Orijinal Kaynak: Binns, Reuben. (2018). "Fairness in Machine Learning: Lessons from Political Philosophy", *Journal of Machine Learning Research 81*: pp. 1–11.

Atıf Şekli: Binns, Reuben. (2021, Mart 21). "Yapay Öğrenmede Adalet: Siyaset Felsefesinden Dersler", Çev. Alp Eren Yüce, *Sosyal Bilimler*, Link: sosyalbilimler.org/yapay-ogrenme-adalet

Yapay Öğrenmede Adalet: Siyaset Felsefesinden Dersler

Reuben Binns
Türkçesi: Alp Eren Yüce

Özet

Bir yapay öğrenme için "adalet", uygulanabilirlik açısından ne anlama gelmektedir? Adalet herkesin bazı faydaları eşit olasılıkta kazanmasını mı garanti etmelidir yoksa bunun yerine en az avantaja sahip olan grubun göreceği zararı azaltmayı mı hedeflemeliyiz? İlgili ideal, ayrımcılığın belirli bir sosyal şeklinin var olmadığı ilişkilerde yer alan bazı alternatif durumlara atıfta bulunularak belirlenebilir mi? Son dönemde literatürde teklif edilen çeşitli tanımlar, ayrımcılık ve adalet gibi terimlerin ne demek olduğuna dair kabullerde bulunur, bunların matematiksel terimler ile nasıl tanımlanacağı ile ilgili farklı varsayımlar yürütür. Ayrımcılığa, eşitlikçiliğe ve adalete ilişkin sorular, bu merkezi kavramları kurumlaştırmak ve savunmak adına kayda değer şekilde çaba harcayan ahlak ve siyaset felsefecilerinin önemli derecede ilgilendirir. Bununla beraber, yapay öğrenmede adaleti formül hâline getirme girişimlerinin bu eski felsefi tartışmaların yankılarını barındırması şaşırtıcı değildir. Bu makale adil yapay öğrenme hakkında ortaya çıkan tartışmaları açıklığa kavuşturmak için ahlak ve siyaset felsefesindeki mevcut tartışmalardan faydalanır.

Anahtar Kelimeler: Adalet, Ayrımcılık, Yapay Öğrenme, Algoritmik Karar Verme, Eşitlikçilik

1. Giriş

Yapay öğrenme, reel dünyadan etiketlenen veriler ile oluşturulan deneme modelleri sayesinde yaşamdaki fenomenleri önceden tahmin edebilmemiz ve sınıflandırabilmemiz için imkânlar sunmaktadır. Bu modellerin çıktıları temel alınarak bir sonuç ortaya koyacak kararlar alındığında, ayrımcılık ve adalet problemlerine ilişkin soru işaretleri kaçınılmaz şekilde ortaya çıkar. Eğer bu modellerin çıktıları bazı ırk, cinsiyet ve din gruplarına karşı sistematik bir şekilde yanlı kararların ortaya çıkmasını sağlıyorsa, o zaman durum ne

olacaktır? Eğer reel dünyada ayrımcılığı vurgulayan yapılar var ise benzer ayrımcılıklar öğrenme süreci içine dâhil olacaktır. Bu da belirli grupların kredi alımı, sigorta başvuruları, mesleki olanakları bakımından adaletsiz bir şekilde reddedilmelerine yol açabilecektir. Bu problemin farkına varılması sonucu, adaletsizliğin tespit edilmesi ve azaltılmasını sağlayan "ayrımcılık-bilincinde veri madenciliği" ve "adaletli yapay öğrenme" şeklinde iki araştırma paradigması ortaya konmuştur (Hajian, Domingo Ferrer, 2013; Kamiran, vd., 2013; Barojas ve Selbst, 2016).

Ortaya konan bu çaba hemen ardından bir formül üretme ihtiyacını doğurmuştur. Ancak yapay öğrenme için uygulanabilir olan "adil olma" ve "ayrımcı olmama" kavramları ne ifade etmektedir? Bu konuda birçok ölçüt sunulmuştur. En yaygın tema koruma altındaki gruplar ile koruma altında olmayan gruplara karşı gösterilen muameleler arasındaki farkların karşılaştırılmasıdır, ancak bu ölçütleri uygulamak için birçok farklı yol vardır. En temel yol gruplar arasında pozitif/negatif sınıflandırma oranlarındaki genel ortalamayı göz önünde bulunduran "farklı etki" veya "istatistiki/demografik denklik" yöntemidir. Ancak bu yöntem de meşruiyet zemininde açıklanması güç olan ayrımcılık kavramına net bir değer biçmekte başarılı olamadığı için verimsiz bir durumdadır (Dwork, vd., 2012). Örneğin yeniden suç işleme oranları, tahmin sistemi içinde kadın ve erkeğe eşit oranda yaptırım uygulama girişiminde bulunulduğunda, eğer erkek daha yüksek oranda yeniden suç işlemişse, kadınların yeniden suç işleme ihtimallerinin düşük olmasına rağmen hapishanede daha uzun süre geçirmek durumda kalacaktır.

Bu ölçütlerin yanında daha detaylı işlenmiş bazı ölçü modelleri de vardır. Bunlar: Her bir grup üzerinde genel bir "doğruluk" tahmin modelini göz önüne alan "Doğru eşitlik" deseni (Angwin, vd., 2016); tahmini modelin doğruluğunu her bir grup için göz önünde bulunduran "koşulsal doğru eşitlik" deseni (Dieterich, vd., 2016); her bir grubun, o gruba verilen asıl oranlar üzerinden arzulanan düzeyde eşit olarak çıktı verip veremeyeceğini denetleyen "fırsat eşitliği" deseni (Hardt, vd., 2016) ve gruplar arasındaki haksız pozitif oranlardaki farklılıkları dikkate alan "farklı muamele" modelidir (Zafar, vd., 2017). Bir başka tanım ise korunan grup üyeleri yerine korunmayan grup üyelerinin bulunduğu hayali senaryoları kapsamaktadır (Kasner, vd., 2017). Sonuçlar farklı olduğu ölçüde ise sistem adaletsiz olacaktır. Örneğin bir adalet sistemi tarafından sınıflandırılan bir kadın, aynı eylemde bulunduğu hayali bir senaryoda erkek olma durumunda da aynı şekilde sınıflandırılmalıdır.

İdeal olarak bir sistemden beklenen tüm bu durumları sezgisel anlamda makul görülebilecek bir adalet duygusu eşliğinde karşılayabilmesidir. Ancak bir şekilde problemli olan belirli ölçütler hem nadir görülen hem de kurgusal olan durumlarda, matematiksel olarak tatmin edici olmaktan çok uzaktadır (Kleinberg, vd., 2016). Bu sebeple adalet metrikleri (ölçümleri) arasındaki zorlu seçimler, teknik çalışmada adaletsizliğe ilişkin yapılan tespit ve hafifletme çalışmaları yürütülmeden önce yapılmış olmalıdır. Vurgulanan başka bir gerilim ise adaletin benzer insanlar üzerinde benzer şekillerde uygulanması

problemidir. Ama bu gerilim sıklıkla, ideal şekilde benzerlik gösteren gruplar arasındaki hedef değişkenler için temel alınan oranlar farklı olduğu durumlarda ortaya çıkar (Dwork, vd., 2012). Dolayısıyla adil yapay öğrenme bu gibi uç noktalardaki bir grup kavramsal etik zorluklarla karşılaşmaktadır. Acaba bu zorluklar belirlenen bağlamda en uygun olan adil ölçümler midir? Hangi değişkenler farklılaşan uygulamalar için meşru zeminleri oluşturur? Bunların nedeni nedir? Bütün eşitsiz örnekler gruplar arasında nesnel olabilmekte midir? Adalet herkesin bir takım çıkarlara mümkün olabilecek en eşit oranda sahip olabilmesini mi garanti etmelidir, ya da en dezavantajlı gruba en az zararı vermeyi mi hedeflemeliyizdir? Buna benzer dengeleme girişimlerinde, karar vericiler yalnızca karar bağlamında empoze edilen kâr-zarar dengesini mi göz önüne almalıdırlar, yoksa başka bağlamlarda yer alan karar verme ve süje arasındaki ilişkiye de değinmeli midirler? Geçmiş, gelecek veya nesiller arası adaletsizlik konuyla nasıl ilişkilendirilebilir?

Ayrımcılığa ve adalete ilişkin bu gibi merkezi sorular, bu merkezi kavramlar hakkında önemli çalışmalar yapan, formüller oluşturan, kavramları ayrıştıran ve tartışan siyaset ve etik filozofları için uzun süredir ilgi odağı sorular olmuştur. Bu sebeple adalet kavramını yapay öğrenme için formül hâline getirme çabalarının bu eski felsefi tartışmalar eşliğinde yankı uyandırması şaşırtıcı değildir. Aslında FAT-ML (*Fairness, Accountability, and Transparency in Machine Learning* — Yapay Öğrenmede Adalet, Hesap Verebilirlik ve Şeffaflık) topluluğundaki bazı taslak çalışmalar ad-hoc yolundaki sınırlılığa rağmen siyaset felsefesine açıkça bir ilham kaynağı olarak atıfta bulunur.[1] Daha kapsamlı gözden geçirmeler ise bu soruların yanıtlanması adına etik karar verme algoritması için takip edilebilecek zengin argümanlar sağlayabilecektir. Bu makale, ortaya çıkan tartışmayı ayrımcılık-duyarlı ve adil yapay öğrenme literatürü içinde açıklığa kavuşturmak ve konumlandırmak için ayrımcılık, adalet ve eşitlikçilikle ilişkili felsefi literatüre genel bir bakış sunmayı amaçlamaktadır. Araştırma boyunca adil yapay öğrenme literatüründe sıkça kullanılan terimler —"ayrımcılık" ve "adalet" de dâhil olmak üzere— ile felsefi literatürde kullanılan ilgili terimler arasında yapılan kavramsal ayrımları ele almayı hedefliyorum. Buradaki amaç yalnızca iki söylem arasındaki benzerlik ve farkları incelemek değildir. Bununla beraber gelecekteki algoritmik adalet araştırmalarına yardım sağlayabilmek ya da henüz irdelenmemiş ilişkili problemler üretebilmek için felsefi tartışmaların soruşturma alanını belirlemek de bu çalışmanın hedefleri arasındadır.

Konuyu tartışmaya felsefi düşüncelerin ayrımcılık tanımlarının ne olduğu, ayrımcılığın neden yanlış olduğu ve eğer yanlışsa ne zaman yanlış olacağı soruları ile başlıyorum. Belirli düşünce tiplerinde ayrımcılığa nelerin sebep olduğunu, önerilen koşulların algoritmik karar verme bağlamında kullanılmasının olası olmadığını gösteriyorum. Eğer doğruysa, bu açıklamalar algoritmik karar vermenin her zaman ahlaki açıdan zararsız olduğu anlamına gelmeyecektir. Yalnızca potansiyel hatanın geleneksel

[1] Dikkate değer örnekler arasında H. Peyton Young, John Rawls ve John Roemer gibi yazarların çalışmalarına yapılan atıflar yer alır (Dwork, vd., 2012; Joseph, vd., 2016).

anlayışta yer alan ayrımcılık düşüncesinde bulunamayacağı anlamına gelecektir. Bu da bizi algoritmik karar verme konusunda problematik olan, eşitlikçilik düşüncesi ideallerini göz önünde bulundurup çeşitli bağlantıları ele alan —ki bunlar temel insani eşitlik hisleri içinde belirli eşitsizlik biçimlerinden olabildiğince kaçınmaya çalışan düşünce biçimleridir— diğer tartışma zeminlerini de göz önünde bulundurmaya itecektir. Bu tartışma, adil yapay öğrenme topluluğunda kullanılan "adalet" kavramının, çeşitli normatif eşitlikçi düşünceler içinde yer edinmiş bir terim olarak en iyi şekilde anlaşılabilen kavram olduğunu söyler. Bilhassa, eşitlikçilik geniş çapta desteklenir bir ilke iken, tam olarak gereksinimi duyulan şey bu kavramın daha çok münazaranın konusu olmasıdır. Bu noktada ben de bu tartışmalardan bazılarına ilişkin genel bir değerlendirme sundum, algoritmik karar verme sistemlerine "adalet" kavramını dâhil eden uygulamalar ile tartışmamı tamamladım.

2. Ayrımcılık Nedir ve Ayrımcılığı Yanlış Kılan Nedir?

Adalet kavramını kısıtlayıp yapay öğrenmenin içine nasıl yerleştirileceğini keşfeden erken dönem çalışmaları "adil" terimindense "ayrımcılık-duyarlı" terimini kullanmışlardır. Bu terminolojik fark, bilgisayar bilimciler için görece önemsiz görünse de konu hakkında yazılmış felsefi metinlerin endişelerini taşıyan, önemli ayrımlara işaret etmektedir. Ahlaki fenomenlerin kavramsal olarak uygun kategorilere ayrılması filozoflar için memnun edicidir. Ayrıca bu ayrım kafa karıştıran durumlar için de faydalı bazı açıklamalar getirir. Bu sebeple ayrımcılığın hangi unsurları kapsadığına ilişkin daha yakından kavramsal bir anlayış edinmek, özel olarak neyin ayrımcılığı yanlış yaptığını (eğer herhangi bir yanlış varsa) ve ayrımcılığın adalet gibi diğer ahlaki kavramlarla olan ilişkisini netleştirmekte yardımcı olabilecektir. Bu da algoritmik sistemlerde ayrımcılık olup olmadığını, ne zaman ayrımcılık olduğunu ve bu konuyla ilgili neler yapılabileceğini açıklayacaktır. Bu durumda düşünülen "algoritmik ayrımcılığın" klasik ayrımcılık formlarından önemli şekillerde ayrıldığı ve bu ayrım konusunda farklı karşıt-önemlerin uygun olduğu ortaya çıkabilecektir.

2.1. Psikolojik Durum Tarifi

Ayrımcılığın paradigma problemi dikkat çekici sosyal grupların üyeliğini temel alan ayrık muameleleri kapsar –örneğin karar verme gücü ile cinsiyet ya da "ırk" gibi gruplar için zarar ya da kâr dağıtımı. Örnekler; iş başvurusunda bulunan erkek adayları eşit niteliklere sahip kadın adaylara oranla daha çok tercih eden işverenleri veya bazı öncelikli çoğunluk gruplarla karşılaştırıldığında belirli azınlık grupların şartlı tahliye mahkûmlarına daha katı koşullar empoze eden şartlı tahliye memurlarını kapsamaktadır. Bu tip paradigma problemlerine odaklanmak ve ayrımcılığı neyin yanlış yaptığına dair bir dizi açıklamaya sahip olmak, karar vericinin niyetine, inancına ve değer yargısına vurgu yapar. Diğerleri arasında Richard Arneson, Thomas Scanlon ve Annabelle Lever tarafından savunulan bu tür psikolojik durum tanımları için, karar vericinin belirli sosyal gruplara karşı sistematik düşmanlık veya tercihlerinin varlığı, ayrımcılığı yanlış yapan şeydir (Arneson, 1989; Lever,

2016; Scanlon, 2009). Bu tür endişeler, karar vericinin kusurlu ahlaki karakteri açısından —örneğin, karar verici kötü niyet ya da düşmanca tavır gösterebilir— veya bu niyetlerin ayrımcılığa yönelik yıkıcı etkileri açısından —örneğin saygı eksikliğinin yarattığı aşağılama durumu gibi— ifade edilebilir.

Böyle bir niyeti olmayan, ancak yine de kazara, farkında olmaksızın bu tür eşitsizlikler yaratan bir karar verici, muhtemelen haksız ayrımcılıktan suçlu olmayacaktır (bu durum başka tartışma zeminlerinde problematik olsa bile). Eğer karar vericinin ilk durumdaki kişilere yönelik, söz konusu eşitsizlikleri öngörmedeki ya da eşitsizlik belirgin hâle geldiğinde durumu telafi etmedeki başarısızlığı, ahlaken benzer ve gerekli şekilde itiraz edilebilir ise; bu gibi vakalar —ki genellikle Birleşik Krallık'ta dolaylı veya kurumsal ayrımcılık ya da ABD'de eşitsizlik etkisi olarak adlandırılırlar— hâlen ayrımcılığın psikolojik durum anlatımı için ayrımcı olarak sayılabilirler. Bununla beraber eğer bu tip durumlara ulaşılmamışsa, dolaylı ayrımcılık yanlış olabildiğinde, ayrımcılığın bir örneği olarak tanımlanamayacaktır (Eidelson, 2015).

Bu düşünce tarzı algoritmik ayrımcılık tasarımı açısından potansiyel zorluklar ortaya çıkarır. Karar vericilerin belirli psikolojik durumlara sahip olması, bir kararın ayrımcı olması için gerekli bir koşul ise, algoritmik karar verme sistemlerinin asla bu şekilde ayrımcı olamayacağı, çünkü bu tür sistemlerin ilgili psikolojik durumlar için elverişsiz oldukları iddia edilebilir. Bu araştırma makine bilinci olasılığının tartışıldığı bir yer değildir, ama bu bilincin (henüz) gerçek olmadığını varsaydığımızda, AI (*artificial intelligence* — yapay zekâ) ve özerk karar verme sistemlerinin; aşağılama, düşmanlık veya saygısızlık gibi karar verme için ayrımcılığın nitelenmesini gerektiren durumlarda birer dayanak noktası oluşturamayacağı görülür.

Bununla beraber, psikolojik durum düşüncesi savunucuları benzer şekilde algoritmik karar verme sistemlerinin, algoritmanın zihinsel durumlarla ilişkilendirilmediği noktalarda ayrımcılık içerebileceğini iddia edebilirler. Öncelikle şunu iddia edebilirler: Eğer karar verme modellerinden sorumlu insan, karar vericiler ve veri bilimciler, sistemin bilinçli şekilde yanlı sonuçlar üretmesine sebebiyet verecek kötü niyetlere sahipseler veya modellerden kaynaklanan istenmeyen eşitsizlikleri ihmâl ederek görmezden geliyor ya da gözden kaçırıyorlarsa suçlu bulunabileceklerdir. İkinci olarak, sosyal epistemologların savunduğu gibi, bazen tek bir bireyden kaynaklanmayan, ancak çeşitli karmaşık şekillerde birden fazla bireysel yargının sonucu olarak bir araya gelen kararları hâlâ ahlaki olarak değerlendiriyor olabiliriz (Gilbert, 2004; Pettit, 2007). Ekonomi ve sosyal seçim teorisindeki yargı kümeleme problemlerinden yararlanan tartışmacılar, kurumsal karar alma senaryolarının uygun şekilde düzenlendiğinde, kişilerin kollektif yargılarından ötürü ahlaki olarak sorumlu tutulabileceklerini söyler. Bu sebeple önceden insan eliyle yapılmış yargılamaları içeren veriler üzerinden yapılandırılan makine öğrenme modelleri, eğer bu bireysel yargıların kendisi, bireylerin ayrımcı olan benzer niyetlerinin bir sonucu ise benzer zeminlerde eleştirilebilirler.

Ancak bu gibi özel vakalar bir yana, ayrımcılığın psikolojik durum düşüncesinin algoritmik karar verme içeriğine doğal olarak dönüştürülemeyeceği görülmektedir. Eğer bu düşünceler doğru ise, algoritmik karar verme potansiyel ve ahlaki olarak problemli iken, haksız ayrımcılık örneği olarak nitelendirilemeyecektir. Alternatif olarak, ayrımcılığın neden (bazen) yanlış olduğunu, karar vericinin psikolojik durum sorumluluğu üzerinden açıklamayan diğer düşünceler, ayrımcılığı algoritmik çeşitliliğe uyarlamak adına daha uygun olabilecektir.

2.2. İnsanlara Birey Olarak Muamele Etme Konusundaki Başarısızlık

Son dönemde yazılan makalelerde ayrımcılık konusunda dikkat çeken benzer bir düşünce, ayrımcılığı yanlış yapan şeyin —yanlışlığın doğrudan ya da dolaylı çeşitleriyle— bireyler hakkında yapılan genellemelerin üyesi oldukları gruplar üzerinden yapılması olduğunu söyler (Lippert-Rasmussen, 2014). Bu itiraz, genellikle istatistiki ayrımcılık olarak adlandırılan kavrama bir yanıt olarak ortaya çıkmıştır (Phelps, 1972). İstatistiki ayrımcılık gruplar hakkında yapılan istatistiki genellemenin, grup üyelerinin özellikleri ve gelecekte sergileyecekleri davranışların tahmini için kullanılmasıdır. Örneğin; bir işveren, sigara içenlerin sigara içmeyenlere göre genellikle daha az çalışkan olduğunu gösteren kanıtları okuyabilir ve sigara içen bir kişinin iş başvurusunu reddedip, işi daha üretken olması beklenen daha az nitelikli, sigara içmeyen bir kişiye verebilir.

Ekonomistlerin bu modellere dayalı olarak tartıştığı gibi, gruplar hakkındaki bu gibi genellemeler, kişiler hakkında doğrudan bir kanıtın eksik olduğu durumlarda, firmalar için riski azaltmak adına verimli aracılar olabilirler (Phelps, 1972). Yine de genellemelerin potansiyel verimlilik faydalarına rağmen, bu kullanım en azından bazı olaylarda yanlış olarak değerlendirilir. İstatistiki ayrımcılığın yanlışlığı hakkındaki sezgiler büyük ölçüde paylaşılırken, uygulamaya yönelik tutarlı itirazları ifade etmenin şaşırtıcı bir şekilde zor olduğu tecrübe edilmiştir. Özellikle müdahalenin kesin istatistiki analiz tarafından desteklenmediği, bir gruba üye olarak sergilenen bir davranışın karar vericinin dikkatini çekmesi ile ilişkilendirildiği durumlar gibi basit olmayan vakaların içine girildiğinde, bu zorluk kendisini iyice gösterir. Algoritmik karar verme steroitler için bir çeşit genelleme şekli olarak görüleceğinden, genellemenin yanlışlığına dayanan herhangi bir ayrımcılık açıklaması mevcut endişelerimiz konusunda özellikle geçerli olacaktır.

Popüler bir düşünceye göre içinde bulunan genellemeler bazı doğruları barındırıyor olsa da istatistiki ayrımcılık yanlıştır, çünkü sistem süje-karar ilişkisinde kişiye birey olarak muamele etmede başarısız olmaktadır. Örneğin; Müslümanların çoğunlukta olduğu ülkelerden gelen yolcuları terörizmin baskınlığına ilişkin genellemeler temelinde daha sıkı sınır kontrolüne tabi tutmak, her bir yolcuyu kendi faziletleri ile birey olarak değerlendirme konusunda başarısız olur. Benzer şekilde, sigara içenlerin ortalama olarak daha az üretken olduğuna dair kanıtlar nedeniyle sigara içen bir kişinin iş başvurusunu reddetmek (bunu tartışmak için yaptığımızı varsayalım), kişinin dâhil olduğu grubun diğer üyelerinin davranışı sonucu ortaya çıkacaktır. Bu da kişiyi haksız bir şekilde cezalandırmak

olacaktır. Bu gibi örneklerde görüldüğü gibi, istatistiki ayrımcılık yöntemi insanlara bireyler olarak muamele etme konusunda başarısız olması, bazı düşünürlerin bu yönteme karşı çıkmalarına neden olmaktadır. Eğer bu doğru ise, burada algoritmik karar verme sistemlerinin varlığına ilişkin ciddi bir zorluk var demektir. Çünkü bu sistemler tasarımları gereği insanlara bireyler olarak muamele edemezler. Aynı özelliklere sahip iki kişi ele alındığında, deterministik model bu kişiler için aynı sonuçları üretecektir. Bu sebeple bu sonuç belirtilen düşünce tarzı içinde doğası gereği ayrımcı olacaktır.

Bununla birlikte, diğer tartışmacılar, kişilere bireyler olarak muamele etmede başarısız olmanın, haksız ayrımcılığın esası olmadığını savunurlar (Schauer, 2009; Dworkin, 1981; Lippert-Rasmussen, 2014). Konuya dair endişelerden birisi, bu kriterin çok geniş kapsamlı olmasıdır. Zira bu kriter yalnızca insan hakları hukukunda yer alan cinsiyet, ırk, din vb. kategorileri kapsamakla kalmaz, her türlü gruba karşı genellemeler yapar. Bu gibi kategoriler haksız ayrımcılığa ilişkin endişeleri kolayca tetiklerken, "sigara kullanımı" gibi diğer kategoriler ise ayrımcılığa ilişkin endişeleri açıkça ortaya koymazlar. Bu da yetki-yetki konusunda ayrımcılık açısından fark yaratan şeyin yalnızca belirli gruplar hakkında genellemeler olduğunu göstermektedir. Diğerleri ise "birey olarak muamele edilme" düşüncesinin yanlış anlaşıldığına itiraz ederler; onlara göre daha yakından incelendiğinde, bireyi gözettiği düşünülen kararlar bile aslında genellemenin kılık değiştirmiş formlarını oluşturmaktadır (Schauer, 2009). Düşünün ki iş veren üreticiliğe yönelik "sigara kullanan/kullanmayan" tahmin yürütme yöntemini kullanmak yerine, adaylardan üreticiliği daha doğru bir şekilde tahmin eden bir testi geçmelerini istiyor; bu noktada Schauer'in de tartıştığı gibi, iş veren test skorlarından, kişinin iş başındaki davranışlarına kadar yine genellemelere güvenmek durumunda kalacaktır. Test skorları "sigara kullanan/kullanmayan" tahmin yürütme yönteminden daha uygun sonuçlar verebilir ama yine de bu sonuçlar da temelde bir çeşit genellemedir (Schauer, 2009, s. 68). Test mükemmel olmadığı sürece, testte kötü performans gösteren bazı adaylar, buna rağmen iş başında görece üretken olabilecektir.

Eğer bu eleştiri biçimi doğruysa, kişilere genellemeler temel alınarak dahil oldukları gruplara istinaden ayrımcı farklı muameleler yapılması bir vaka durumu oluşturmaz. Bu noktada yaygın şekilde genellemenin eleştirisinin ortaya çıkışı, aslında kesin olmayan genelleme araçlarının eleştirisine indirgenebilir. Eğer sınır güvenlik sistemleri (veya işe alım süreçleri) masum olan turistlere daha az yük oluşturacak şekilde, daha uygun tahminler ortaya koyabilirse, o zaman ayrımcılık suçlaması biraz güç kaybedebilir. Tabii ki daha uygun tahmin mekanizmaları yüksek ihtimalle daha pahalıdır ve buradaki takas genellemelerin zararı ve yararı arasında olacaktır; ama iki durumda da bu görüş anti-ayrımcılığın insanlara birey olarak muamele edilmesini gerektirmemektedir. Daha ziyade, istatistiki ayrımcılığa uğrayanların kim olduğu ve kaç kişilik bir grup şeklinde, üyesi oldukları grup üzerinden haksız şekilde yargılandıkları sorularına bağlı olarak, istatistiki ayrımcılık kabul edilebilir olmaktadır. Ayrıca buna genellemenin doğruluğunu geliştirici maliyetler de eklenmektedir. Eğer bu düşünce biçimi doğruysa, bu durum algoritmik

sistemlerin taraftarları için hoşnut bir vargı olacaktır, çünkü onlar kaçınılmaz şekilde genellemelerin farklı formlarını temel alırlar.

Şimdiye dek, psikolojik durum hatası veya genellemeler açısından, algoritmik ayrımcılık düşüncesinde bir takım zorluklar olduğunu takdim ettim. Bu zorluklardan bazıları ayrımcılığın felsefi açıklamalarında saklıdır. Diğer zorluklar ise insan ve algoritmik karar verici arasındaki benzer olmayan yapıdan kaynaklanmaktadır. Eğer (algoritmik) ayrımcılıktaki yanlışlık karar vericinin ahlaki açıdan şüpheli niyetlerinden veya insanlara birey olarak muamele etmedeki başarısızlıktan kaynaklanmıyorsa, o zaman neden kaynaklanıyor olabilir? Daha genel bir dizi eşitlikçi norm, algoritmik adalet teorisi için diğerlerinden teorilerden daha başarılı bir algoritmik adalet teorisi sunabilecektir.[2]

3. Egalitarianism (Eşitlikçilik)

Genel olarak değindiğimizde, *egalitarianism* (eşitlikçilik) insanlara eşit muamele edilmesi ve (bazen) belirli değerli şeylerin eşit olarak dağıtılması gerektiği fikridir. Ayrımcılığın hatalı olmasının eşitlikçilikle alâkalı olduğu çok açık şekilde görünüyor olabilir. Bununla beraber, belki de şaşırtıcı şekilde, yukarıda bahsedilen ayrımcılık kuramcıları bu bağlantıya direnç gösterirler. Bu teorisyenlerden birisi "anti-ayrımcılık yasaları ve eşitlikçilik arasındaki herhangi bir bağlantının, en iyi ihtimalle ihmâl edilebilir olduğunu ve her durumda bir gerekçe olarak gösterilmesinin yetersiz olduğunu" iddia eder (Holmes, 2005). Bu sırada, diğer teorisyenler ise aksini savunur. Onlara göre eşitlikçi normlara doğrudan bir başvuru, ayrımcılık konusunda yanlış olan her şeyi tatmin edici bir şekilde açıklayabilir (Segall, 2012). Bizim mevcut amaçlarımız için ise, bu tartışmadan güvenli bir şekilde kaçınılabilir. Bu kaçınma, tartışmanın felsefi proje olarak ilgi çekici olmaması ya da önemsiz olması sebebiyle değildir. Daha ziyade, buradaki amacımız eşitlikçi normların algoritmik sistemlerin neden ve ne zamanlarda adaletsiz olarak göz önüne alınabilecekleri sorusunu, düşünsel olarak nasıl yanıtlayabileceğine ilişkin bir irdeleme yapmaktır. Bu tip bir adaletsizliğin meşruiyete ilişkin ayrımcılığın bir formu olarak değerlendirilip değerlendirilemeyeceği, bizim sorunumuz değildir. Bu sebeple bu bölüm eşitlikçiliğe ilişkin bazı ana tartışmalar konusunda genel bir görüş sağlar, bu tartışmaların adil yapay öğrenme için önemini ortaya koyar.

3.1 Eşitlikçiliğin Değer Birimi ve Adaletin Alanları

Adalet sorunlarının ortaya çıktığı yapay öğrenme içeriklerinde, sistem bireyleri sürekli olarak olumlu ve olumsuz etkiye sahip olduğu varsayılan farklı çıktı sınıfları ile eşler. Örneğin onaylanan/reddedilen mali bir kredi, yüksek ya da düşük sigorta fiyatları, hapishanede daha fazla ya da daha az geçirilen süre bu sınıflama örneklerinden bazılarıdır. Bu çıktı sınıflarını, bir dereceye kadar eşit şekilde dağıtılması gereken temelde değerli bazı

[2] Bununla beraber, ayrımcılığın felsefi açıklamaları algoritmik kararlarda kolayca uygulanmasa da algoritmik bir sistemi ayrımcı olarak adlandırmak (veya özel olarak, cinsiyetçi, ırkçı gibi adlandırmak), onun retorik gücü ya da günlük hayatta kullanışlı bir söylev olmasıyla gerekçelendirilebilir.

nesne grupları için aracılar ya da engelleyiciler olarak varsayarız. Peki, ama tam olarak bu çıktılara yüklenen değerler bağlamında nedir bu eşitlikçiliğin değer birimi? Eşitlikçilik çeşitli rekabetçi görüşler içinde refah ve yapılabilirlik kavramları kapsamında açıklanmıştır. Bu noktada refah kavramı memnuniyet ya da tercihe dayalı tatmin (Cohen, 1989); gelir ve varlık gibi kaynaklar (Rawls, 2009; Dworkin, 1981) olarak anlaşılırken, yapılabilirlik ise, kesin olan şeyleri yapmak için gerekli kaynaklar ve beceriler olarak anlaşılmaktadır (Sen, 1992). Diğerlerinin önerisine göre ise vatandaşlar eşit politik ve demokratik statülere sahip olduğu sürece, refah, kaynaklar veya beceriler konusundaki eşitsizlikler kabul edilebilir eşitsizliklerdir (Anderson, 1999).

Eşitlikçiliğin ne olduğu sorusu (bazen atıfta bulunulduğu şekliyle "neyin eşitliği" tartışması), farklı algoritmik karar çıktılarının etkisine ilişkin varsayımlarımızla ilgilidir. Birçok vakada —kredinin otomatik şekilde paylaştırılması ya da sigorta prim ayarları gibi— karar çıktıları kaynakların dağıtımını birincil olarak etkilemektedir. Diğer vakalarda ise —algoritmik engel koyma veya kullanıcıyı çevrimiçi tartışmadan men etmek gibi— kararlar direkt olarak refah veya becerinin dağıtımıyla ilişkili olabilirler. Bu noktada eşitlikçiliğin her bir değer biriminin önemi bağlamlar arası anlaşılır farklılıklar gösterebilir, bu da algoritmik karar verme sistemlerinin, potansiyel diferansiyel etkilerini nasıl hesaba kattığımızı etkiler.

Bununla beraber bu tartışma, farklı insanların benzer sonuçlardaki çıktılarda yer alan zarar ve faydalara farklı şekillerde değerler atfedebileceği sezgisine dayalıdır. Yine de adil yapay öğrenme ile ilgili mevcut çalışmaların çoğu, farklı popülasyonlar arasındaki karar sonuçlarının tek tip değerlendirmesini benimser. Bu durum, bazı vakalarda farklı alt grupların olası eşit gruplar olarak varsayılmasını sağlayan güvenli bir yol olabilir. Bu sayede bu gruplar iyi ya da kötü şeklinde özel çıktı sınıfları olarak değerlendirilebilirler — örneğin kredinin reddedilmesi ya da işe alınma gibi. Ancak diğer vakalarda, özellikle derece sırası konusunda içinde evrensel olarak açık ortak bir konsensüs bulunmayan çoklu çıktı sınıfların olduğu, kişiselleştirme ve tavsiye sistemlerinde, bu varsayım kusurlu olabilir.

Bağlantılı bir tartışma, tek bir eşitlikçilik düşüncesinin farklı sosyal bağlamlar açısından uygulanıp uygulanmaması gerektiği sorusunu konu edinir veya aralarında yeniden dağıtımın uygun olmayabileceği ve farklı ölçülemez adalet mantığı içinde uygulanabilir içsel adalet alanlarının olup olmadığı sorusunu ele alır (Walzer, 2008). Örneğin seçimlerde oy kullanmak gibi mülki ve demokratik haklar göz önüne alındığında, eşitlikçiliğin amacı sadece eşitlik için yarışmayı sağlayacak fırsat eşitliğinin sağlanmasından ziyade, yararın tam olarak eşit şekilde dağıtımıdır. Bu fikir mesleki testler doğruyken, seçmen kayıt testlerinin doğru olmadığı sezgisini açıklar. Oylamanın yapılması için bir tür testin öncelik olarak talep edilmesi, testi almak konusunda herkesin eşit fırsata sahip olacağı hissi eşliğinde, fırsat eşitliğini garanti altına alabilir. Ama yetenek ve çaba eşit olarak dağıtılmadığı için, bazı insanlar testte başarısız olacaktır ve sonuçların eşit olmaması durumu oluşacaktır. Fakat birileri demokrasinin gerekli bir bileşeni olarak, oy

kullanma hakkının yeteneğe ya da çabaya bağlı olmaması gerektiğini tartışabilecektir. Bununla birlikte, sosyal pozisyonlar ve ekonomik mallar için rekabet söz konusu olduğunda, fırsat eşitliğini sağlamak sorunuyla ilgilenebiliriz, ama sonuç eşitliği konusunda daha az endişe duyarız. Diğer şeyler eşitken, başvuran en nitelikli kişinin işi almasını, en çalışkan/yetenekli kişinin diğerlerinden daha fazla ekonomik faydayı hak etmesini adil olarak değerlendiririz (aslında mevcut sistemlerin eşit şartlar oluşturmadığına, ahlaki olarak bazı seviyelerdeki gelirlerin yeniden dağıtılması gerektiğine inansak bile).

Farklı adalet alanlarına tabi olan farklı içerikler, kesin adalet doğrulama metotlarının uygunluğu konusuyla doğrudan ilgilidir. Fırsat eşitliği, "ekonomik adalet" alanında yer alacak modelleri uygulamak için uygun bir metrik olabilir; örneğin, iş mülakatları için adayların seçimi ya da sigorta hesaplamaları gibi. Bununla beraber medeni hukuk alanına dâhil olan bağlamlarda, eşitsizlik çıktıları (ya da "benzer etki") gibi daha duygusuz metrikleri dayatmak isteyebiliriz. Bu sosyal bloklaşma hissinin önemli olduğu, havaalanı güvenlik kontrolü vakaları için geçerli olabilir. Bu gruplar arasında temel oranlarda ciddi farklar olsa da tahmin sisteminin bir sonucu olarak hiçbir grup yoğun olarak irdelenmez (Hellman, 2008). Bu sebeple bir bağlamda uygun olan adalet metriğinin diğer bağlam için uygun olacağını varsayamayız.

3.2. Şans ve Fiil Oranlılığı

Eşitlikçi düşüncede ikinci bir ana tartışma bölümü, hangi eşitsizliklerin kabul edilebilir olduğunu belirlemede, seçim (Huseby, 2016) ve fiil oranlılığı (Temkin, 1986) gibi kavramların rolünü değerlendirmektedir. İnsanlar hangi koşullarda ve ne ölçüye kadar kendilerini, içinde bulundukları eşitsiz statü için sorumlu tutmalıdırlar? "Şans eşitlikçisi" diye adlandırılan bir grup bu soruyu, "ideal bir çözüm insanların özgür seçimleri ve bilgilendirildikleri riskleri almaları sonucu oluşan eşitsizliklere izin vermeli ama mantıksız bir şans ile ortaya çıkan sonuçları göz ardı etmelidir" önerisini getirerek yanıtlamayı amaçlar (Arneson, 1989). Özgür iradeli bireyler olarak, bazen bizi diğerlerinden daha iyi veya daha kötü hâle getiren seçimler yapma ve bu seçimlerin sonuçlarını katlanma kabiliyetimiz vardır. Yaptığımız seçimler belirli ödül ve cezaları hak ediyor olabilir. Bununla beraber, eşitlikçiler, bir bireyin kontrolü dışındaki koşulların sonucu olarak ortaya çıkan eşitsizliklerin düzeltilmesi için (örneğin, zayıflatıcı bir sağlık koşuluyla doğmak veya kişinin ten renginin sistemik olarak daha kötü muameleye yol açtığı bir kültürde doğmak), tartışmalarda bulunurlar. Buna ek olarak tercih edilen/edilmeyen ya da hak edilen/edilmeyen bu eşitsizlikler arasında ayrım yapan bir ilke tanımlamak, eşitlikçileri yüzyıllardır kızdıran aldatıcı bir görüş olmuştur.

Şans eşitlikçisinin yeniden dağıtımı sürdürme amacı, eşitsizliğin yalnızca safi şansa bağlı olduğu durumlarda ve eşitsizliğin kişisel seçimler gibi öncesinden bilgilendirme yapılan kumar oyunları gibi durumların gölgesinde kaldığı hâllerde ilginç sorular ortaya çıkarır. Bu ilginç sorular adil yapay öğrenme modellerine eklenmelidir. ABD'deki tekerrür

suçuna ilişkin risk skorlarının oluşturulması konusundaki üst düzey anlaşmazlıklar, özellikle COMPAS sistemi, öncelikli olarak Afrika kökenli Amerikalı ve Kafkasyalı süjeler üzerindeki farklı etkilere odaklanmıştır (Angwin, vd., 2016). Ama COMPAS skorlama sisteminin potansiyel olarak reddedilebilir özelliklerinden birisi "ırk" kavramını bir değişken olarak kullanması değildir (ki kullanmamıştır). Aksine COMPAS sistemi bireyin tercihinin sonucu olmayan değişkenlerin kullanımıdır, örneğin bir ailenin, bir sosyal çevrenin ya da suç oranlarının daha yüksek olduğu semtin bir parçası olmak. Bu örnekler ABD'de ki ırk durumu ile ilişkili oldukları için itiraz edilebilir örnekler olabilirler. Bununla beraber bu örnekler genel olarak kişisel seçimlerin sonucu olmadığı için daha geniş kapsamda da itiraz edilebilir örneklerdir. Bu nedenle, bu örneklerden kaynaklanan herhangi bir eşitsizlik, şans eşitliği görüşüne göre tolere edilebilmelidir.

Buna ek olarak, şans eşitlikçisi eleştirmenlerinin tartıştığı gibi, bazen eşitsizlikler seçim sonucu olsa bile yine de tazmin edilebilir olmalıdırlar. Örneğin, Elizabeth Anderson'ın öne sürdüğü üzere, standart şans eşitliği, bakıma muhtaç kişilerin bakıcılarının savunmasızlığına yol açar, çünkü standart şans eşitliği tam zamanlı gelir getiren bir işte çalışmak yerine, başkasına muhtaç olan kimselerle ilgilenmeyi seçen kişilerin durumlarını tazmin etmez (Anderson, 1999). Bu durumu Thayson ve Albertson (2017) "maliyetli kurtarış" olarak adlandırmaktadır. Onların görüşüne göre, şans eşitlikçiliği yalnızca avantaj ve dezavantaj yaratma sorumluluğu konusunda duyarlı olmalıdır –bu olanakları dağıtma konusunda sorumlu olmamalıdır. Böylece, gönüllü şekilde kendilerini eşitsiz pozisyonlara yerleştiren bireyler, bazı zamanlarda eğer tercihleri bazı önemli sosyal hedeflere hizmet ediyorsa, tazminat hak edebilirler. COMPAS vakasına dönecek olursak, bir kimsenin sosyal çevresi ya da yaşadığı semt o kişinin bireysel seçimi olsa bile (ekonomik olarak avantaj sağlayan seçimler için belki de daha muhtemel olan), bu tercihler yine de olumsuz sonuçlardan korunmayı hak ediyor olabilir. Bu koruma da Anderson, Thayson ve Albertson'un yukarı da özetlediği olaylardaki durumlarda uygulanabilir. Örneğin, bir kimse yaşadığı toplumda olumlu farklar yaratabilmek için, yüksek oranda suç işlenen bir semtte ikamet etmeyi seçebilir.

3.3. Deontik Adalet

Eşitlikçi siyaset felsefesi, eşitsizlik konusundaki belirli örnek olayları analiz etmek için uygulanırken, belirtilen soyut ilkelerin amprik kanıtlar ile desteklenmesi gerekir. Bu ampirik kanıtlar belirli olayların neden ve nasıl elde edildiğine ilişkindir. Bu durum, eşitlikçiliğin, Derek Parfit'in terminolojisi ile deontik olabileceği anlamını yansıtır. Bir başka deyişle bu durum ilişkilerin eşitsizliği ile ilgilenmekten ziyade, daha çok durumu ortaya çıkartan yolla ilgilenir (Parfit, 1997). Analitik felsefenin bittiği noktada, bazı grupların adil olmayan şekilde dezavantajlı gruplar olduğu özel durumları anlamak için sosyoloji, tarih, ekonomi ve diğer yeni oluşan disiplinlere ihtiyaç duyulur (Fang, Moro, 2010; Hooks, 1992). Ancak bundan sonra, sunulan eşitsizliğin adil olup olmadığını ve hangi ölçekte adaletsiz olduğunu anlamlı bir şekilde değerlendirebiliriz. Fakat tarihi ve

sosyolojik bağlamları göz önünde bulundurmak felsefi düşünceyi canlandırıp yeni sorular meydana getirebilecektir. Bu konudaki bir örnek sorumluluk niteliğidir. Eşitsizliklerin ilk ortaya çıkışından kimler sorumlu tutulmalıdır? Eşitsizliklerin doğrulanması için kimler sorumlu tutulmalıdır? Geçmişte bir kurum tarafından ortaya konan tarihi adaletsizliklerin giderilmesi yeniden talep edilebilir mi? Özel bir gruba yönelik eşitsiz muamelenin belirli bir örneği, o grup için daha geniş bir eşitsizlik bağlamında gerçekleşirse, yumuşak ya da avantajlı muamele modelleriyle kıyaslandığında, muamele daha mı kötü olacaktır?

Belirli vakalardan soyutlanmak, tarihi ve sosyal trendleri daha geniş çerçevede değerlendirmek, özellikle ilgi çekici olan belirli eşitsiz muamele formlarını neyin meydana getirdiğini daha iyi açıklamamızı sağlayabilecektir. İstatistiki bulgulara dayansa da ırksal profil oluşturmayı diğer profil formları oluşturma biçimlerinden daha kötü yapan şeyin ne olduğu tartışılırken, Kasper Lippert-Rasmussen şunları söyler (2014; s. 300):

> İstatistiki gerçekler genellikle nasıl davranmayı seçtiğimiz hakkındaki gerçeklerdir. Nasıl davranacağımız konusundaki seçimimizi ahlaki olarak değerlendirebildiğimiz için, poliçeleri değerlendirirken istatistiki bulguları doğru varsayamayız. Öncelikli bir soru olarak, bu istatistiki gerçeklerin meşru olarak değerlendirilip değerlendirilemeyeceğini sormamız gerekir.

Bu görüşe göre, ırksal profil oluşturmanın belirli yanlışları, ancak birinci sırada yer alan istatistiki düzenlemelere sebep olan sosyal süreçlere başvurularak anlaşılabilir. Lippert-Rasmussen'in örneğinde, ABD'de suç konusunda ırksal profil yaratmanın işe yaramasının sebebi (eğer gerçekten de işe yarıyorsa), beyaz çoğunluğun yaptıklarının ya da yapamadıklarının, benzer istatistiki çıkarımların güvenilirliğini azaltıyor olmasına bağlı olabilir.

Benzer şekilde bu gibi "deontik", tarihi ve sosyolojik çıkarımlar, belirli algoritmik karar verme içeriklerinde bulunan adaletin kararlaştırılması konusunda hayati olabilecek kritik altyapı bilgileri sağlayabilir. Bu çerçeveler dâhilinde modeller oluşturulurken, eşitsizliklerin tarihi nedenleri ve daha geniş anlamda var olan sosyal yapılar göz ardı edilemez. Basit bir seviyede bu, karakter seçiminin benzer bilgiler ile donatılması gerektiğini ifade eder. Bununla beraber eğer birden fazla adalet analizi mevcut ise, bu çerçeveler adalet analizi ve hafifletici unsurlar için hangi eşitsizliğin öncelik oluşturacağına karar verebilir. Daha geniş biçimde söylemek gerekirse, deontik çıkarımlar, farklı ve uyuşmaz adalet metrikleri içinde ortaya çıkan ahlaki gerilimlerin konumlandırılmasına ve aydınlanmasına yardımcı olabilirler. COMPAS tekerrür puanlama sistemi konusundaki tartışmaya dönecek olursak, sistemin eşit olmayan taban oranlarına sahip olmaması, bu sistemin doğruluk eşitliği konusunda, eşitlenmiş yanlış pozitif oranlarla birlikte elde edilmesinin matematiksel olarak imkânsız olduğu anlamına gelmektedir. Deontik eşitlikçi bakış açısı ise bu gibi eşit olmayan taban oranları için tarihi nedenlere odaklanmayı önerir. Bu kendi içinde, adalet metriği için uygulanmayı bekleyen sorunun doğrudan bir çözümlemesini yapmasa da konu hakkında verilecek cevapların,

uyuşmaz dilemmalar yaratan daha geniş sosyal durumlardan sorumlu çağdaş ve tarihi bileşenleri kapsaması gerektiğini tavsiye eder.

3.4. Dağıtıcı Zararlar Temsili Zararlar Ayrımı

Son olarak, eşitlikçi adalet anlayışının bazı yönlerinin doğrudan dağıtıcı olmadığını söylemek önemlidir. Bu bağlamda eşitlikçi adalet anlayışları belirli kararlar veren belirli insanlara kâr ve zararın dağıtımı konusuyla ilgilenirler. Bunlar daha çok farklı kimliklerin, kültürlerin, etnik kökenlerin, dillerin ya da sosyal kategorilerin temsiline ilişkin olabilir. Örneğin, birden çok resmi dilin kullanıldığı ülkeler her bir dilin temsilini garanti etmek gibi bir yükümlülüğe sahip olabilirler. Bu yükümlülüğün bu dil gruplarının üyelerine yönelik eşit olmayan yararlar ve zararlarla ilgili herhangi özel bir iddiadan kaynaklanması da gerekmemektedir (Taylor, 1994). Benzer tartışmalar resmî belgelerdeki kültürel temsiller hakkında da yapılabilir. Hatta kamuya ait parayı kullanan kuruluşların yayın politikaları için de bu tartışma yürütülebilir. Özel kurumlar bile gönüllü olarak benzer görevleri kendi üzerlerine dayatabilirler. Eşit kültürel tanınma ve dağıtımcı eşitlikçiliğin ne ölçüde farklı kavramlar olduğu konusunda bir tartışma da vardır. Bazıları tanınma ve dağıtmanın, tatmin edici adalet teorisine indirgenemeyecek iki unsur olduğunu söylerken, diğerleri zenginlik ya da kaynakların yeniden dağıtımına ilişkin bir anlaşmazlığın, belirli bir sosyal grubun ya da bireyin özelliklerine ilişkin sosyal değer biçme unsuruna indirgenebileceğini iddia eder (Fraser, Honneth, 2003).

Temsili eşitlik konusundaki bu gibi düşünceler algoritmik yanlılık üzerine en dikkat çekici, tartışmalı örnekleri yakalamaktadır. Örneğin kelime gömme üretiminde kullanılan cinsiyet ve dil külliyatının diğer yanlılıkları üzerine rapor edilmiş birçok çalışma temsili adaletin birer örneğidir (Bolukbasi, vd., 2016; Caliskan-Islam, vd., 2016). Bu gibi vakalarda zorunlu olmaksızın, problem herhangi bir sosyal grubun bir üyesine ilişkin özel bir zarar problemi değildir. Daha ziyade bu, doğal dil sınıflandırıcıları veya arama motoru sonuçları gibi dijital kültür yapılarında temsil edilen belirli gruplar için kullanılan yöntemin problemidir. Bu noktada, farklı grup etkileri ve benzer insanların benzer davranışı gibi düşünceler uygun olmadığı için adalet ve yanlılık konusunda farklı bakış açıları gerekebilir. Bunun yerine amaç, farklı grupların herhangi bir sıralamadaki eşit temsilini garanti etmek (Zehlike, vd., 2017) veya gruplardaki normların yaptırımını ayarlayan farklı normatif/ideolojik bakışlara gereken önemi vermek olabilir (Binns, vd. 2017).

4. Sonuç

Yapay öğrenmeye ilişkin mevcut bakış açıları veri hazırlama, model öğrenimi ve ileri işlem aşamaları konusunda yapılan müdahalelere odaklanmaktadır. Süreçleri yürütmeyi hedefleyen veri bilimcilerinin tipik görev alanları göz önüne alındığında bu anlaşılır bir durumdur. Ancak burada bakış açısından kaynaklanan bir tehlike vardır. Bu bakış açısı dar anlamda hukuktan türetilmiş ve bağlamdan yoksun olan, önceden belirlenmiş korumalı sınıfların statik setlerine odaklanır. Bu sınıfların neden korunduğunu ve onların nasıl özel

bazı yargı bakış açılarına ilişkin uygulamalar olduğunu göz ardı eder. Ayrımcılığa ve adalete ilişkin felsefi düşünceler ise bu daha temel sorular üzerinde düşünmeye teşvik eder ve daha ileri değerlendirmeler için neyin ilişkili olabileceğine ve neden ilişkili olabileceğine dair takip edilecek yollar önerir.

Bu durum pratikte yer alabilecek etkili ve sistematik adil yapay öğrenme yaklaşımlarını kısıtlayabilecek bir dizi pratik zorluklar yaratabilir. Özel yapay öğrenme görevleri bağlamında, bu gibi eşitlikçi teoriler arasındaki farkları tercüme etme ve açığa çıkarma girişimleri muhtemelen çetrefilli olacaktır. Basit vakalarda, modelleri eğitmek için kullanılan vektör özellikleri sezgisel olarak hem seçilmiş hem de seçilmemiş olarak sınıflandırılabilen kişilik özelliklerini içerebilir (ve bu nedenle farklı muamele için meşru veya gayri meşru gerekçeler, örneğin şans eşitliği). Ancak daha sık olarak, adalet konusunda ilişkili felsefi meselelerin gerekliliğini doğru şekilde yakalamış uygun yaklaşımlar, situ verisinde tipik olarak mevcut olmayan faktörlere dayanabilir. Bu gibi kayıp veriler etki altındaki bireylerin korunmuş özelliklerini barındırabilir (Veale, Binns, 2017) ama bununla beraber kişisel sorumluluğun, kişi kusurunun ya da kişinin fiili karşılığındaki hak ettiği cezanın ölçümüne ilişkin bilgileri de içerebilir –örneğin kişilerin sosyo-ekonomik durumları, yaşam deneyimleri, kişisel gelişimleri ve bu kişiler arasındaki ilişkiler gibi. Yalnızca yasal koruma altındaki kategorilerin eğitim verileri ve listelerine dayanarak bu tip sonuçlar çıkarmak için yapılan girişimlerin, kendine özgü yaşamlar ve farklılaşan sosyal bağlamlarda ortaya çıkan adalet sorularını yargılamak için bir yol oluşturması muhtemel değildir.

References

Elizabeth S Anderson. What is the point of equality? *Ethics*, 109(2):287–337, 1999.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *Pro Publica*, 2016.

Richard J Arneson. Equality and equal opportunity for welfare. *Philosophical studies*, 56(1): 77–93, 1989.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International Conference on Social Informatics*, pages 405–415. Springer, 2017.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

Aylin Caliskan-Islam, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. *arXiv preprint arXiv:1608.07187*, 2016.

Gerald A Cohen. On the currency of egalitarian justice. *Ethics*, 99(4):906–944, 1989.

William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, 2016.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

Ronald Dworkin. What is equality? part 1: Equality of welfare. *Philosophy & public affairs*, pages 185–246, 1981.

Benjamin Eidelson. *Discrimination and Disrespect*. Oxford University Press, 2015.

Hanming Fang and Andrea Moro. Theories of statistical discrimination and affirmative action: A survey. Technical report, National Bureau of Economic Research, 2010.

Nancy Fraser and Axel Honneth. *Redistribution or recognition?: a political-philosophical exchange*. Verso, 2003.

Margaret Gilbert. Collective epistemology. *Episteme*, 1(2):95–107, 2004.

Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25 (7):1445–1459, 2013.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

Deborah Hellman. *When is discrimination wrong?* Harvard University Press, 2008.

Elisa Holmes. Anti-discrimination rights without equality. *The Modern Law Review*, 68(2):175–194, 2005.

Bell Hooks. Yearning: Race, gender, and cultural politics. 1992.

Robert Huseby. Can luck egalitarianism justify the fact that some are worse off than others? *Journal of Applied Philosophy*, 33(3):259–269, 2016.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, 2016.

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Techniques for discrimination-free predictive models. In *Discrimination and Privacy in the Information Society*, pages 223–239. Springer, 2013.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

Annabelle Lever. Racial profiling and the political philosophy of race. *The Oxford Handbook of Philosophy and Race*, page 425, 2016.

Kasper Lippert-Rasmussen. *Born free and equal?: a philosophical inquiry into the nature of discrimination*. Oxford University Press, 2014.

Derek Parfit. Equality and priority. *Ratio*, 10(3): 202–221, 1997.

Philip Pettit. Responsibility incorporated. *Ethics*, 117(2):171–201, 2007.

Edmund S Phelps. The statistical theory of racism and sexism. *The american economic review*, 62(4):659–661, 1972.

John Rawls. *A theory of justice*. Harvard university press, 2009.

Thomas M Scanlon. *Moral dimensions*. Harvard University Press, 2009.

Frederick F Schauer. *Profiles, probabilities, and stereotypes*. Harvard University Press, 2009.

Shlomi Segall. What's so bad about discrimination? *Utilitas*, 24(1):82–100, 2012.

Amartya Sen. *Inequality reexamined*. Clarendon Press, 1992.

Charles Taylor. *Multiculturalism*. Princeton University Press, 1994.

Larry S Temkin. Inequality. *Philosophy & Public Affairs*, pages 99–121, 1986.

Jens Damgaard Thaysen and Andreas Albertsen. When bad things happen to good people: luck egalitarianism and costly rescues. *Politics, Philosophy & Economics*, 16(1):93–112, 2017.

Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

Michael Walzer. *Spheres of justice: A defense of pluralism and equality*. Basic books, 2008.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa* ir: A fair top-k ranking algorithm. *arXiv preprint arXiv:1706.06368*, 2017.