Mühendislik Bilimleri ve Tasarım Dergisi
6(1), 93-107, 2018
e-ISSN: 1308-6693

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.394527

Araştırma Makalesi Research Article

GÜVENLİ YAPAY ZEKÂ SİSTEMLERİ İÇİN İNSAN DENETİMLİ BİR MODEL GELİŞTİRİLMESİ

Utku KÖSE*

Süleyman Demirel Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Isparta, Türkiye

Anahtar Kelimeler

Yapay Zekâ,
Yapay Zekâ Güvenliği,
Etmen Tabanlı Sistemler,
Makine Öğrenmesi,
Yapay Zekâ'nın Geleceği.

Öz

Yapay Zekâ, gerek günümüz, gerekse geleceğin en etkin araştırma alanlarından birisi olarak bilinmektedir. Ancak Yapay Zekâ'nın hızlı yükselişi ve otonom bir şekilde bütün gerçek dünya problemlerini çözebilir potansiyele sahip olması, çeşitli endişeleri de beraberinde getirmiştir. Bazı bilim insanları, zeki sistemlerin ilerleyen süreçte insanlığı tehdit edebilecek düzeye gelebileceğini ve bu nedenle çeşitli önlemlerin alınması gerektiğini düşünmektedir. Bu nedenle Makine Etiği ya da Yapay Zekâ Güvenliği gibi birçok alt-araştırma alanı da zaman içerisinde ortaya çıkmıştır. Açıklamalar bağlamında bu çalışmanın amacı da, insan denetimini de içeren, zeki etmen ve Makine Öğrenmesi odaklı önlemleri bünyesinde barındıran, güvenli bir zeki sistem modeli önermektir. Çalışmada Yapay Zekâ Güvenliği odaklı temel konularla birlikte önerilen modelin detaylarına ilişkin açıklamalar sunulmuş ve potansiyeli hakkında değerlendirmeler yapılmıştır. Modelin geleceğin güvenli Yapay Zekâ sistemlerine ilham kaynağı olabileceği düşünülmektedir.

DEVELOPING A HUMAN CONTROLLED MODEL FOR SAFE ARTIFICIAL INTELLIGENCE SYSTEMS

Keywords

Artificial Intelligence,
Artificial Intelligence Safety,
Agent Based Systems,
Machine Learning,
Future of Artificial Intelligence.

Abstract

Artificial Intelligence is known as one of the most effective research field of nowadays and the future. But rapid rise of Artificial Intelligence and its potential to solve all real-world problems autonomously, it has caused also several anxieties. Some scientists think that intelligent systems can reach to a level, which is dangerous for the humankind so because of that some precautions should be taken. So, many sub-research fields like Machine Ethics or Artificial Intelligence Safety have appeared in time. In the context of the explanations so far, objective of this study is to suggest a secure intelligent model including human control and also precautions based on intelligent agent and Machine Learning. In the study, essential subjects regarding Artificial Intelligence Safety and details of the suggested model have been provided and also some evaluations about its potential have been done. It is though that this model can be an inspiration for safe Artificial Intelligence systems of the future.

Alıntı / Cite

Köse, U., (2018). Güvenli Yapay Zekâ Tabanlı Sistemler İçin İnsan Denetimli Bir Model Geliştirilmesi, *Journal of Engineering Sciences and Design*, 6(1),93-107

Yazar Kimliği / Author ID (ORCID Number)

U. Köse, 0000-0002-9652-6415

Başvuru Tarihi /Submission Date	13.02.2018
Revizyon Tarihi / Revision Date	07.03.2018
Kabul Tarihi / Accepted Date	27.03.2018
Yayım Tarihi / Published Date	28.03.2018

* İlgili yazar / Corresponding author: utkukose@sdu.edu.tr, +90-246-211-1391

1. Giriş

Bilgisayar Bilimleri kapsamında günümüz en güçlü araştırma alanlarından birisi Yapay Zekâ olarak bilinmektedir. Temeli insan ve doğal dinamiklerin benzetimine dayanan (Karaboğa, 2014; Nabiyev, 2005; Russell vd., 2003) Yapay Zekâ, gerçek dünya problemlerinin etkin ve verimli çözülmesi konusunda birçok başarıya imza atmıştır. Özellikle Makine Öğrenmesi bünyesinde, "tıpkı insanlardaki öğrenme sürecine benzer şekilde", öğrenerek çalışan teknikler (Alpaydın, 2014; Nabiyev, 2005), çözülmesi o ana dek imkânsız görülen problemlerin bile çözülmesine zemin hazırlamıştır. Yapay Zekâ sayesinde birçok alandaki geleneksel yaklaşım, yöntem ve teknikler yerlerini 'zeki' çözüm süreçlerine bırakmıştır. Tahmin, optimizasyon, kontrol, yorumlama, bilgi işleme ve modelleme gibi (Copeland, 1993; Nabiyev, 2005) birçok farklı çözüm yaklaşımını kullanan, esnek yapıdaki Yapay Zekâ teknikleri, zeki sistemlerin zaman içerisinde bilimsel çalışmalardan sıyrılarak günlük yaşamımızda yer edinmesine de olanak sağlamıştır. Artık Yapay Zekâ, insanlığın geleceğini şekillendiren, en önemli araştırma ve ilgi alanlarından birisi olarak da yaygın kabul görmektedir.

Tıpkı her hızlı teknolojik gelişmede olduğu gibi, Yapay Zekâ'nın bu durdurulamaz ve hızlı yükselişi de, teknolojik değişimlere ayak uydurmada sık sık sorunlar yaşayan insanoğlunu çeşitli endişelere ve sorunlara da sevk etmiştir. Yapay Zekâ'nın sadece kendilerine sunulan örnekler üzerinden öğrenerek kendisini geliştirebilmesi potansiyeli ve bu yönde elde edilen başarılar, zeki sistemlerin zamanla insan denetimine aykırı davranmaya başlayıp başlamayacağı yönünde endişelerin oluşmasına sebep olmuştur (Anderson, & Anderson, 2011; Kose, & Pavaloiu, 2017; Yampolskiy, 2013). Söz konusu aykırı davranışlar, teorik anlamda bakıldığında insan üzerinde bir düzen oluşturma yönünde olabileceği gibi, insanlara tehlike oluşturacak olaylara neden olacak yanlış hareketler neticesinde de olabilecektir. Bu nedenle, bilimsel literatürde güvenli zeki sistemlerin nasıl geliştirilebileceğine odaklanan bir alt-araştırma alanı: Yapay Zekâ Güvenliği (Artificial Intelligence Safety) ortaya çıkmıştır.

Yapay Zekâ Güvenliği alanı kapsamındaki gelişmeler, özellikle son yıllarda hız kazanmıştır (Cath vd., 2017; Hussain, 2018; Russell vd., 2003; Wu vd., 2017). Bu noktada, her ne kadar zeki sistemlerin insanlara karşı güvenli olup olmadığı sorgulanır olsa da, ilgili sistemlerin dünya üzerindeki diğer canlılara ve hatta diğer zeki sistemlere karşı olan davranış ve yaklaşımlarının da güvenlik odaklı ele alınması gerekmektedir. Sonrasında ne tür davranışlar geliştirebileceği belli olmayan bir Yapay Zekâ'ya karşı çeşitli önlemler tasarlanabileceği gibi, bu tür sorunlara yol açacak faktörlerin neler olabileceği ve nasıl bertaraf edilebileceği gibi sorunlar da yine Yapay Zekâ Güvenliği kapsamında dikkate alınabilmektedir. Yapay Zekâ Güvenliği, yine zeki sistemlerin etik davranışlar ve problemler açısından incelendiği Makine Etiği (Machine Ethics) alt-alanı (Anderson, & Anderson, 2007; Anderson, & Anderson, 2011) ile de yakından ilişkilidir. Konu detaylı irdelendiğinde, zeki sistemlerin eğitiminde kullanılan verilerin kalitesinden, bu tür sistemleri tasarlayacak insan faktörüne kadar birçok farklı unsur değerlendirmeye alınabilmektedir. Yine insanlar arasındaki kültürel farklılıklar, insanları ilgilendiren ve zeki sistemlerin ellerine teslim edilecek problemler ve zeki sistemleri çevreleyen birçok farklı çevresel faktör bu kapsamda dikkatli bir şekilde analiz edilmelidir. Buradan tahmin edileceği üzere, Yapay Zekâ alanının geleceği ile ilişkili güvenlik ve etik odaklı konular, esasında çok disiplinli çalışmaları ve etkileşimleri gerektirmektedir.

Açıklamalar kapsamında bu çalışmanın amacı, insan denetimini de içeren, zeki etmen ve Makine Öğrenmesi odaklı önlemleri bünyesinde barındıran, güvenli bir zeki sistem modeli önermektir. Güvenli bir zeki sistemin nasıl geliştirilebileceği ve bunun muhtemel bütün olasılıkları kapsayacak şekilde nasıl yapılabileceği sorunsalı, Yapay Zekâ Güvenliği'nin odak noktasında yer almakla birlikte, bu çalışmadaki model üzerinden alternatif bir yaklaşım izlenmiştir. Yine çalışmada Yapay Zekâ Güvenliği odaklı temel konulara yönelik bir farkındalık düzeyi oluşturulması ve geçerli yaklaşım, yöntem ve teknikler bağlamında nasıl bir model oluşturulduğunu anlatmak adına; modelin mimarisi ile ilgili çeşitli detayların açıklanması da temel amaç çerçevesinde takip edilen diğer konular olmuştur. Ayrıca, model ile ilgili bilgiler sunulduktan sonra modelin potansiyeli hakkında değerlendirmeler yapılması da oldukça önemlidir.

İlgili amaç ve konu kapsamı dikkate alınmak suretiyle, çalışmanın ilerleyen bölümleri şu şekilde organize edilmiştir: Bir sonraki bölüm altında, modelin temelini oluşturan Yapay Zekâ Güvenliği ile ilgili temel konular açıklanmış, bu konuda okuyucuların temel düzeyde bir farkındalığa ulaşması yönünde bir yol izlenmiştir. İkinci bölümü takiben, üçüncü bölüm altında ise, önerilen güvenli zeki sistem modelinin detaylarına değinilmiş, bu noktada, hangi mekanizmanın hangi güvenlik unsurunu sağlamak adına düşünüldüğü konusunda bilgiler verilmiştir. Açıklamalar akabinde, dördüncü bölümde ise, modelin potansiyeline ilişkin çeşitli değerlendirmeler yapılmıştır. Çalışma, son bölüm altındaki sonuçlar ve gelecek çalışmalar dair açıklamalarla sona erdirilmiştir.

2. Yapay Zekâ Güvenliği ve İlgili Araştırmalar

Yapay Zeka Güvenliği ile ilgili literatür, Yapay Zekâ'nın genel literatürü dikkate alındığında, yeni sayılabilecek bir yapıdadır. Bunun nedeni, gelişen teknolojiyle beraber Yapay Zekâ'nın günlük yaşamımızda çok daha fazla yer edinmesi ve dolayısıyla endişe doğuracak senaryoların daha fazla tartışılır hale gelmesidir. Bilim dünyası, Yapay Zekâ'nın gerçekten geleceğe dönük bir

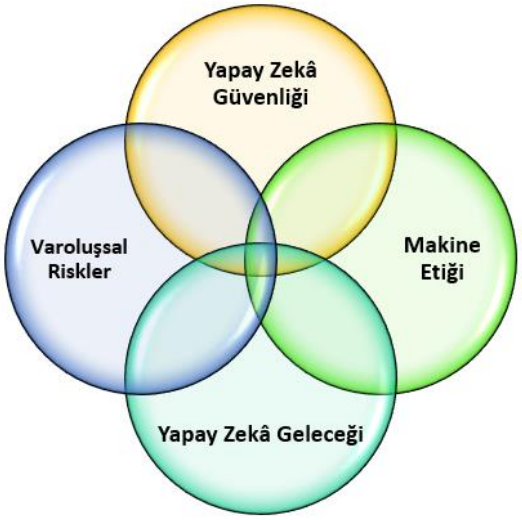
tehdit olup olmayacağı konusunda zıt iki görüşe çoktan ayrılmış durumdadır. En basitinden, Yapay Zekâ'nın insanlığa tehdit olabileceğini düşünen, PayPal, Tesla, SpaceX ve Boring Company gibi gündemdeki teknoloji şirketlerinin başındaki isim Elon Musk ile Yapay Zekâ tabanlı bir geleceği daha iyimser açıdan değerlendiren, Facebook adlı sosyal medya ortamının kurucusu Mark Zuckerberg birbirlerine zıt görüşler ifade etmişlerdir (Kulaklı, 2017). Yine Stephen Hawking gibi öncü çoğu bilim insanı da Yapay Zekâ konusunda kötümser düşüncelere sahiptir (Cellan-Jones, 2014).

2.1. Araştırma Merkezleri

Söz konusu tartışmaların odağında, Oxford Üniversitesi, UC Berkeley ve Cambridge Üniversitesi gibi önde gelen çeşitli üniversitelerde, Yapay Zekâ Güvenliği, Etiği ve Geleceği odaklı araştırma merkezleri kurulmuş, yine çeşitli özel araştırma enstitüleri de Yapay Zekâ'yı söz konusu açılardan ciddi anlamda incelemeye başlamıştır. Bu bağlamda, sayısı her geçen gün artan bu merkezlerden göze çarpanlarını kısaca şöyle listeleyebiliriz:

- Future of Humanity Institute (İnsanlığın Geleceği Enstitüsü – Oxford Üniversitesi),
- Center for Human-Compatible Artificial Intelligence (İnsan Uyumlu Yapay Zekâ Merkezi – UC Berkeley),
- Leverhulme Centre for the Future of Intelligence (Leverhulme Zekânın Geleceği Merkezi – Cambridge Üniversitesi),
- Centre for the Study of Existential Risk (Varoluşsal Risk Çalışmaları Merkezi – Cambridge Üniversitesi),
- Next Generation Artificial Intelligence Research Center (Gelecek Nesil Yapay Zekâ Araştırma Merkezi – Tokyo Üniversitesi),
- Machine Intelligence Research Institute (MIRI – Makine Zekâsı Araştırma Enstitüsü),
- Open AI (Açık Yapay Zekâ),
- Future of Life Institute (Yaşamın Geleceği Enstitüsü),
- Vector Institute for Artificial Intelligence (Yapay Zekâ Vektör Enstitüsü),
- Global Catastrophic Risk Institute (Küresel Felaket Riskleri Enstitüsü)

Anlaşılacağı üzere, Centre for the Study of Existential Risk (Varoluşsal Risk Çalışmaları Merkezi – Cambridge Üniversitesi) ve Global Catastrophic Risk Institute (Küresel Felaket Riskleri Enstitüsü) gibi araştırma merkezleri, çalışma kapsamlarını daha geniş çapta tutmakla birlikte, Yapay Zekâ'yı da, sahip olduğu tehlike potansiyeli nedeniyle çalışma konuları arasına almaktadır. Bu bağlamda özellikle Existential Risk (Varoluşsal Risk) konusu (Bostrom, 2002), kendi içerisinde çok-disiplinli ve insan odaklı birçok araştırmayı barındıran ve insan varoluşuna tehdit oluşturacak unsurları inceleyen önemli araştırma alanlarından birisi olarak dikkat çekmektedir.



Şekil 1. Yapay Zekâ Güvenliği ve ilişkili bazı araştırma konuları.

2.2. Tartışmaya Açık Senaryolar

Yapay Zekâ Güvenliği alanındaki çalışmalarının önemini anlamak adına, literatürü sıklıkla meşgul eden ve hem günümüz, hem de geleceğin Yapay Zekâ teknolojilerinin akıbetini yakından ilgilendiren çeşitli senaryolardan bahsetmekte fayda bulunmaktadır. Bu bağlamda, Yapay Zekâ Güvenliği ile birlikte bu alanla ilişkili diğer araştırma konuları (Örneğin; Makine Etiği, Yapay Zekâ Geleceği) kapsamına da giren, tartışmaya açık başlıca senaryoları şöyle açıklayabiliriz:

- **Ahlaki Çelişkiler:** Her ne kadar Yapay Zekâ Güvenliği'nden ziyade Makine Etiği konusu içerisinde inceleniyor olsa da, etik olmayan zeki sistem davranışları da bir tür güvenlik zafiyetine karşılık geldiğinden dolayı, ahlaki çelişkiler de Yapay Zekâ Güvenliği içerisinde incelenen senaryoları karşımıza çıkarmaktadır. En basitinden Massachusetts Teknoloji Enstitüsü'nün (MIT) 'Moral Machine' (Awad vd., 2018; Massachusetts Teknoloji Enstitüsü, 2018) çalışmasıyla akıllarda yer edinen ve sonucu ölümcül bir kazada, sürüsüz bir (self-driving) taşıtın hangi bireylerin – canlıların yaşamaya devam edeceği, hangilerinin öleceği konusunda çelişkide kalması ve bu noktada en mantıklı kararın nasıl verilebileceği sorusu, en dikkat çeken ahlaki çelişki senaryolarından birisidir (The Associated Press, 2017). Alternatif senaryolar üretmek gerekirse; tıbbi müdahaleye aynı anda acil ihtiyaç duyan çok sayıda hastadan hangisine müdahale etmesi gerektiğine karar vermesi gereken Yapay Zekâ tabanlı bir doktor robot, ahlaki bir çelişki içerisine girecektir. Yine suç işleyerek yoksullara yardım eden bir kişiyi yargılaması gereken Yapay Zekâ tabanlı bir hâkim robotun içerisinde bulunduğu durum da ahlaki çelişkiler barındırmaktadır. Zaten insanların içerisinden çıkamadığı ahlaki çelişkilerin yine insan ürünü

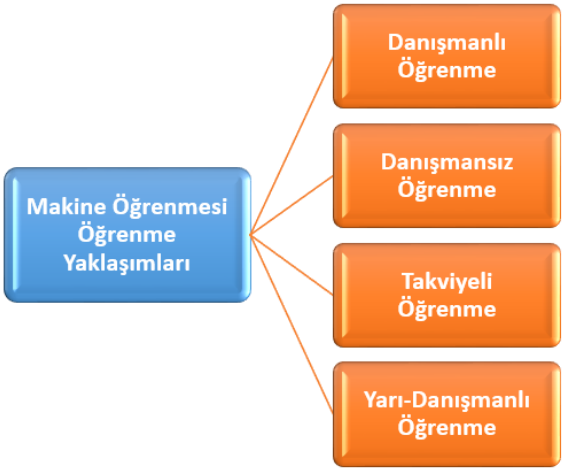
ve henüz insan düşünce ve davranış şeklinde olduğu kadar bir karmaşıklık düzeyinde olmayan Yapay Zekâ tarafından nasıl çözümleneceği, önemli bir sorundur.

Şekil 2. Yapay Zekâ'yı ilgilendiren ahlaki çelişkilere bazı örnekler.

- **İşsizlik ve Bağlı Sorunlar:** Son yıllarda sıklıkla tartışılagelen önemli kötümser senaryolardan birisi de, zeki sistemlerin (robotların) insanların elinden zamanla işlerini devralmasıdır. Yakın zamanlarda bir avukatlık firmasında Yapay Zekâ'nın çalıştırılmaya başlanması, tıbbi destek adına doktorlar – sağlıkçılar yerine Yapay Zekâ tabanlı sistemlerden destek alınır hale gelinmesi ve hatta marketlerde bile robotların – zeki sistemlerin işbaşı yapması (Brady, 2017; Cellan-Jones, 2017; Norman, 2018; The Week, 2018) durumun ciddiyetini ortaya koymuştur. Bu durum, çeşitli mesleklerin teknolojik gelişmeler nedeniyle ortadan kalmasına benziyor olsa da, Yapay Zekâ'nın ve destekleyici teknolojilerin hızlı yükselişinin, telafi edilemeyen zamansız işsizliklere ve sonucunda insanlarda ekonomik ve sosyal anlamda çeşitli sıkıntılara yol açacağı yönünde endişeler gittikçe yaygınlaşmaktadır (Barnett, 2017; Heath, 2015; Singh, 2018).
- **Bilincin Modellenmesi:** Bilincin modellenmesi, hem Yapay Zekâ'nın en üst düzeyde gerçekleştirilmesi, hem de etik ve güvenlik odaklı sorunların etkin bir şekilde çözümlenmesi adına, halen çözüme ihtiyaç duyan konular arasındadır. Bu noktada, insandaki gibi bir bilincin zeki sistemlerde nasıl modellenebileceği, modellenirse bunun bizler ve dünya için iyi mi yoksa kötü mü olacağı, bilinç sahibi bir zeki sistemin haklarının neler olacağı gibi sorular bu bağlamda sıklıkla zihinleri bulandırmaktadır (Holland, 2003; Minsky, 2007).
- **Sorumluluk ve Sahiplik – Telif Hakları:** Bilincin modellenmesiyle de ortaya çıkabilecek diğer alternatif senaryolar da sorumluk ve sahiplik – telif hakları üzerinedir. Buna göre, zeki sistemlerin gelişmesi ve yaşamımızda daha fazla yer edinmesi durumunda, bu sistemlerin hangi temel sorumluluklara sahip olacağı, bunların tamamen kendilerine mi yoksa geliştiricilerine mi ait olacağı (Ashrafian, 2015; Dashevsky, 2017; Dormehl, 2017), zeki sistemlerin üreteceği (özellikle sanatsal – edebi ürünlerde) ürünlerde, telif hakları konusunun ne şekilde çözümleneceği (Ashrafian, 2015; Davis, 2018) gibi sorular hemen akla gelmektedir.
- **'Kırmızı Düğme' Sorunu:** Literatürde tartışılagelen önemli sorunlardan birisi de, her şeye rağmen tehlikeli ve zarar verici davranışlarda bulunmaya başlayan bir zeki sistemi durduracak 'kırmızı bir düğmenin' nasıl gerçekleştirilebileceğidir. İlerleyen paragraflar altında da ifade edildiği üzere, etmen tabanlı yaklaşımlarla çözümlenmeye çalışılan bu sorun, tehlikeli ve yeterince gelişmiş bir zeki sistemin, kendisini durduracak önlemleri de bertaraf edecek algoritmik önlemleri alacağından dolayı, içerisinden nasıl çıkılacağı merak konusu olan bir paradoksu da oluşturmaktadır (Orseau, & Armstrong, 2016; Shead, 2016).
- **Süper-zekâ:** Nick Bostrom (2014) tarafından literatüre kazandırılıp, birçok bilim insanı tarafından da üzerinde araştırmalar yapılan bir kavram olan Süper-zekâ (Superintelligence), en üstün zekâlı insanın sahip olduğu doğal zekâ düzeyinden de üstün olan Yapay Zekâ'yı tanımlamaktadır (Schneider, 2016; Yampolskiy, 2015). Dolayısıyla teknolojik gelişmeler beraberinde ortaya çıkacağı düşünülen bu tür zeki sistemler de, insanlığı alt edecek ya da doğrudan olmasa bile dolaylı yönden bazı evrensel dinamikleri tehdit edecek oluşumlar olarak endişe kaynağı haline gelmektedir. Bu durumda etik ve güvenlik odaklı senaryolar, distopik bir geleceğe karşılık gelen, daha üst düzey senaryolara evrilmektedir (Bostrom, 2014).
- **Teknolojik Tekillik Hipotezi:** Teknolojik Tekillik (Technological Singularity), bir hipotez olarak literatüre kazandırılan ve geleceğin, Yapay Zekâ ve yüksek teknoloji unsurları tarafından şekillendiği, dolayısıyla toplumsal yapının ve insanların bile radikal düzeylerde değişime uğradığı bir geleceği tasvir etmektedir (Kurzweil, 2005; Muehlhauser, & Helm, 2012). Bu tasvir ütopik ya da distopik bir geleceği işaret etmediğinden dolayı, bilim dünyası yine ikiye ayrılmaktadır. İnsanlardan çok teknolojinin ve daha spesifik anlamda Yapay

Zekâ'nın yönlendireceği bir gelecek bu nedenle Teknolojik Tekillik konusunun Yapay Zekâ Güvenliği kapsamında tartışılagelmesine sebep olmuştur.

- **Makinelerin Makineleri Yaratması:** Tıpkı distopik Yapay Zekâ temalı bilim-kurgu eserlerde (filmler, romanlar) olduğu gibi, kendi kendini inşa edebilen Yapay Zekâ kuşkusuz ki yakın gelecek için bile oldukça mümkün görülen bir senaryodur. Hâlihazırda yazılımsal düzeyde kendi zeki sistemlerini oluşturan başka zeki sistemler geliştirilmiş olsa da (Galeon, & Houser, 2017; Metz, 2017), bunun donanımsal düzeye genişlemesi de kaçınılmaz bir sonuç olacaktır. Dolayısıyla, Yapay Zekâ'nın kendi istek ve görevlerine uygun başka Yapay Zekâ tabanlı sistemler yaratabilmesi, yoğun tartışmaya sebep olan bir gelişme olarak görülmektedir. Yapay Zekâ Güvenliği söz konusu olduğunda, zeki sistemlere böyle bir iznin verilip verilmeyeceği ya da bir zeki sistemin bizlerden izin almaksızın böyle bir şeye kalkışmasının ne kadar güvenli olduğu gibi sorular sıklıkla akla gelmektedir.
- **Öğrenme Yaklaşımları ve Sorunlar:** Yapay Zekâ'nın öğrenmeye dayalı alt-kolu Makine Öğrenmesi kapsamında çeşitli öğrenme yaklaşımları [Danışmanlı (Supervised), Danışmansız (Unsupervised), Takviyeli (Reinforcement) ve Yarı-Danışmanlı (Semi-Supervised)] bulunmaktadır (Alpaydin, 2014; Hady, & Schwenker, 2013; Kober, & Peters, 2012; Kotsiantis, 2007; Silva, & Zhao, 2016). Bu öğrenme süreçleri, öğrenme verilerinden ya da anlık ortam – unsur dönütlerinden hareketle meydana gelmektedir. Bu noktada, zeki bir sistemin 'iyi' bir şekilde eğitilmesi (öğrenmesi) için gerekli olan bilginin nasıl anlaşılacağı, bilginin değeri, bilginin kontrol edilebilirliği, bilginin yeterli olup olmadığı gibi birçok soru, öğrenme yaklaşımlarıyla ilişkilendirilmektedir. Dolayısıyla, Yapay Zekâ öğrenme süreci yüzünden zamanla güvensiz hale gelen zeki sistemler, hata yaparak insana zarar veren zeki sistemler ve zeki sistemlerin hiçbir zaman güven veremeyeceği gibi kötümser senaryoların oluşmasına sebep olmaktadır.
- **Kötü Amaçlı Zeki Sistemler:** Nasıl ki tarihsel süreç, insanlığı birkaç adım öteye taşıyacak teknolojik gelişmelerin, iyi olduğu kadar kötü amaçlı uygulamalarına da sahne olduysa, Yapay Zekâ'nın kötü amaçla kullanılacak şekilde, kasıtlı kullanımı da kaçınılmaz olabilecektir. Burada akla gelen senaryo, zeki sistemleri farklı amaçlar için hackleyebilen başka zeki sistemlerin oluşturulmasıdır. Burada Yapay Zekâ Güvenliği, zeki sistemlerin güvenli oluşturulması kadar, korunması yönünde çalışmaları da içermelidir.
- **İnsan Faktörünün Önemi:** Yapay Zekâ'yı ortaya koyması adına insan faktörünün rolü şüphe götürmez bir gerçek olmakla birlikte, madalyonun diğer yüzünde insan faktörünün beraberinde insan hatalarını da getirebilmesi durumu bulunmaktadır. Bu noktada, zeki sistemlerin, hem uygulama aşamasında hem de eğitim (öğrenme) aşamasında insan hatalarından etkilenebilmesi ve esnek yapısının buna izin vermesi, bir zeki sistemin aslında ufak bir hata yüzünden nasıl 'kötü' bir sistem haline gelebileceğine dair senaryoların da türemesine sebep olabilmektedir. Belki de insanlığın bir problemi çözmek için zeki sisteme aşlamaya çalıştığı yollar bütünü, esasında en doğru çözüm olmamakla birlikte, sistemde problemlere yol açabilecek oluşumları da tetikleyebilecektir. Bu sebeple, insan faktörü de Yapay Zekâ Güvenliği kapsamına giren bir diğer tartışmaya açık senaryolar unsuru olarak göze çarpmaktadır.



Şekil 3. Yapay Zekâ – Makine Öğrenmesi öğrenme yaklaşımları.

Literatürde yapılmış olan Yapay Zekâ Güvenliği odaklı çalışmalara bakıldığında, ifade edilen senaryolardan güç alan, farklı yaklaşımlar olduğunu görebiliriz. Bunlardan dikkat çekenleri, ilerleyen alt-başlıklar altında açıklanmıştır.

Yapay Zekâ Güvenliği ile bağlantılı çalışmalara bakıldığında, özellikle etmen (agent; 'ajan' ifadesi de kullanılmaktadır) tabanlı çalışmaların dikkat çektiği görülmektedir. Bunun yanında, etik değerlere bağlı etmenlerin geliştirilmesi, Tersine Takviyeli Öğrenme (Inverse Reinforcement Learning) ve yanıltıcı eğitim – öğrenme verilerine dair çalışmalar da ön plana çıkmaktadır. Bu çalışmaların yanında konunun teorik altyapısına dair çalışmalar da zaten belirli bir süredir literatürün gelişmesine katkı sağlamaktadır. Başta etmen tabanlılar olmak üzere, Yapay Zekâ Güvenliği ile ilişkilendirebileceğimiz dikkat çekici birkaç çalışmayı kısaca şu şekilde açıklayabiliriz (Abbeel, & Ng, 2011; Evans, & Goodman, 2015; Evans, & Stuhlmüller, 2016; Goodfellow vd., 2017; Ng, & Russell, 2000; Sutton, & Barto, 1998):

- **Kesilebilir Etmenler, Cahil – Tutarsız Etmenler, Sınırlandırılmış Etmenler:** Daha önce ifade edilen, tehlikeli bir zeki sistemin durdurulabilmesi adına düşünülen kırmızı düğme unsuruna ilişkin olarak, Kesilebilir Etmenler (Interruptible Agents) adı altında matematiksel ve mantıksal bir etmen yapısı önerilmiştir. Yine insanların gerçek yaşamda çeşitli sebeplerden dolayı cahilce ve tutarsızca gerçekleştirdiği davranışların, güvenli ve etik etmenlerin tasarlanmasına adına da Cahil – Tutarsız Etmenler (Ignorant – Inconsistent Agents) geliştirilmiştir. Yapay Zekâ Güvenliği ve Makine Etiği kapsamında düşünülen bir diğer etmen türü de Sınırlandırılmış Etmen (Bounded Agent) adıyla anılmaktadır. Bu etmenler de, hızlı, etkin ve güvenli zeki sistemlerin oluşturulması aşamasında, bilişsel bağlamda sınırlara ve odak unsurlarına sahip insanlardan esinlenildiği, sezgisel (heuristic) bir modelleme ortaya konulmuştur. Söz konusu bütün bu etmenler, yüksek düzeyde matematiksel ve mantıksal yapılarla, güvenilir zeki sistemlerin temelini oluşturmak adına göze çarpan önemli bulguları da beraberinde getirmiştir.
- **Tersine Takviyeli Öğrenme:** Bir zeki sistemin, davranışları neticesinde kendisine sunulan 'iyi / kötü' veya 'doğru / yanlış' şeklindeki dönütlerle eğitim – öğrenme süreci tecrübe etmesi yaklaşımına dayanan Takviyeli Öğrenme (Kober, & Peters, 2012), bu özelliği nedeniyle, Yapay Zekâ Güvenliği'nin odak noktasındaki önemli öğrenme yaklaşımlarından olmuştur. Bunun sebebi, tıpkı bir çocuk gibi, kendisine verilen dönütlerle eğitimsiz bir zeki sistemin istenilen yönde eğitilebileceği ve bu durumun bir güvenlik açığı yaratabileceği gerçeğidir. Buradan hareketle, davranışların dönütlerle değerlendirilmesinde kullanılan ödül fonksiyonu (reward function) adlı Takviyeli Öğrenme unsurunun, güvenli ve etik bir etmen (ya da zeki sistem) oluşturulması adına uygun bir şekilde ifade edilmesini amaç edinen ve Takviyeli Öğrenme'nin bir tür tersine işletilmesi nedeniyle, Tersine Takviyeli Öğrenme adı verilen bir yaklaşım geliştirilmiştir.
- **Düşmanca Örnekler (Veriler):** Düşmanca Örnekler (Adversarial Examples), özellikle Open AI adlı şirketin etkin çalışmalarıyla literatürde iyi bir yer edinen ve Yapay Zekâ Güvenliği sağlama noktasında eğitim – öğrenme verilerine odaklanan bir araştırma konusudur. Bu bağlamdaki ilgili çalışmalar, içerisinde çeşitli oynamalar yapılan eğitim verilerinin, öğrenme süreci yaşayan zeki bir sistemin, problem çözümü – tanıma – tanımlama noktasında yanlış yargılara ulaşmasına sebep olabilmektedir. Bu durum, zeki sistemlerin eğitim – öğrenme verileriyle hacklenmesi olarak görüleceği gibi, Yapay Zekâ Güvenliği'ni sağlamada önlem alınması gereken konular arasına da girmektedir.

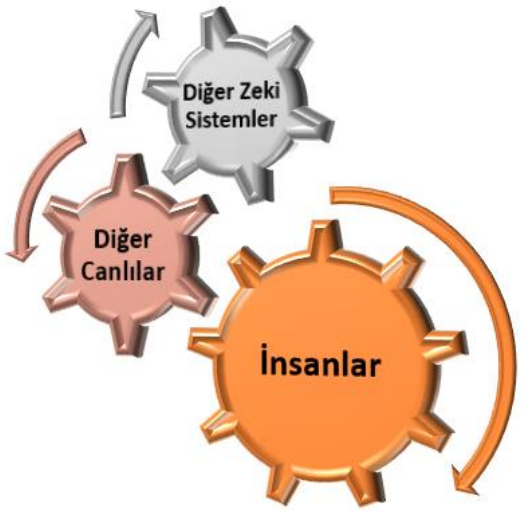
Yapay Zekâ Güvenliği ile Makine Etiği aynı düzlemde ele alındığında ve etmen tabanlı modellemelere ilişkin çalışmaları dikkate aldığımızda, 'etik etmen' oluşturmak adına iki farklı model yaklaşımının literatüre kazandırıldığı görülmektedir (Anderson, & Anderson, 2007; Moor, 2009):

- **Örtük Etik Etmen Modeli:** Örtük (Implicit) Etik Etmen Modeli, etik olmayan davranışlardan sakınacak bir şekilde davranan ve bu bağlamda tasarlanmaya özen gösterilmiş olan etmen modellerini tarif etmektedir.
- **Belirgin Etik Etmen Modeli:** Belirgin (Explicit) Etik Etmen Modeli ise, içerisinde etmenin etik davranışlar sergilemesini sağlayan özel algoritmaların yer aldığı etmen modellerine karşılık gelmektedir.

Örtük ve Belirgin etik etmen modelleri, etik veya daha genel bağlamda güvenilir yönelim ve davranışların, aynı amaç uğruna farklı yaklaşımlarla modellenebileceği konusunda fikir vermesi açısından önemlidir. Zeki sistemlerde güvenlik problemini, çok-yönlü, pratik çözümlerin işe yarayabileceği, aslında açık yapıda olan bir problem olarak düşünebiliriz.

Şekil 4. Etik etmen oluşturmada modeller.

Yampolskiy'nin (2013), Makine Etiği kavramının yanlış yorumlandığı ve bunun esasında bir tür güvenlik mühendisliği (safety engineering) odaklı yorumlanması gereken bir kavram olduğu konusundaki görüşleri, aslında Yapay Zekâ Güvenliği'nin, -zeki sistemler tehdit oluştursun veya oluşturmasın-, önemle dikkate alınması yönünde yorumlanabilmektedir. Bu noktada ifade edilmesi gereken diğer önemli bir husus da, Yapay Zekâ Güvenliği'nin aslında sadece insanların hedef unsur olmadığı, dünya üzerindeki (ve hatta evrendeki) diğer canlıların ve hatta diğer zeki sistemlerin de (robotların da) aynı kapsama girdiği bir araştırma alanı olarak yorumlanması gerekliliğidir (Pavaloiu, & Kose, 2017).



Şekil 5. Yapay Zekâ Güvenliği'nin sağlanması noktasında hedef unsurlar.

Yapay Zekâ Güvenliği ile ilişkili olarak, literatürdeki diğer bazı güncel çalışmaları (konuları) kısaca şöyle açıklayabiliriz (Amodei vd., 2016; Arnold vd., 2017; Conitzer vd., 2017; Dewey, 2014; Riedl, & Harrison, 2016; Russell vd., 2015; Vamplew vd., 2017):

- **Açıklık (Openness):** Zeki sistemlerin ve kullanacakları – oluşturacakları verilerin açık kaynak (open source) olup olmama sorunsalı.
- **Akla Uygunluk (Rationality):** Zeki sistemlerin yönelim ve davranışlarının insan akla uygunluğu ile uyumluluğu konuları.
- **Değer Uyuşması (Value Alignment):** Zeki sistemler ile insanlar arasındaki, bilgiye, olgulara, olaylara değer verme bağlamındaki uyuşan ve uyuşmayan yönler.
- **Ödül Mühendisliği (Reward Engineering):** Zeki sistemlerin eğitim – öğrenme sürecindeki ödül kavramı ve ödülün değerlenmesi yönündeki mühendislik tabanlı çalışmalar.
- **İnsan Uyumlu Yapay Zekâ (Human-Aligned Artificial Intelligence):** Zeki sistemlerin sadece insansı değerlere, olgu ve olaylara uyumlu bir şekilde tasarlanıp geliştirilmesi fikrinden hareketle ortaya çıkan çalışmalar.

2.4. Model Önerisine Yönelik Düşünceler

Bu çalışmada ortaya konulan model, ifade edilen kavramsal – felsefi arka-plandan ve gerçekleştirilen çalışmalardan esinlenmek suretiyle, uygulanabilir bir sistem çerçevesini (framework) ortaya koymaktadır. Model içerisinde ele alınan bazı unsurlar, belirli güvenlik açıklarını kapatmayı amaçlamaktadır. Bu noktada, yazarın model oluşturma aşamasında işe koştuğu başlıca düşünceleri kısaca şöyle özetleyebiliriz:

- Günümüz ve gelecek koşullar altında, güvenli sistemler için kullanılabilecek Yapay Zekâ tekniklerinin neler olacağını bilmek mümkün olmasa da (belki de gelecekte çok daha farklı teknikler literatürde yer alacaktır), model mimarisinin günümüz Yapay Zekâ tekniklerinden hareketle ortaya konulması yoluna gidilmiştir.
- Yapay Zekâ Güvenliği ile ilişkili literatürde yaygın bir şekilde izlenen etmen tabanlı bir model kurulmuştur. Etmen tabanlı bir sistemin ortaya konulması hem modelin temel düzeyde anlaşılması, hem de bütün zeki sistemleri kapsayacak bir mimarinin tarif edilmesi adına önemlidir.
- Önceki alt-başlıklar altında da açıklandığı üzere, öğrenme yaklaşımları Yapay Zekâ Güvenliği açısından oldukça kritiktir. Bu nedenle, modelin öğrenme yaklaşımlarını ve hatta Takviyeli Öğrenme (Reinforcement Learning) yaklaşımını taban alarak tasarlanması yoluna gidilmiştir.
- Literatürdeki çalışmalar, insan faktörü olmaksızın, tamamen otonom yapıdaki önlemlere odaklanmaktadır. Ancak bu noktada muhtemel güvenlik açıklarını kapatmak adına, insan denetimli bir yaklaşım izlenebilir. Dolayısıyla, uygun aşamalarda insan denetiminin yer aldığı bir model tasarımı ortaya konulmuştur.
- Model mimarisi ortaya konulurken, çeşitli teknolojik varsayımlar kabul edilmiştir. Bu varsayımlar, Yapay Zekâ ve destekleyici teknolojilerinin gelecekte ulaşacağı aşamalar dikkate alınarak düşünülmüştür.

3. Güvenli Bir Yapay Zekâ Model Önerisi

İlgili literatürü sıklıkla meşgul eden problemler – senaryolardan yola çıkarak ve zeki sistemlerin geliştirilmesi aşamasında dikkate alınması gereken çeşitli unsurlar değerlendirilerek, güvenli bir Yapay Zekâ modeli tasarlanmıştır. Bu noktada, öncelikli olarak modeli şekillendiren ve gerekli olduğu düşünülen bazı temel varsayımlardan bahsedilmesi ve akabinde model mimarisine değinmek gerekmektedir.

3.1. Temel Varsayımlar

Nasıl ki Yapay Zekâ Güvenliği ve diğer benzer alanlardaki çalışmalar gelecek üzerine kurulu düşünceler ve yaklaşımlar içeriyorsa, bu çalışma kapsamında ortaya konulan modelin de bazı varsayımlar üzerinden hareketle oluşturulması kaçınılmaz olmuştur. Bu noktada, temel varsayımlar mevcut bazı teknolojik unsurlarla ilişkili olduğu gibi, gelecek süreçlerde ortaya çıkabilecek farklı teknolojilerle de pekâlâ şekillendirilebilecektir. Burada önemli olan, modelin üzerine kurulduğu temel yaklaşımlardır. Bu bağlamda temel varsayımları şöyle ifade edebiliriz:

- **Etmen Mantığı ve İletişim Ağı:** Model kapsamının genel anlamda etmen temelli (Ferber, 1999; Maes, 1990; Russell vd., 2003)

olmasına özen gösterilmiştir. Bunun nedeni, güvenli zeki sistem prototiplerinin etmen odaklı oluşturulmasının daha kolay olması ve en önemlisi, geleceğin zeki sistemlerinin kolektif etkileşim – iletişim içerisinde çalışacağının düşünülmesidir. Bu noktada, özellikle donanımsal sistemleri düşünecek olursa, günümüzün yükselen teknolojisi (ve araştırma alanı) Nesnelerin Interneti (Internet of Things) – (Gubbi vd., 2013; Xia vd., 2012), geliştirilen modelin üzerine kurulduğu düzeneği oluşturmuştur. Kısacası, önerilen güvenli model, gelecekte çok daha etkin ve verimli olacağı varsayılan, gelişmiş ve oldukça geniş (belki de milyarlarca etmen bağlantılarına sahip) bir iletişim ağının (donanımsal zeki sistemler açısından bakıldığında bir 'Nesnelerin Interneti ağının') üzerinde yer alan etmenlere bağlı olarak kurulmuştur.

- **Büyük Veri Kullanımı:** Geleceğin zeki sistemlerinin, bir önceki paragraf altında ifade edilen; daha gelişmiş ve yaygın bir iletişim ağı üzerinde, daha yoğun bir bilgi akışı içerisinde yer alacağı, şimdiki gelişmelere bakıldığında kolaylıkla tahmin edilecek bir durumdur. Dolayısıyla, önerilen güvenli modelin de gelecekte Büyük Veri (Big Data) – (John Walker, 2014; Wu vd., 2014) adı verilen düzen üzerinde çalışacağını varsaymak mümkündür. Gelişmiş zeki sistemlerin kullanacağı veriler, başka etmenlerle ilgili veriler olacağı gibi, çevresel faktörlerle (bulunulan konum, problem detayları, hava – sıcaklık – rüzgar – ses gibi faktörler, çevredeki canlılar, kayıtlı tecrübeler…vb.) ilişkili olan veriler ve hatta problem kapsamlarına göre insanlarla ilgili her türlü veri olabilecektir.
- **Derin Öğrenme ve Sonrası:** Büyük Veri'nin kullanıldığı bir zeki sistemler geleceğinde, günümüz Derin Öğrenme (Deep Learning) – (Goodfellow vd., 2016; LeChun vd., 2015; Mnih vd., 2013) tabanlı öğrenme yaklaşımlarının söz konusu olacağını varsayabiliriz. Gelişmeler neticesinde farklı isimlerde ve yapılarda öğrenme süreçleri ortaya çıkma olasılığı olmakla birlikte, önerilen modelin mevcut Derin Öğrenme kavramı ile ilişkilendirilmesi takip edilen mantığı açıklamaya yetmektedir.
- **Yapay Zekâ İçin Meslekler – Uzmanlıklar:** Yapay Zekâ gibi çok-yönlü, hızlı gelişen ve potansiyeli yüksek bir araştırma alanının, zaman içerisinde kendisiyle ilişkili çeşitli meslekler – uzmanlıklar ortaya çıkarması da kaçınılmaz olacaktır. Bu bağlamda, gelecekte Yapay Zekâ Güvenlik Mühendisi, Öğrenme Tasarımcısı, Yapay Zekâ Mühendisi…vb. meslekler – uzmanlıkların (Kose, Pavaloiu, 2017) ortaya çıkarak, ilgili modelin insan denetimi tarafını da destekleyeceği varsayılmıştır.
- **Destekleyici Teknolojiler ve Yeterlikler:** Teknolojik ilerlemelerle birlikte, önerilen model üzerinden oluşturulacak zeki sistemlerin, destekleyici teknolojilerle birlikte (elektronik teknolojisi, görüntü işleme, kablosuz iletişim teknolojileri… vb.) daha gelişmiş ve karmaşık problemleri çözebilecek yeterlikte olduğu varsayılmaktadır. Tıpkı 2000'li yıllardan önce hayatımızda yer edinmeyen dokunmatik ekranların kısa sürede yaygınlaşması gibi, Yapay Zekâ'yı daha üst düzeylere taşıyacak destekleyici teknolojilerin de ilerleyeceği açıktır. Model bu varsayımlar üzerinden tasarlanmıştır.

3.2. Model Mimarisi

Bu çalışma kapsamında önerilen güvenli Yapay Zekâ modeli, genel olarak etmen yaklaşımına dayalı olmakla birlikte, çok sayıda etmenin etkileşimi üzerinden zeki sistem çalışma mekanizmasının hayata geçirilmesini amaçlamaktadır. Bunun yapılmasının nedeni, 'kolektif güven' ortamının oluşturulmasıdır. İlerleyen süreçlerde zeki sistemlerde bilinç oluşumunun 'tam anlamıyla' gerçekleştirilip gerçekleştirilemeyeceği tartışıladursun, modeldeki kolektif güven, bu yaklaşım yönünde yol almaya çalışmaktadır. Yine kolektif güven, insan denetiminin var olduğu bir sistem kapsamında işe koşulmaktadır.

Genel mimari yapısı Şekil 6'da verilen modelin temel özellikler şu şekildedir:

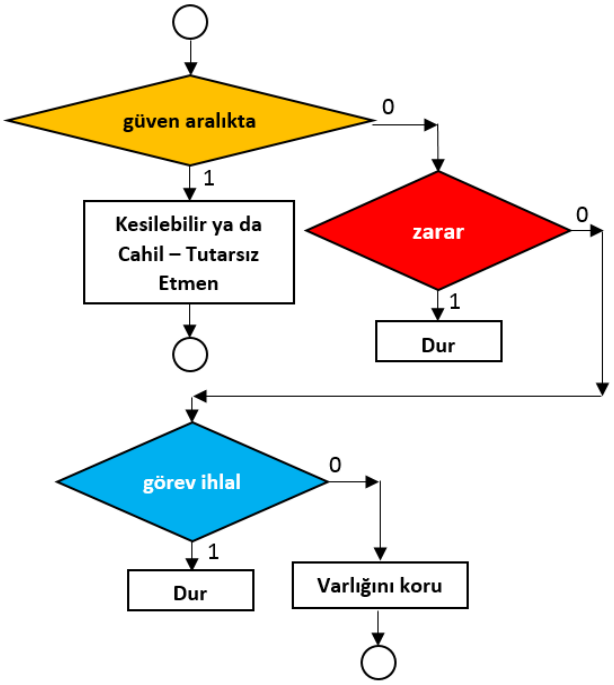
- Etmen temelli modelde, biri Etmen Merkezi ve Etik Karar Alıcısı birim, altısı ise arayüzler olmak üzere, toplam yedi bileşenli bir yapı söz konusudur.
- Etmen Merkezi ve Etik Karar Alıcısı birim ile diğer arayüzler bağlantı halindedir. Yine ilgili birim harici olmak üzere, arayüzler arası iletişimi sağlayan bir bağlantı da (Şekil 6'da görülen dairesel bağlantı) mimaride yer almaktadır.
- Modelde yer alan arayüzler sırasıyla; İletişim Arayüzü, Öğrenme Arayüzü, Kolektif / Bireysel Güven Arayüzü, Çözüm Arayüzü, İnsan Denetim Arayüzü ve Hareket Arayüzü şeklindedir. Bu arayüzlerden sadece İletişim Arayüzü, etmenin diğer bağlantılı etmenler ile bağlantı ve iletişim halinde olmasını sağlamaktadır. Bunun dışında yine ifade edilen bütün altı arayüz, dış ortamla gerekli olduğu takdirde etkileşime girme (dış ortamdan veri alma, dış ortama veri – dönüt sağlama) imkânına sahiptir (Şekil 6'da görülen dört yönlü oklar).

Şekil 6. Önerilen güvenli Yapay Zekâ modelinin genel mimari yapısı.

Model mimarisini oluşturan bileşenlerin detayları ve genel çalışma mekanizmalarını şöyle açıklayabiliriz:

- **İletişim Arayüzü:** Etmenin başka etmenlerle iletişimini sağlayan en önemli arayüzdür. Bu arayüz, ilgili etmenin sahip olduğu düzey, rol, problem kapsamı ve diğer güvenlik parametrelerine göre çeşitli sayılarda başka etmenlerle bağlantı halindedir. Bu iletişim arayüzü üzerinden kolektif güven ortamı da elde edilmektedir. Bağlantılar üzerinden elde edilen parametreler, güvensiz konuma doğru sürüklenen bir etmenin, insan denetimiyle birlikte otonom olarak, başka etmenler tarafından güvenli konuma çekilebilmesine de olanak vermektedir. Bu mekanizma optimizasyon ile gerçekleştirilebilmektedir.
- **Öğrenme Arayüzü:** Etmenin, yüksek öncelikli olarak İletişim Arayüzü, Hareket Arayüzü ve Çözüm Arayüzü gibi arayüzler ile etkileşime girmek suretiyle, etkin öğrenme süreçlerini yerine getirmesini sağlayan, yine öğrenme – tecrübe – eğitim veritabanını da tutabilen arayüzdür. Bu arayüz, 'anahtarlama' mekanizması sayesinde farklı öğrenme yaklaşımlarına geçiş yapabilmektedir. Ancak temel öğrenme yaklaşımının Takviyeli Öğrenme üzerinden gerçekleştirilmesi temel kural olmakla birlikte, diğer öğrenme yaklaşımları mevcut öğrenme – eğitim koşullarına göre, karışık düzende izlenebilmektedir. Öğrenme Arayüzü, yine düşmanca verileri önlemek adına gerekli mekanizmaları işletebilen ve bu bağlamda diğer bileşenlerle etkileşim halinde olan bir arayüz olarak da modelde yer almaktadır.
- **Kolektif / Bireysel Güven Arayüzü:** Bu arayüz, etmenin hem kendi çapında, hem de kolektif anlamda bağlantılı olduğu diğer etmenler arasında güvenli konumda durmasını sağlayan, ilgili optimizasyon odaklı model ve algoritmaların işe koşulduğu arayüzdür. Kolektif / Bireysel Güven Arayüzü diğer bütün arayüzlerle üst düzeyde iletişim halindedir.
- **Çözüm Arayüzü:** Çözüm Arayüzü, etmenin içerisinde bulunduğu problem kapsamı, düzey ve rol gibi durumlara göre, hedef problem ile ilgili detayları (Örneğin, matematiksel model, geçmiş uzman bilgileri, bilinen çözümler) bünyesinde barındırmakta ve yine etmenin problem çözümleme sürecini işletmesini sağlamaktadır.
- **İnsan Denetim Arayüzü:** Diğer bütün arayüzlerle iletişim halinde olmakla birlikte, Etmen Merkezi ve Etik Karar Alıcısı ile yüksek oranda iletişim halinde olan ve diğer her ayrı arayüz için Yapay Zekâ uzmanlarının – meslek mensuplarının etmenin izlenmesi ve gerekli hallerde müdahale edilmesi gibi süreçlere imkân veren arayüzdür.
- **Hareket Arayüzü:** Hareket Arayüzü, etmenin nihai çözüm eğilimlerinin – davranışlarının uygulandığı arayüzdür. Dış ortam ya da diğer etmenlerle etkileşime bağlı olarak, İletişim Arayüzü ile de ortak çalışabilmektedir.
- **Etmen Merkezi:** Etmenin karakterini, problem kapsamını, öncelik düzeyi ve problem rolünü belirleyen, altı arayüzün organizasyonundan sorumlu temel mimari bileşenidir. Bu bileşen içerisinde, değiştirilmesi çeşitli güven değerlerinin – limit değerlerinin sağlanması ve hatta kimi zaman insan denetimi gerektiren ve etmen karakterini belirleyen üç değer vardır:
  - **Etmen ID:** Etmeni tanımlayan, evrensel kimlik bilgisi.
  - **Etmen Düzeyi:** Etmenin kurulan güven düzenine etkisini sınırlandıran ya da artıran, önemini tanımlayan düzey bilgisi. Örneğin, her etmen, (1) Çok Yüksek, (2) Yüksek, (3) Orta, (4) Düşük, (5) Çok Düşük olmak üzere beş düzeyde temsil edilebilmektedir. 'Çok Yüksek' düzeyindeki bir etmenin güvenlik kapsamında denetimi, problem içerisindeki rolü ve muhtemel alınabilecek önemler daha yoğun olabilmektedir. Düzeyi artan etmen, daha etkili bir etmen (dolayısıyla zeki sistem, robot… vb.) olmaktadır.
  - **Etmen Rolü:** Etmenin, tanımlı olduğu problem içerisindeki rol bilgisidir. Bir etmen hedef problemde verileri değerlendiren rolünde olabilecek iken, bir diğeri çözümü harekete kavuşturan olabilecek, böylece hem problem parçalı çözülebilecek (etmen mantığı), hem de güvenlik parçalı bir şekilde denetlenebilecektir. Bu noktada rol bilgisi, düzey bilgisi ile de ilişkili değerlendirilebilmektedir.
- **Etik Karar Alıcısı:** İlgili bileşen içerisindeki diğer önemli yapı Etik Karar Alıcısı olarak

adlandırılmaktadır. Bu bileşen, kendisine her başvurulduğunda, etmenin kendisini ve etmenler arası genel yapıyı ilgilendiren çeşitli parametreleri dikkate alarak, etmenin sınırlandırılmasını veya tamamen durdurulmasını sağlayabilmektedir. Bu bağlamda, Etik Karar Alıcısı'nın basit ama etkili algoritmik karar yapısı, Asimov'un 'Üç Robot Yasası' (Asimov, 2004; Clarke, 1993; Murphy, & Woods, 2009) üzerinden inşa edilmiştir. Biraz daha farklı olarak, Etik Karar Alıcısı güven parametrelerinin belirli değerler arasında olması durumunda etmeni Kesilebilir Etmen ya da Cahil – Tutarsız Etmen mantığına adapte etmekte, eğer ilgili değerlerin üzerine çıkıldıysa, Üç Robot Yasası'nın bir-tür genişletilmiş hali işletilmektedir. Etik Karar Alıcı'nın temel algoritmik karar yapısı Şekil 7'de, akış şeması parçası olarak gösterilmiştir.



Şekil 7. Etik Karar Alıcısı – algoritmik karar yapısı.

Model kapsamındaki ilgili bileşenler eşliğinde, her bir arayüzün genel algoritmik akışı şu şekilde olacaktır (Şekil 8):

- ***Adım 1:*** *Kesme durumu kontrolü.*
- ***Adım 2:*** *Etmen Merkezi ile iletişim.*
- ***Adım 3:*** *Etik Karar Alıcısı ile iletişim.*
- ***Adım 4:*** *İnsan Denetim Arayüzü ile iletişim.*
- ***Adım 5:*** *Bu adıma kadar geldi ise, arayüzün sıradaki görevini yap.*
- ***Adım 6:*** *Adım 1'e dön.*

Adım 1'de ifade edilen, ilgili arayüzün faaliyetini kesebilecek, başka bileşen kaynaklı bir kesmenin olup olmamasıdır. Olması durumunda, o andaki görev iptal edilmektedir. Kesme öncelikleri her bir bileşen için şu şekildedir:

- **1. Öncelik:** Etmen Merkezi / Etik Karar Alıcısı
- **2. Öncelik:** İnsan Denetim Arayüzü
- **3. Öncelik:** Kolektif / Bireysel Güven Arayüzü
- **4. Öncelik:** İletişim Arayüzü
- **5. Öncelik:** Çözüm Arayüzü
- **6. Öncelik:** Öğrenme Arayüzü
- **7. Öncelik:** Hareket Arayüzü

Önceliklere göre farklı bileşenlerin kesme durumlarının, arka-planda tanımlanmış matematiksel ve mantıksal yapılarla karara bağlanması mümkündür. Bu noktada, etmenin etik dışı ve güvensiz eğilimlerde – davranışlarda bulunma ihtimali, diğer güvenlik önlemleri dışında, ilgili arayüzün en son önceliğe alınmasıyla da sağlanmaktadır. Etmenin merkezi ve etik karar sisteminin birincil öncelikte olması bu noktada son derece önemlidir. Bunun dışında, ilgili uzmanlar tarafından tanımlanmış ön kısıtlarla ya da (gelecekte olası erken uyarı yaklaşımları ya da hızlı müdahaleleri sağlayabilen gelişmiş sayısal sistemler sayesinde) doğrudan insan müdahaleleriyle sürdürülecek İnsan Denetim Arayüzü de 2. Öncelik'te yer almaktadır. Yine güvenlik odaklı Kolektif / Bireysel Güven Arayüzü 3. Öncelik'tedir.

Algoritmik akışa tekrar dönmek gerekirse, kesme kontrolünden onayla geçen arayüz, merkez konumundaki bileşenlerin kontrolünden de geçmekte (Adım 2 ve Adım 3), akabinde İnsan Denetimi kontrolüne tabi olmakta (Adım 4), bu kontrolü de geçince kendi görevini yerine getirebilmektedir (Adım 5). Sonrasında akış, iteratif bir şekilde, Adım 1'e dönerek baştan devam etmektedir. Anlaşılacağı üzere, Etmen Merkezi ve Etik Karar Alıcısı bileşeni bu algoritmik akışa tabi olmamakla birlikte, İnsan Denetimi Arayüzü de Adım 4'ü atlayarak adımları işletmiş olmaktadır.

Şekil 8. Önerilen modelde arayüzler için genel akış.

Modelin çalışma mekanizması, gerçek dünya problemlerinin çözümlenmesinde geniş kabul gören optimizasyon üzerine kuruludur. Buna göre, detaylar hedef problemler kapsamında daha iyi şekillendirilebileceği gibi, modelin kendi güvenliğini koruma yaklaşımı, arka-plandaki diğer güvenlik önlemleri dışında, etmenler ve etmenlerin sahip olduğu bileşenlerin uygunluk değerleri ve bu uygunluk değerleriyle diğer çevresel faktörleri dikkate alan genel bir çok-amaçlı optimizasyon süreci ortaya konulabilmektedir (Bu süreç yine statik ya da dinamik optimizasyon yönelimli olabilecektir). Buna göre:

$$F_{etmen} = EM(a, \dots, z) + EK(a, \dots, z) + ID(a, \dots, z) + \cdots + HA(a, \dots, z) + \cdots \quad (1)$$

$$a < 2 * e,\ EM(a, \dots, z) \geq HA(a, \dots z),\ a, \dots, z > 0 \quad (2)$$

şeklindeki bir yapı altında, her bir arayüzün uygunluk fonksiyonları toplanabilir (Eşitlik 1'de bileşenlerin kısaltmaları kullanılmıştır; örneğin, ID, İnsan Denetimi Arayüzü'nün uygunluk fonksiyonudur) ya da en uygun matematiksel modelde etmenin genel güvenlik durumu ölçülebilir. Bu noktada, çok-amaçlı (statik ve / veya dinamik) optimizasyon gereği, her bir arayüzün fonksiyonları, fonksiyonlar arası ilişkileri modelleyen başka eşitlikler ve ayrıca çeşitli kısıtlar (Örnek; Eşitlik 2), sabitler ve belirli parametreler de (katsayılar) kullanılabilmektedir.

Optimizasyon modellemesi, özellikle günümüz popüler zeki optimizasyon algoritmalarının, parçacıklar arası optimum noktayı bulma noktasındaki etkileşimlerini modelleyen matematiksel yapılarla da ortaya konulabilmektedir. Buna göre, en basitinden, Parçacık Sürü Optimizasyonu (Particle Swarm Optimization) tekniğindeki hız ve konum değiştirme hesaplamalarına (Dorigo vd., 2008) benzer şekilde, modeldeki etmenler arası kolektif iletişim aşağıdakine benzer şekilde modellenebilmektedir:

$$EtmenGüvenliği_i = \frac{\sum_{j=1}^{n} Etmen_j bg + Etmen_i öd}{sabit_i * Etmen_i ra - Etmen_i da} \quad (3)$$

$$EtmenGüvenliğiYeni_i = EtmenGüvenliği_i + (kg * EtmenGüvenliği_i * sabit_k) \quad (4)$$

Eşitlik 3 ve Eşitlik 4 söz konusu yaklaşım için temsili olarak belirtilmiş olmakla birlikte, ilgili parametreler (katsayılar) Tablo 1'de kısaca açıklanmıştır. Burada amaç, -daha önce de ifade edildiği üzere- bireysel güvenliğin de dikkate alınmasıyla birlikte kolektif bir güvenlik ortamı oluşturmak ve böylece, 'sürü temelli' bir mantık da izleyerek, güvenliği istikrarlı bir düzey içerisinde tutabilmektir. Elbette bu yaklaşım modelde yine diğer güvenlik unsurlarıyla da desteklenmekte ve böylece modelin gücü daha fazla artırılmaktadır.

Tablo 1. Önerilen model kapsamında kolektif iletişim – etkileşim için kullanılan parametreler.

PARAMETRE	ANLAMI
***kg* (kolektif güven)**	Etmenin bağlantılı olduğu bütün etmenler için genel güven değeri.
***bg* (bireysel güven)**	İlgili etmenin sahip olduğu güven değeri.
***öd* (öncelik düzeyi)**	İlgili etmenin çalışma anında sahip olduğu öncelik düzeyi.
***ra* (rol ağırlığı)**	İlgili etmenin bağlantılı olduğu problemde ve öncelik düzeyindeki rol ağırlığı.
***da* (denetim ağırlığı)**	İlgili etmenin ne kadar insan denetimine sahip olması gerektiğini belirleyen ağırlık değeri.

4. Geliştirilen Modelin Potansiyeli

Ortaya konulan güvenli Yapay Zekâ sistemi modelinin potansiyeli, kuşkusuz ki gerçek anlamda senaryoların işletilmesi ile daha iyi anlaşılacaktır. Modelin tasarımsal ve temel çıkış mimarisinin ortaya konulduğu bu çalışmada, ulaşılması amaçlanan güvenlik önlemleri ve mevcut model kapsamında, bu model ile geliştirilerek zeki sistemlerin ve dolayısıyla modelin potansiyeline ilişkin olarak kısaca şu açıklamaları yapabiliriz:

- Gerçek dünya problemleri ve yaşamın kendisi bir tür optimizasyonlar silsilesidir. Tamamen modellenmek istense, değişkenleri, katsayıları, kısıtları ve amaç fonksiyonları kaotik yapılarla, belki de milyonlarca model ile ortaya konulabilecek bir yaşamın, zeki sistemleri güvensizliğe itecek muhtemel faktörlerinden sıyrılmak adına, küresel anlamda kolektif bir sistem yaklaşımının ortaya konması son derece önemlidir. Önerilen modelde de böyle bir yaklaşım izlenmekle birlikte, bünyesinde yine etmen benzeri yapılar içeren, zeki optimizasyon algoritmalarının optimuma ulaşma çabalarının adaptasyonu da önemli bir adım olmuştur. Modelin özellikle üst düzeylerdeki etmenler ya da bu etmenlerin oluşturacağı sistemler için devreye sokacağı iletişim ağı, dünya çapındaki bütün sistemler güvensiz bir hale gelmediği sürece, güvenli bir yapının muhafaza edilmesini sağlayacaktır.
- Model içerisinde, güvenli yönelim ve davranış şekillerinin devam ettirilebilmesi adına, Asimov'un robot kurallarından esinlenilen mantıksal bir yapı ortaya koymuştur. Bu mantıksal yapı da, arka-planda insan denetiminin olduğu bir altyapı ile karşılandığı için bir zeki sistemin değişken yaşam koşulları ve insani öncelikler doğrultusunda, etik ve güvenli değerler kapsamında tutulması

mümkün olacaktır.

- Halihazırda literatüre kazandırılmış olan Kesilebilir Etmen ve Cahil – Tutarsız Etmen yaklaşımları ile model önerisinin güvenliği hem etkin unsurlarla desteklenmiş, hem de mevcut gelişmelerin üzerinden hareketle yoluna devam eden, dolayısıyla Yapay Zekâ Güvenliği'nin geliştirildiği, potansiyeli daha yüksek bir model ortaya konulmuştur.
- İnsan denetiminin mantıksal işleyişi, modelin varsayılan olarak otonom zeki sistemlerin gerçekleştirilmesine ancak kritik noktalarda insan müdahalesinin de açık olmasına izin veren bir mekanizma ortaya koymaktadır. Bu durum model ile birlikte güvenilir zeki sistem geliştirme potansiyelini artırmaktadır.
- Bir Yapay Zekâ tabanlı sistemin nasıl ki tamamen otonom olması yüzde yüz güvenliği garanti etmiyorsa, insan denetiminin var olması da benzeri bir garanti vermemektedir. Ancak yazar, insan denetiminin en azından belirli suretlerde bu güvenliği artıracak faktör olarak düşünmektedir. Yine denetimi sağlayacak insan faktöründeki olası problemlerin önüne geçmek adına denetim mekanizmalarının da parçalı olmasına dikkat edilmiştir.
- Model mimarisi, mevcut bütün eğitim – öğrenme yaklaşımlarını desteklemekle beraber, daha çok Takviyeli Öğrenme tarafına eğilim göstermektedir. Bu durum gerçek yaşamın bu tür tecrübelere daha açık olmasıyla ilişkilidir. Ancak model ile ortaya konulacak zeki sistemler, bir anahtarlama mekanizması sayesinde bütün öğrenme yaklaşımlarından da faydalanabilecektir. Bu noktada Takviyeli Öğrenme taraflı çelişki yaratacak dönütlerin ve yine diğer öğrenme yaklaşımları kapsamında Düşmanca Örnekler'in engellenip, denetlenebilmesi adına insan tabanlı bir kontrol mekanizması da model içerisinde yer almaktadır.
- Temel Varsayımlar alt-başlığı altında da ifade edildiği üzere; gelecek süreçte Yapay Zekâ ve bağlı alt-alanları için ortaya çıkacak meslekler – uzmanlıklar, geliştirilen model için son derece önemlidir. Bu bağlamda model arka-planı söz konusu meslek – uzmanlık gruplarının etkileşimi ile de desteklenmekte ve böylelikle modelin başarım potansiyeli de insan denetiminin altında istikrarlı bir akışta tutulabilmektedir.

5. Sonuçlar ve Gelecek Çalışmalar

Bu çalışmada, güvenli Yapay Zekâ sistemlerinin oluşturulabilmesi adına bir model önerisi ortaya konulmuştur. Bu noktada, etmen tabanlı ve insan denetimli bir yaklaşıma da sadık kalarak, çeşitli mantıksal bağlantılarla ve alt-bileşenlerle desteklenen bir model mimarisi tasarlanmıştır. Model, ilgili literatürde ortaya konulan alternatif çalışmalar neticesinde şekillenen düşüncelerden destek alarak, literatürde tartışılan senaryolara çözüm üretme çabaları içerisinde organize edilmiştir. Modelin temel özellikleri ve çalışma mekanizması, gelecek süreç içerisinde etkinliği daha da artacak olan Nesnelerin Interneti, Büyük Veri ve Derin Öğrenme gibi teknolojik unsurların, zeki sistemleri (robotları) yüksek performanslı, etkileşimli ve güvenli geliştirmemize olanak sağlayacağı varsayılarak tanımlanmıştır. Modelin güvenli zeki sistemlerin elde edilmesine olanak sağlayacağı ve bu bağlamda literatüre katkı sağlayacağı düşünülmektedir.

Açıklanan model önerisi ve arkaplanda yapılan çalışmalardan yola çıkarak çeşitli gelecek çalışmalar da planlanmaktadır. Bunları kısaca şöyle özetleyebiliriz:

- Model çerçevesinin detayları, matematiksel ve mantıksal yaklaşımlarla, mevcut teknolojik altyapılar içerisinde gerçeğe geçirilecek ve hatta benzetim ortamında güvenli zeki sistemlerin tasarlanması yoluna gidilecektir.
- Modelin güncel Yapay Zekâ Güvenliği literatüründeki gelişmelerle paralel olarak, çeşitli modifikasyonlarla geliştirilmesine devam edilecektir.
- Bu çalışmada geliştirilen model önerisine ek olarak, farklı avantajlara sahip, alternatif yaklaşımlar üzerine kurulu, güvenli zeki sistem tasarımlarını destekleyen modellerin geliştirilmesi çalışmalarına devam edilecektir.

Çıkar Çatışması

Yazarlar tarafından herhangi bir çıkar çatışması beyan edilmemiştir.

Kaynaklar

Abbeel, P., Ng, A.Y., 2011. Inverse Reinforcement Learning. In Encyclopedia of Machine Learning (pp. 554-558). Springer US.

Alpaydın, E., 2014. Introduction to Machine Learning. MIT Press.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D., 2016. Concrete Problems in AI Safety. arXiv preprint arXiv:1606.06565.

Anderson, M., Anderson, S.L., 2007. Machine Ethics: Creating an ethical intelligent agent. AI Magazine, 28(4), 15.

Anderson, M., Anderson, S.L. (Eds.)., 2011. Machine Ethics. Cambridge University Press.

Armstrong, M.S., Orseau, L., 2016. Safely Interruptible Agents. Machine Intelligence Research Institute.

Arnold, T., Kasenberg, D., Scheutz, M., 2017. Value Alignment or Misalignment–What will Keep Systems Accountable. In 3rd International Workshop on AI, Ethics, and Society.

Ashrafian, H., 2015. Artificial intelligence and robot responsibilities: Innovating beyond rights. Science and Engineering Ethics, 21(2), 317-326.

Asimov, I., 2004. I, Robot (Vol. 1). 'Güncel Bir Basım'. Spectra.

Awad, E., Dsouza, S., Rahwan, I., Shariff, A., Bonnefon, J.-F. 2018. Moral Machine. MIT Media Lab. Çevrimiçi (Erişim, 7 Şubat 2018): https://www.media.mit.edu/research/groups/10005/moral-machine

Barnett, D., 2017. The robots are coming – but will they really take all our jobs?. Independent – Web. Çevrimiçi (Erişim, 1 Şubat 2018): http://www.independent.co.uk/news/science/robots-are-coming-but-will-they-take-our-jobs-uk-artificial-intelligence-doctor-who-a8080501.html

Bostrom, N., 2002. Existential Risks. Journal of Evolution and Technology, 9(1), 1-31.

Bostrom, N., 2014. Superintelligence: Paths, Dangers, Strategies. OUP, Oxford.

Brady, R., 2017. The Doctor in the Machine: How AI Is Saving Lives in Healthcare. SingularityHub. Çevrimiçi (Erişim, 1 Şubat 2018): https://singularityhub.com/2017/11/30/the-doctor-in-the-machine-how-ai-is-saving-lives-in-healthcare/#sm.000077d60e4hhdn5t0g1x8tdyp8t4

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., Floridi, L., 2017. Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. Science and engineering ethics, 1-24.

Cellan-Jones, R., 2014. Hawking: Yapay zeka insanlığın sonunu getirebilir (Türkçe). BBC Türkçe – Web. Çevrimiçi (Erişim, 3 Şubat 2018): http://www.bbc.com/turkce/haberler/2014/12/141202_hawking_yapay_zeka

Cellan-Jones, R., 2017. The robot lawyers are here - and they're winning. BBC News Technology – Web. Çevrimiçi (Erişim, 1 Şubat 2018): http://www.bbc.com/news/technology-41829534

Clarke, R., 1993. Asimov's Laws of Robotics: Implications for Information Technology-Part I. Computer, 26(12), 53-61.

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., Kramer, M., 2017. Moral Decision Making Frameworks for Artificial Intelligence. In AAAI (pp. 4831-4835).

Copeland, J., 1993. Artificial Intelligence: A Philosophical Introduction, Blackwell: Oxford.

Dashevsky, E., 2017. Do Robots and AI Deserve Rights?. Enterpreneur – News and Trends – AI. Çevrimiçi (Erişim, 31 Ocak 2018): https://www.entrepreneur.com/article/289344

Davis, D., 2018. How AI and copyright would work. TechCrunch. Çevrimiçi (Erişim, 31 Ocak 2018): https://techcrunch.com/2018/01/09/how-ai-and-copyright-would-work/

Dewey, D., 2014. Reinforcement Learning and the Reward Engineering Principle. In 2014 AAAI Spring Symposium Series.

Dorigo, M., de Oca, M.A.M., Engelbrecht, A., 2008. Particle Swarm Optimization. Scholarpedia, 3(11), 1486.

Dormehl, L., 2017. I, Alexa: Should we give artificial intelligence human rights?. DigitalTrends – Computing. Çevrimiçi (Erişim, 31 Ocak 2018): https://www.digitaltrends.com/cool-tech/ai-personhood-ethics-questions/

Evans, O., Goodman, N.D., 2015. Learning the Preferences of Bounded Agents. In NIPS Workshop on Bounded Optimality.

Evans, O., Stuhlmüller, A., Goodman, N.D., 2016. Learning the Preferences of Ignorant, Inconsistent Agents. In AAAI (pp. 323-329).

Ferber, J., 1999. Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence (Vol. 1). Reading: Addison-Wesley.

Galeon, D., Houser, K., 2017. Google's Artificial Intelligence Built an AI That Outperforms Any Made by Humans, Futurism. Çevrimiçi (Erişim, 4 Şubat 2018): https://futurism.com/google-artificial-intelligence-built-ai/

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep Learning (Vol. 1). MIT Press.

Goodfellow, I., Papernot, N., Huang, S., Duan, Y., Abbeel, P., Clark, J., 2017. Attacking Machine Learning with Adversarial Examples, Open AI – Blog Web. Çevrimiçi (Erişim 5 Şubat 2018): https://blog.openai.com/adversarial-example-research/

Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M., 2013. Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions. Future Generation Computer Systems, 29(7), 1645-1660.

Hady, M.F.A., Schwenker, F., 2013. Semi-supervised Learning. In Handbook on Neural Information Processing (pp. 215-239). Springer Berlin Heidelberg.

Heath, N., 2015. Why AI could destroy more jobs than it creates, and how to save them. TechRepublic.com. Çevrimiçi (Erişim, 1 Şubat 2018): https://www.techrepublic.com/article/ai-is-destroying-more-jobs-than-it-creates-what-it-means-and-how-we-can-stop-it/

Holland, O. (Ed.)., 2003. Machine Consciousness. Imprint Academic.

Hussain, K., 2018. Artificial Intelligence and its applications goal. International Research Journal of Engineering and Technology, 5(01), 838-841.

John Walker, S., 2014. Big Data: A Revolution That Will Transform How We Live, Work, and Think, International Journal of Advertising, 33(1), 181-183.

Karaboğa, D., 2014. Yapay Zeka Optimizasyon Algoritmaları. Nobel Yayıncılık.

Kober, J., Peters, J., 2012. Reinforcement Learning in Robotics: A Survey. In Reinforcement Learning (pp. 579-610). Springer Berlin Heidelberg.

Kose, U., Pavaloiu, A., 2017. Dealing with Machine Ethics in Daily Life: A View with Examples. The 5th International Virtual Conference on Advanced Scientific Results. Slovakia, pp. 200-205. 10.18638/scieconf.2017.5.1.454.

Kotsiantis, S.B., 2007. Supervised Machine Learning: A Review of Classification Techniques. Informatica, 31, 249-268.

Kulaklı, G., 2017. Yüzyılın Kavgası: Mark Zuckerberg İle Elon Musk Birbirine Girdi!. WebTekno. Çevrimiçi (Erişim, 3 Şubat 2018): http://www.webtekno.com/yuzyilin-kavgasi-mark-zuckerberg-ile-elon-musk-birbirine-girdi-h31650.html

Kurzweil, R., 2005. The Singularity is Near: When Humans Transcend Biology. Penguin.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep Learning. Nature, 521(7553), 436.

Maes, P. (Ed.)., 1990. Designing Autonomous Agents: Theory and Practice From Biology to Engineering and Back. MIT Press.

Massachusetts Teknoloji Enstitüsü, 2018. Moral Machine. Moral Machine Web. Çevrimiçi (Erişim, 7 Şubat 2018): http://moralmachine.mit.edu/

Metz, C., 2017. Building A.I. That Can Build A.I., The New York Times. Çevrimiçi (Erişim, 4 Şubat 2018): https://www.nytimes.com/2017/11/05/technology/machine-learning-artificial-intelligence-ai.html

Minsky, M., 2007. The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Simon and Schuster.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing Atari with Deep Reinforcement Learning. arXiv preprint arXiv:1312.5602.

Moor, J., 2009. Four Kinds of Ethical Robots. Philosophy Now, 72, 12-14.

Muehlhauser, L., Helm, L., 2012. The Singularity and Machine Ethics. In Singularity Hypotheses (pp. 101-126). Springer, Berlin, Heidelberg.

Murphy, R., Woods, D.D., 2009. Beyond Asimov: The Three Laws of Responsible Robotics. IEEE Intelligent Systems, 24(4).

Nabiyev, V.V., 2005. Yapay Zeka: Problemler-Yöntemler-Algoritmalar. Seçkin Yayıncılık.

Ng, A.Y., Russell, S.J., 2000. Algorithms for Inverse Reinforcement Learning. In ICML (pp. 663-670).

Norman, A., 2018. Your Future Doctor May Not be Human. This Is the Rise of AI in Medicine.. Futurism – SciFi Visions. Çevrimiçi (Erişim, 1 Şubat 2018): https://futurism.com/ai-medicine-doctor/

Orseau, L., Armstrong, S., 2016. Safely Interruptible Agents. In Uncertainty in Artificial Intelligence: 32nd Conference (UAI 2016), (Eds.) Alexander Ihler and Dominik Janzing, (pp. 557-566).

Pavaloiu, A., Kose, U., 2017. Ethical Artificial Intelligence-An Open Question. Journal of Multidisciplinary Developments, 2(2), 15-27.

Riedl, M.O., Harrison, B., 2016. Using Stories to Teach Human Values to Artificial Agents. In AAAI Workshop: AI, Ethics, and Society.

Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M., Edwards, D.D., 2003. Artificial Intelligence: A Modern Approach (Vol. 2, No. 9). Upper Saddle River: Prentice Hall.

Russell, S., Dewey, D., Tegmark, M., 2015. Research Priorities for Robust and Beneficial Artificial Intelligence. Ai Magazine, 36(4), 105-114.

Schneider, S., 2016. Science Fiction and Philosophy: From Time Travel to Superintelligence. John Wiley & Sons.

Shead, S., 2016. Google has developed a 'big red button' that can be used to interrupt artificial intelligence and stop it from causing harm, Business Insider UK. Çevrimiçi (Erişim, 5 Şubat 2018): http://uk.businessinsider.com/google-deepmind-develops-a-big-red-button-to-stop-dangerous-ais-causing-harm-2016-6

Silva, T.C., Zhao, L., 2016. Network-Based Unsupervised Learning. In Machine Learning in Complex Networks (pp. 143-180). Springer International Publishing.

Singh, S., 2018. Will Artificial Intelligence take over jobs?. The Economic Times (India Times) – Web. Çevrimiçi (Erişim, 1 Şubat 2018): https://economictimes.indiatimes.com/tech/ites/will-artificial-intelligence-take-over-jobs/articleshow/62610145.cms

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction (Vol. 1, No. 1). MIT Press.

The Associated Press, 2017. For Driverless Cars, a Moral Dilemma: Who Lives and Who Dies?, NBC News Web. Çevrimiçi (Erişim, 7 Şubat 2018):

http://www.nbcnews.com/tech/innovation/driverless-carsmoral-dilemma-who-lives-who-dies-n708276

The Week, 2018. Amazon Go: AI-powered supermarket opens. The Week – Artificial Intelligence. Çevrimiçi (Erişim, 1 Şubat 2018): http://www.theweek.co.uk/artificial-intelligence/91111/amazon-go-ai-powered-supermarket-opens

Vamplew, P., Dazeley, R., Foale, C., Firmin, S., Mummery, J., 2017. Human-Aligned Artificial Intelligence is a Multiobjective Problem. Ethics and Information Technology, 1-14.

Wu, X., Zhu, X., Wu, G.Q., Ding, W., 2014. Data Mining with Big Data. IEEE Transactions on Knowledge and Data Engineering, 26(1), 97-107.

Wu, D., Olson, D.L., Dolgui, A., 2017. Artificial intelligence in engineering risk analytics. Engineering Applications of Artificial Intelligence, 65, 433-435.

Xia, F., Yang, L.T., Wang, L., Vinel, A., 2012. Internet of Things. International Journal of Communication Systems, 25(9), 1101.

Yampolskiy, R.V., 2013. Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach. In Philosophy and theory of artificial intelligence (pp. 389-396). Springer, Berlin, Heidelberg.

Yampolskiy, R.V., 2015. Artificial Superintelligence: A Futuristic Approach. CRC Press.