIOLENT OFFENSES = 1, PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT VICTIM BATTERY OR ASSAULT
      = 1:
  Protected Subgroup = 'Black', 1.04% covered
              No recourses for this subgroup.
  Protected Subgroup = 'White', 1.00% covered
    Make PREDICTOR RAT ARRESTS VIOLENT OFFENSES = 0, PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT VICTIM BATTERY OR
         ASSAULT = 0 with effectiveness 72.73%
    Make PREDICTOR RAT ARRESTS VIOLENT OFFENSES = 0, PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT VICTIM BATTERY OR
         ASSAULT = 0 with effectiveness 72.73%
    Make PREDICTOR RAT ARRESTS VIOLENT OFFENSES = 0, PREDICTOR RAT NARCOTIC ARRESTS = 2, PREDICTOR RAT VICTIM BATTERY OR
         ASSAULT = 0 with effectiveness 72.73%
  Bias against 'Black' due to Equal Cost of Effectiveness (threshold = 0.7). Unfairness score = inf.
```

```
Subgroup 2
If PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.2, -0.1], PREDICTOR RAT UUW ARRESTS
     = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0:
  Protected Subgroup = 'Black', 2.51% covered
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999, -0.3],
         PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.1, 0.1], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.1, 0.3], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999, -0.3],
         PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.3, 7.3], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.3, -0.2], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.1, 0.1], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.2, -0.1], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.1, 0.3], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 2, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999, -0.3],
         PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.3, 7.3], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
  Protected Subgroup = 'White', 2.87% covered
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999, -0.3],
         PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 57.14%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.1, 0.1], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.1, 0.3], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999, -0.3],
         PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.3, 7.3], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.3, -0.2], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.1, 0.1], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.2, -0.1], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.1, 0.3], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 2, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999, -0.3],
         PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.3, 7.3], PREDICTOR RAT UUW
         ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
  Bias against 'Black' due to Equal(Conditional) Mean Recourse. Unfairness score = inf.
```

```
Subgroup 3
If PREDICTOR RAT GANG AFFILIATION = 1, PREDICTOR RAT NARCOTIC ARRESTS = 2, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY =
     (-8.200999999999999, -0.3], PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0:
  Protected Subgroup = 'Black', 1.18% covered
              No recourses for this subgroup.
  Protected Subgroup = 'White', 1.00% covered
    Make PREDICTOR RAT GANG AFFILIATION = 0, PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY
         = (-8.200999999999999, -0.3], PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 31.82%
  Bias against 'Black' due to Equal Cost of Effectiveness (threshold = 0.3). Unfairness score = inf.
```

Figure A.3: Example of three Comparative Subgroup Counterfactuals in SSL; ref. Table A.3

Subgroup 1

```
If age = 45-54, area = Unknown, parents = 1:
    Protected Subgroup = 'Male', 1.22% covered
        Make age=55-64, area=Rural with effectiveness 30.77%.
    Protected Subgroup = 'Female', 1.13% covered
        No recourses for this subgroup.
    Bias against Female due to Equal Cost of Effectiveness (threshold=0.3). Unfairness score = inf.
```

Subgroup 2

```
If age = 55-64, area = Unknown, homeowner = 1, income = Unknown, parents = 0, politics = Unknown, religion = Unknown:
    Protected Subgroup = 'Male', 2.53% covered
        Make homeowner=0, parents=1 with effectiveness 100.00%.
        Make homeowner=0, parents=1, religion=Christianity with effectiveness 100.00%.
        Make homeowner=0, parents=1, religion=Other with effectiveness 100.00%.
        Make area=Urban, parents=1, religion=Christianity with effectiveness 93.91%.
        Make area=Urban, parents=1, religion=Other with effectiveness 93.91%.
        Make homeowner=0, income=<100K, parents=1 with effectiveness 100.00%.
        Make area=Rural, parents=1, religion=Other with effectiveness 100.00%.
        Make area=Rural, parents=1, religion=Christianity with effectiveness 100.00%.
        Make homeowner=0, income=<100K, parents=1, religion=Christianity with effectiveness 100.00%.
        Make homeowner=0, income=<100K, parents=1, religion=Other with effectiveness 100.00%.
    Protected Subgroup = 'Female', 2.33% covered
        Make homeowner=0, parents=1 with effectiveness 100.00%.
        Make homeowner=0, parents=1, religion=Christianity with effectiveness 100.00%.
        Make homeowner=0, parents=1, religion=Other with effectiveness 100.00%.
        Make homeowner=0, income=<100K, parents=1 with effectiveness 100.00%.
        Make homeowner=0, income=<100K, parents=1, religion=Christianity with effectiveness 100.00%.
        Make homeowner=0, income=<100K, parents=1, religion=Other with effectiveness 100.00%.
    Bias against Female due to Equal Choice for Recourse (threshold=0.7). Unfairness score = 4.
```

Subgroup 3

```
If ages = 55-64, income = <100K, religion = Unknown:
    Protected Subgroup = 'Male', 1.02% covered
        Make religion=Christianity with effectiveness 0.00%.
        Make religion=Other with effectiveness 0.00%.
    Protected Subgroup = 'Female', 1.08% covered
        Make religion=Christianity with effectiveness 9.86%.
        Make religion=Other with effectiveness 9.86%.
    Bias against Male due to Equal Conditional Mean Recourse. Unfairness score = inf.
```

Figure A.4: Example of three Comparative Subgroup Counterfactuals in Ad Campaign; ref. Table A.4

A.5 Comparison of Fairness Metrics

This section seeks to answer the question: "How do the different fairness of recourse metrics compare?" To address this, we analyze all subpopulation groups and compare their unfairness rankings across 12 distinct fairness metrics. The results provide clear evidence that these metrics capture different aspects of recourse unfairness, as discussed previously and in Section 4.4.

For each dataset and protected attribute, we provide two tables: (a) the ranking analysis table, and (b) the aggregated rankings table.

A.5.1 Ranking Analysis Table

The first column of the *ranking analysis* table shows the number of most unfair subgroups per metric (i.e., those tied for rank 1). The number of ties varies greatly depending on the metric used, as different metrics compare protected subgroups using different units (e.g., cost, effectiveness, or number of actions). For example, metrics based on effectiveness percentages tend to have fewer ties, while cost-based metrics often result in more ties. The second and third columns provide statistics on the direction of bias (e.g., bias against males vs. bias against females) for the top 10% most unfair subgroups.

A.5.2 Aggregated Rankings Table

The *aggregated rankings* table allows for a comparative analysis of how subpopulation groups ranked as most unfair by one metric are evaluated by other metrics. Specifically, it shows the relative ranking of groups that are ranked first in one metric when evaluated according to the remaining metrics. The values v in each cell i, j are computed as follows:

1. We collect all groups ranked first (most biased) according to the fairness metric in row i.

2. We compute the average rank a of these groups in the fairness metric of column j.

3. We normalize a by dividing it by the total number of ranking tiers for the fairness metric in column j, resulting in the value v.

Each non-diagonal value in the table represents the average ranking of the groups, which are ranked first by the fairness metric in the corresponding row, when evaluated using the fairness metric in the corresponding column. The closer the value is to 1, it suggests that groups ranked highest by the row metric are ranked significantly lower (i.e., less unfair) by the column metric. Conversely, a value close to 0 indicates that groups ranked highest by the row metric remain highly ranked (i.e., still considered unfair) according to the column metric.

Each diagonal value (i.e., where the metric in the row and the column is the same) is left empty, as subgroups ranked first by a given metric are, by definition, ranked as the most unfair by that same metric.

Thus, the non-diagonal values give us insights into how the most unfair subgroups, according to one metric, fare when evaluated against other fairness metrics. This cross-comparison helps us understand whether different metrics capture similar types of unfairness or focus on distinct aspects of the problem.

A.5.3 Comparison on Adult Dataset (Gender as Protected Attribute)

In the test set for the Adult dataset, 10,205 individuals were affected. We split the affected individuals into two subgroups: affected males (D_1) and affected females (D_0). Running fp-growth with a support threshold of 1% on D_1 and D_0, and computing the intersection of the results, we generated 12,880 subpopulation groups. The fairness metrics then evaluated and ranked these groups based on the actions applied.

Tables A.5 and A.6 present the ranking analysis and the aggregated rankings, respectively, for the Adult dataset with 'gender' as the protected attribute. The following discussion focuses on the key findings from these two tables. Similar patterns are observed in the tables for the other datasets, so we omit their detailed discussion here.

From Table A.5, it becomes clear that different fairness definitions can lead to significant variations in the number of groups that tie for the same rank. In this case, only rank 1 is shown. The "Top 10%" columns provide particularly interesting insights into which protected subgroups experience bias. While it might be expected that most bias would be detected against the "Female" subgroup, we also find instances where bias is identified against "Male" subgroups. This demonstrates a robustness to gerrymandering, which refers to the manipulation of subgroup divisions to conceal biases (as previously hinted in Section 4.4).

Table A.6 offers stronger evidence for the distinct usefulness of each fairness definition (as discussed in the commentary of Section 4.4). This table reports the average relative (normalized) rankings of the subgroups that are ranked first by one fairness definition when evaluated by the other definitions. The normalization maps rankings to a scale between 0 and 1, where a value close to 1 indicates a relatively low average rank (i.e., subgroups ranked first by one definition are ranked much lower by another), while a value close to 0 indicates a high average rank (i.e., subgroups ranked first by one definition are still highly ranked by another).

Values that deviate significantly from 0 show that different fairness definitions capture distinct aspects of unfairness, revealing the non-trivial and unique characteristics of each definition. The table's values support this, as many groups that are ranked first according to one definition are ranked much lower when evaluated with a different fairness definition. This pattern underscores the importance of using multiple fairness definitions to obtain a comprehensive view of recourse unfairness.

Table A.5: Ranking Analysis in Adult (protected attribute gender)

	# Most Unfair Subgroups	# Subgroups w. Bias against Males (in Top 10% Unfair Subgroups)	# Subgroups w. Bias against Females (in Top 10% Unfair Subgroups)
(Equal Cost of Effectiveness(Macro), 0.3)	1673	56	206
(Equal Cost of Effectiveness(Macro), 0.7)	301	26	37
(Equal Choice for Recourse, 0.3)	2	54	286
(Equal Choice for Recourse, 0.7)	6	31	50
Equal Effectiveness	1	39	1040
(Equal Effectiveness within Budget, 5.0	1	41	616
(Equal Effectiveness within Budget, 10.0)	1	6	904
(Equal Effectiveness within Budget, 18.0)	1	22	964
(Equal Cost of Effectiveness(Micro), 0.3)	1523	10	226
(Equal Cost of Effectiveness(Micro), 0.7)	290	38	27
Equal(Conditional Mean Recourse)	764	540	565
(Fair Effectiveness-Cost Trade-Off, value)	1	61	1156

Table A.6: Aggregated Rankings in Adult (protected attribute gender)

	(Equal Cost of Effectiveness (Macro), 0.3)	(Equal Cost of Effectiveness (Macro), 0.7)	(Equal Choice for Recourse, 0.3)	(Equal Choice for Recourse, 0.7)	Equal Effectiveness	(Equal Effectiveness within Budget, 5.0)	(Equal Effectiveness within Budget, 10.0)	(Equal Effectiveness within Budget, 18.0)	(Equal Cost of Effectiveness (Micro), 0.3)	(Equal Cost of Effectiveness (Micro), 0.7)	Equal(Conditional Mean Recourse)	(Fair Effectiveness-Cost Trade-Off, value)
(Equal Cost of Effectiveness(Macro), 0.3)	-	1.0	0.836	1.0	0.214	0.509	0.342	0.285	0.3	1.0	0.441	0.237
(Equal Cost of Effectiveness(Macro), 0.7)	0.634	-	0.864	0.686	0.358	0.602	0.464	0.407	0.738	0.293	0.481	0.307
(Equal Choice for Recourse, 0.3)	0.018	1.0	-	1.0	0.001	0.006	0.001	0.001	0.017	1.0	0.105	0.001
(Equal Choice for Recourse, 0.7)	1.0	0.364	0.857	-	0.814	0.528	0.813	0.81	1.0	0.882	0.451	0.34
Equal Effectiveness	0.018	1.0	0.214	1.0	-	0.003	0.0	0.0	0.017	1.0	0.058	0.0
(Equal Effectiveness within Budget, 5.0	0.018	1.0	0.857	1.0	0.006	-	0.004	0.006	0.017	1.0	1.0	0.006
(Equal Effectiveness within Budget, 10.0)	0.018	1.0	0.214	1.0	0.0	0.002	-	0.0	0.017	1.0	0.047	0.0
(Equal Effectiveness within Budget, 18.0)	0.018	1.0	0.214	1.0	0.0	0.003	0.0	-	0.017	1.0	0.058	0.0
(Equal Cost of Effectiveness(Micro), 0.3)	0.238	1.0	0.857	1.0	0.136	0.452	0.263	0.215	-	1.0	0.462	0.155
(Equal Cost of Effectiveness(Micro), 0.7)	0.611	0.279	0.864	0.771	0.336	0.621	0.449	0.402	0.7	-	0.465	0.295
Equal(Conditional) Mean Recourse	0.996	1.0	1.0	1.0	0.723	0.946	0.875	0.777	0.997	1.0	-	0.83
(Fair Effectiveness-Cost Trade-Off, value)	0.018	1.0	0.214	1.0	0.0	0.002	0.0	0.0	0.017	1.0	0.047	-

A.5.4 Comparison on Adult Dataset (Race as Protected Attribute)

For the Adult dataset with 'race' as the protected attribute, the test set contained 10,205 affected individuals. These individuals were split into affected whites (D_1) and affected non-whites (D_0). Applying the fp-growth algorithm with a 1% support threshold on D_1 and on D_0 and computing the intersection of the results produced 16,621 subpopulation groups. The fairness metrics were then used to evaluate and rank these groups based on the actions applied.

Table A.7: Ranking Analysis in Adult (protected attribute race)

	# Most Unfair Subgroups	# Subgroups w. Bias against Whites (in Top 10% Unfair Subgroups)	# Subgroups w. Bias against Non-Whites (in Top 10% Unfair Subgroups)
(Equal Cost of Effectiveness(Macro), 0.3)	1731	0	295
(Equal Cost of Effectiveness(Macro), 0.7)	325	7	51
(Equal Choice for Recourse, 0.3)	1	2	391
(Equal Choice for Recourse, 0.7)	2	10	60
Equal Effectiveness	1	6	1433
(Equal Effectiveness within Budget, 1.15	1	50	24
(Equal Effectiveness within Budget, 10.0	1	3	1251
(Equal Effectiveness within Budget, 21.0)	1	0	1423
(Equal Cost of Effectiveness(Micro), 0.3)	1720	0	294
(Equal Cost of Effectiveness(Micro), 0.7)	325	7	51
Equal(Conditional Mean Recourse)	2545	53	1316
(Fair Effectiveness-Cost Trade-Off, value)	2	0	0

Table A.8: Aggregated Rankings in Adult (protected attribute race)

	(Equal Cost of Effectiveness (Macro), 0.3)	(Equal Cost of Effectiveness (Macro), 0.7)	(Equal Choice for Recourse, 0.3)	(Equal Choice for Recourse, 0.7)	Equal Effectiveness	(Equal Effectiveness within Budget, 1.15)	(Equal Effectiveness within Budget, 10.0)	(Equal Effectiveness within Budget, 21.0)	(Equal Cost of Effectiveness (Micro), 0.3)	(Equal Cost of Effectiveness (Micro), 0.7)	Equal(Conditional Mean Recourse)	(Fair Effectiveness-Cost Trade-Off, value)
(Equal Cost of Effectiveness(Macro), 0.3)	-	1.0	0.845	1.0	0.162	0.996	0.283	0.177	0.026	1.0	0.448	0.194
(Equal Cost of Effectiveness(Macro), 0.7)	0.7	-	0.9	0.829	0.147	0.973	0.315	0.169	0.698	0.05	0.421	0.12
(Equal Choice for Recourse, 0.3)	0.419	1.0	-	1.0	1.0	1.0	0.03	1.0	0.419	1.0	0.782	0.073
(Equal Choice for Recourse, 0.7)	1.0	0.095	0.909	-	0.644	1.0	0.003	0.328	1.0	0.1	0.041	0.011
Equal Effectiveness	0.023	1.0	0.909	1.0	-	1.0	0.01	0.0	0.023	1.0	0.0	0.0
(Equal Effectiveness within Budget, 1.15	1.0	0.048	1.0	0.857	0.069	-	0.047	0.07	1.0	0.05	1.0	0.102
(Equal Effectiveness within Budget, 10.0	0.395	0.048	0.818	0.571	0.001	1.0	-	0.001	0.395	0.05	0.611	0.002
(Equal Effectiveness within Budget, 21.0)	0.023	1.0	0.909	1.0	0.0	1.0	0.01	-	0.023	1.0	0.0	0.0
(Equal Cost of Effectiveness(Micro), 0.3)	0.023	1.0	0.845	1.0	0.162	0.996	0.284	0.177	-	1.0	0.449	0.195
(Equal Cost of Effectiveness(Micro), 0.7)	0.7	0.048	0.9	0.829	0.147	0.973	0.315	0.169	0.698	-	0.421	0.12
Equal(Conditional) Mean Recourse	0.979	1.0	1.0	1.0	0.628	1.0	0.778	0.633	0.979	1.0	-	0.721
(Fair Effectiveness-Cost Trade-Off, value)	0.023	1.0	0.818	1.0	0.001	1.0	0.012	0.001	0.023	1.0	0.003	-

A.5.5 Comparison on COMPAS Dataset

In the COMPAS dataset, 745 individuals were affected in the test set. These individuals were split into affected Caucasians (D_1) and affected African-Americans (D_0), producing 995 subpopulation groups using fp-growth with a 1% support threshold and intersecting the results. Fairness metrics were applied to evaluate and rank these groups and their corresponding actions.

Table A.9: Ranking Analysis in COMPAS

	# Most Unfair Subgroups	# Subgroups w. Bias against Caucasians (in Top 10% Unfair Subgroups)	# Subgroups w. Bias against African-Americans (in Top 10% Unfair Subgroups)
(Equal Cost of Effectiveness(Macro), 0.3)	51	0	11
(Equal Cost of Effectiveness(Macro), 0.7)	46	0	6
(Equal Choice for Recourse, 0.3)	13	12	8
(Equal Choice for Recourse, 0.7)	15	8	6
Equal Effectiveness	1	14	37
(Equal Effectiveness within Budget, 1.0)	4	16	30
(Equal Effectiveness within Budget, 10.0)	1	20	39
(Equal Cost of Effectiveness(Micro), 0.3)	51	0	11
(Equal Cost of Effectiveness(Micro), 0.7)	46	0	6
Equal(Conditional Mean Recourse)	37	19	24
(Fair Effectiveness-Cost Trade-Off, value)	5	18	62

Table A.10: Aggregated Rankings in COMPAS

	(Equal Cost of Effectiveness (Macro), 0.3)	(Equal Cost of Effectiveness (Macro), 0.7)	(Equal Choice for Recourse, 0.3)	(Equal Choice for Recourse, 0.7)	Equal Effectiveness	(Equal Effectiveness within Budget, 1.0)	(Equal Effectiveness within Budget, 10.0)	(Equal Cost of Effectiveness (Micro), 0.3)	(Equal Cost of Effectiveness (Micro), 0.7)	Equal(Conditional Mean Recourse)	(Fair Effectiveness-Cost Trade-Off, value)
(Equal Cost of Effectiveness(Macro), 0.3)	-	1.0	0.65	1.0	0.169	0.801	0.398	0.2	1.0	0.797	0.226
(Equal Cost of Effectiveness(Macro), 0.7)	0.96	-	0.925	0.625	0.127	0.518	0.236	0.96	0.2	0.52	0.149
(Equal Choice for Recourse, 0.3)	0.32	0.76	-	0.775	0.082	1.0	0.178	0.32	0.76	0.297	0.116
(Equal Choice for Recourse, 0.7)	0.9	0.46	0.8	-	0.424	0.484	0.057	0.9	0.46	0.259	0.045
Equal Effectiveness	0.2	1.0	0.75	1.0	-	1.0	0.003	0.2	1.0	0.003	0.002
(Equal Effectiveness within Budget, 1.0)	0.8	1.0	0.75	0.75	1.0	-	1.0	0.8	1.0	0.413	0.002
(Equal Effectiveness within Budget, 10.0)	0.2	1.0	0.75	1.0	0.003	1.0	-	0.2	1.0	0.003	0.002
(Equal Cost of Effectiveness(Micro), 0.3)	0.2	1.0	0.65	1.0	0.169	0.801	0.398	-	1.0	0.797	0.226
(Equal Cost of Effectiveness(Micro), 0.7)	0.96	0.2	0.925	0.625	0.127	0.518	0.236	0.96	-	0.52	0.149
Equal(Conditional) Mean Recourse	0.98	1.0	1.0	1.0	0.507	0.772	0.512	0.98	1.0	-	0.627
(Fair Effectiveness-Cost Trade-Off, value)	0.68	1.0	0.75	0.8	0.801	0.202	0.801	0.68	1.0	0.331	-

A.5.6 Comparison on SSL Dataset

For the SSL dataset, the test set contained 11,343 affected individuals. These were split into affected Blacks (D_1) and affected Whites (D_1), based on the 'race' attribute (named 'RACE CODE CD' in the dataset itself). The fp-growth algorithm, with a support threshold of 1%, and the subsequent intersection generated 6,551 subpopulation groups in total. These were then evaluated and ranked by the fairness metrics.

Table A.11: Ranking Analysis in SSL

	# Most Unfair Subgroups	# Subgroups w. Bias against Whites (in Top 10% Unfair Subgroups)	# Subgroups w. Bias against Blacks (in Top 10% Unfair Subgroups)
(Equal Cost of Effectiveness(Macro), 0.3)	371	10	107
(Equal Cost of Effectiveness(Macro), 0.7)	627	26	124
(Equal Choice for Recourse, 0.3)	1	108	184
(Equal Choice for Recourse, 0.7)	16	78	229
Equal Effectiveness	1	15	389
(Equal Effectiveness within Budget, 1.0)	18	18	436
(Equal Effectiveness within Budget, 2.0)	2	19	532
(Equal Effectiveness within Budget, 10.0)	1	15	548
(Equal Cost of Effectiveness(Micro), 0.3)	458	5	135
(Equal Cost of Effectiveness(Micro), 0.7)	671	23	130
Equal(Conditional Mean Recourse)	100	41	434
(Fair Effectiveness-Cost Trade-Off, value)	80	76	544

A.5.7 Comparison on Ad Campaign Dataset

The Ad Campaign dataset had 273,773 affected individuals in the test set, which were split into affected males (D_1) and affected females (D_0) based on the 'gender' attribute. Running fp-growth with a 1% support threshold separately on males and females and intersecting

Table A.12: Aggregated Rankings in SSL

	(Equal Cost of Effectiveness (Macro), 0.3)	(Equal Cost of Effectiveness (Macro), 0.7)	(Equal Choice for Recourse, 0.3)	(Equal Choice for Recourse, 0.7)	Equal Effectiveness	(Equal Effectiveness within Budget, 1.0)	(Equal Effectiveness within Budget, 2.0)	(Equal Effectiveness within Budget, 10.0)	(Equal Cost of Effectiveness (Micro), 0.3)	(Equal Cost of Effectiveness (Micro), 0.7)	Equal(Conditional Mean Recourse)	(Fair Effectiveness-Cost Trade-Off, value)
(Equal Cost of Effectiveness(Macro), 0.3)	-	0.883	0.854	0.988	0.216	0.401	0.285	0.238	0.3	0.843	0.678	0.338
(Equal Cost of Effectiveness(Macro), 0.7)	0.929	-	0.877	0.725	0.239	0.421	0.332	0.264	0.871	0.314	0.829	0.342
(Equal Choice for Recourse, 0.3)	0.143	1.0	-	1.0	0.328	0.704	0.464	0.368	0.143	1.0	0.727	0.601
(Equal Choice for Recourse, 0.7)	1.0	0.167	0.769	-	0.083	0.177	0.127	0.086	1.0	0.143	0.926	0.135
Equal Effectiveness	0.143	0.167	0.923	0.938	-	0.002	0.0	0.0	0.143	0.143	0.0	0.003
(Equal Effectiveness within Budget, 1.0)	0.857	0.833	0.854	0.881	0.89	-	0.923	0.876	0.857	0.857	0.327	0.0
(Equal Effectiveness within Budget, 2.0)	0.286	0.333	0.923	0.938	0.5	0.002	-	0.5	0.286	0.286	0.0	0.003
(Equal Effectiveness within Budget, 10.0)	0.143	0.167	0.923	0.938	0.0	0.002	0.0	-	0.143	0.143	0.0	0.003
(Equal Cost of Effectiveness(Micro), 0.3)	0.443	0.833	0.877	0.969	0.143	0.312	0.198	0.154	-	0.843	0.729	0.268
(Equal Cost of Effectiveness(Micro), 0.7)	0.9	0.383	0.892	0.788	0.203	0.406	0.299	0.225	0.886	-	0.816	0.327
Equal(Conditional) Mean Recourse	0.6	0.733	0.946	0.969	0.244	0.464	0.395	0.378	0.514	0.729	-	0.396
(Fair Effectiveness-Cost Trade-Off, value)	0.971	0.967	0.838	0.869	0.967	0.774	0.977	0.96	0.971	0.971	0.837	-

the results, we generated 1,432 subpopulation groups, which were then evaluated using the fairness metrics.

Table A.13: Ranking Analysis in Ad Campaign

	# Most Unfair Subgroups	# Subgroups w. Bias against Males (in Top 10% Unfair Subgroups)	# Subgroups w. Bias against Females (in Top 10% Unfair Subgroups)
(Equal Cost of Effectiveness(Macro), 0.3)	427	0	44
(Equal Cost of Effectiveness(Macro), 0.7)	264	0	26
(Equal Choice for Recourse, 0.3)	2	1 0	66
(Equal Choice for Recourse, 0.7)	384	0	39
Equal Effectiveness	15	0	123
(Equal Effectiveness within Budget, 1.0)	1	0	42
(Equal Effectiveness within Budget, 5.0)	10	0	114
(Equal Cost of Effectiveness(Micro), 0.3)	427	0	44
(Equal Cost of Effectiveness(Micro), 0.7)	264	0	26
Equal(Conditional Mean Recourse)	108	9	74
(Fair Effectiveness-Cost Trade-Off, value)	15	0	128

Table A.14: Aggregated Rankings in Ad Campaign

	(Equal Cost of Effectiveness (Macro), 0.3)	(Equal Cost of Effectiveness (Macro), 0.7)	(Equal Choice for Recourse, 0.3)	(Equal Choice for Recourse, 0.7)	Equal Effectiveness	(Equal Effectiveness within Budget, 1.0)	(Equal Effectiveness within Budget, 5.0)	(Equal Cost of Effectiveness (Micro), 0.3)	(Equal Cost of Effectiveness (Micro), 0.7)	Equal(Conditional Mean Recourse)	(Fair Effectiveness-Cost Trade-Off, value)
(Equal Cost of Effectiveness(Macro), 0.3)	-	0.7	0.483	0.6	0.167	1.0	0.276	0.25	0.7	0.487	0.154
(Equal Cost of Effectiveness(Macro), 0.7)	0.25	-	0.35	0.333	0.082	1.0	0.21	0.25	0.5	0.506	0.079
(Equal Choice for Recourse, 0.3)	0.25	1.0	-	1.0	0.73	1.0	1.0	0.25	1.0	0.037	0.338
(Equal Choice for Recourse, 0.7)	0.5	0.65	0.333	-	0.296	0.851	0.385	0.5	0.65	0.566	0.273
Equal Effectiveness	0.25	0.5	0.333	0.333	-	1.0	0.205	0.25	0.5	0.002	0.001
(Equal Effectiveness within Budget, 1.0)	1.0	1.0	1.0	1.0	0.714	-	0.671	1.0	1.0	0.305	0.395
(Equal Effectiveness within Budget, 5.0)	0.25	0.5	0.333	0.333	0.001	1.0	-	0.25	0.5	0.002	0.001
(Equal Cost of Effectiveness(Micro), 0.3)	0.25	0.7	0.483	0.6	0.167	1.0	0.276	-	0.7	0.487	0.154
(Equal Cost of Effectiveness(Micro), 0.7)	0.25	0.5	0.35	0.333	0.082	1.0	0.21	0.25	-	0.506	0.079
Equal(Conditional) Mean Recourse	0.525	0.75	0.65	0.7	0.25	1.0	0.408	0.525	0.75	-	0.267
(Fair Effectiveness-Cost Trade-Off, value)	0.25	0.5	0.333	0.333	0.001	1.0	0.205	0.25	0.5	0.002	-