2021

3(1), 1-24

Yapay Zekâ Risklerinin Etik Yönünden Değerlendirilmesi

Ahmet Efe*[1]

Anahtar Sözcükler

Yapay Zeka
Derin Öğrenme
Teknoloji Riski
Etik

Makale Hakkında

Gönderim Tarihi
13 Ocak 2021
Kabul Tarihi
25 Şubat 2021
Yayın Tarihi
30 Haziran 2021

Makale Türü
Araştırma Makalesi

Öz

Yapay Zekâ (YZ), makine öğrenmesi, gelişmiş öğrenme ve ileri performans aracılığıyla süreçlerdeki ekonomiklik, etkinlik ve verimliliği ciddi ölçüde güçlendirme potansiyeline sahip yenilikçi teknolojilere dayanmaktadır. Ancak YZ ile ilgili bu potansiyel kabul edildiğinde, sadece teknik değil, aynı zamanda etik ve dini sonuçları da kaçınılmaz olmaktadır. Pek çok yenilikçi mühendis ve akademisyen, bilimsellik, genel bilgelik ve sosyal beceriler de dâhil olmak üzere, YZ'nin derin öğrenme sayesinde insan beyninden daha akıllı olacağından ciddi riskleri barındırdığını iddia etmektedir. Bu nedenle bunu ekonomik, sosyal, politik ve dini hayatımızdaki karar ve uygulamaları etkileyebilecek çok önemli bir faktör olarak görmek gerektiği savunulmaktadır. Bu çalışmamızda, YZ ile gelinen ve gelmesi düşünülen seviyelerin, geliştirilecek maddi süreçler kadar, manevi ve etik değerleri ne ölçüde etkileyebileceği ve muhtemel risklere karşı ne tür önlemler alınması gerektiği üzerinde analiz ve değerlendirmeler yapılmaktadır. YZ etik ve manevi değerlerden yoksun olarak programlanırsa veya kötü niyetli korsan veya teröristlerin eline geçtiğinde insanlık için yarardan çok zarar verebileceği savunulmaktadır.

Analysis of Artificial Intelligence Risks from Ethical Aspect

Keywords

Artificial Intelligence
Deep Learning
Technological Risk
Ethics

Article Info

Received
January 13, 2021
Accepted
February 25, 2021
Published
June 30, 2021

Article Type
Research Paper

Abstract

Artificial Intelligence which is leaning on the innovative technology has the potential to significantly strengthen the economy, efficiency and productivity of processes through machine learning, advanced learning and advanced performance. However, when this potential with AI is accepted, not only technical but also ethical and religious consequences are inevitable. Many innovative engineers and academicians claim that through deep learning artificial intelligence will be smarter than the human brain, including science, general wisdom, and social skills and hence poses great risks. For this reason, it is argued that it should be seen as a very important factor that can affect the decisions and practices in our economic, social, political and religious lives. In this study, analyzes and evaluations are made on the extent to which the levels reached or expected to come with AI can affect moral and ethical values as well as the material processes to be developed, and what measures should be taken against possible risks. It is argued that if AI is programmed as lacking ethical and moral values and captured by wrongdoer pirates or terrorists, it can do more harm than its expected benefits for humanity.

Atıf: Efe, A. (2021). Yapay zekâ risklerinin etik yönünden değerlendirilmesi. *Bilgi ve İletişim Teknolojileri Dergisi, 3*(1), 1-24.
Cite: Efe, A. (2021). Analysis of artificial intelligence risks from ethical aspect. *Journal of Information and Communication Technologies, 3*(1), 1-24.

*Sorumlu Yazar/Corresponding Author: icsiacag@gmail.com
[1] Dr., CISA, CRISC, PMP, Internal Auditor, Ankara Development Agency, Ankara/Turkey, icsiacag@gmail.com, https://orcid.org/-0002-2691-7517

Extended Abstract

Introduction

While most of the current practices are used to positively influence humanity, any powerful tools can be used for harmful purposes if they fall into the wrong hands. It is also possible that autonomous systems that are not encoded correctly will get out of control in the future. Today, hands-on AI products that perform a narrow task such as facial recognition, natural language processing or internet searches arouse excitement (Haenlein & Kaplan, 2019). From the Artificial Intelligence levels given in Figure 1, the second level application can now be made. However, the main race is for third-level super AI, which has great benefits as well as serious risks. Experts and investors in this field are working for super artificial general intelligence, where systems can fulfill any task that smart people can perform, and most likely defeat human beings in each. Even the most innocent accounting software package can normally be used to reconcile accounts faster but can also be used to commit corporate fraud in the wrong hands. Assuming that programming designed to facilitate such fraud is not intentionally included in coding, it may result in morally objectionable consequences (fraud) arising from its use (Jobin et al., 2019). There are also legal practices and legal ramifications in the AI domain. It is expected that AI technology will be adopted significantly in the sector in the coming period.

Method

The basic assumption of our research; AI will gradually deepen and expand, cover all sectors, and this field will inevitably progress. Our research question is that "What are the ethical problems that may arise with AI, and what are the ways to eliminate the serious risks associated with them in advance". For this reason, in our study firstly, by focusing on the risks and dangers that occur with AI, a detailed conceptual, theoretical and logical analysis study is carried out in addition to the relevant literature review within the scope of ethical and moral values. Care was taken for our study to have descriptive, relational and problem-solving research qualities.

Findings

When we want to take a look at some of the key risks associated with artificial intelligence, these are, respectively, autonomous weapons, social manipulation, breach of privacy and social grading, machine-target misalignment, discrimination, the possibility of a new religion, and AI can reduce the influence of religion.

1. Autonomous Weapons

Autonomous weapons undoubtedly provide a great advantage because they offer much more economical, effective and efficient attack and defense opportunities. The superiority in the higher number of soldiers is now obsolete. However, as with autonomous weapons programmed to death, AI programmed to do something dangerous can pose a serious risk (Hancock, 2017).

2. Social Manipulation

As stated by Goggin (2019), social media is very effective in marketing through self-driving algorithms. Because they know who we are, what we love, and your weaknesses, their chances of persuasion also increase.

3. Violation of Privacy and Social Rating

Cameras are available almost everywhere. While the GPS signals of mobile phones can determine who you are with and which groups you are most in contact with, facial recognition algorithms can also understand who you are (Hill, 2020).

4. Misalignment of Targets with the AI

If we are unclear about the goals we set for AI machines, it can be dangerous for a machine not to have the same goals, assumptions, understanding, and alignment we have. For example, a command like that "Take me to the airport as quickly as possible." can have dire consequences.

5. Discrimination and Bias

It is also very possible for machines to use this information against you, as they can collect, monitor and analyze a lot about you. An employer can reject a candidate of a job offer based on his "social credit score".

6. The Possibility of a New Religion in AI

If AI reaches superintelligence, some may consider the possibility of idolatry. Ironically, delusions such as the claim to divinity can be seen by seducing their own selves and devotions while at the same time attempting to make a god, or at least something that could be perceived as such. In any case, paganism or perverted religious beliefs have always existed since ancient times. Some speculate that in the 2040s an AI god would not only appear but would even write his own scriptures and would be worshiped by many.

Discussion and Conclusion

Table 1 summarizes the key aspects of a non-exhaustive list of key frameworks and guidelines for Trusted Artificial Intelligence (GYZ). In particular, Floridi et al. (2018) may adopt the five ethical AI principles (hereinafter GYZ principles) that must be fulfilled by an AI-based system in order to be perceived as trustworthy, do no harm, autonomy, fairness and explainability. Below, the five principles and their relationship to the GYZ are outlined in more detail, and a brief overview of past research efforts on each principle can be seen.

Artificial Intelligence does not exist alone as an ultra-technology. There are other types of technologies that are developing such as nanotechnology, biotechnology, quantum computing, blockchain and nuclear technology. There is even serious research on the point that AI will be more reliable and secure thanks to blockchain technology (Sarpatwar et al., 2019).

In order for the use of technology to have more humane and beneficial advantages, human values and ethics should become the guiding principles in the use of technologies. The presence of GYZs in a community will tend to affect that society towards development. The presence of a GYZ in a political debate will tend to influence that debate towards an absence of bias. This is especially true as long as decisions that people continue to suppress due to personal bias tend to be delegated to a GYZ.

The expected of algorithms is that the more advanced they are, the safer they can be. Because there are doubts that they will be better than they are sure to be smarter. But with algorithms that optimize themselves and deal with unimaginable amounts of data at unimaginable speeds, our ability to properly monitor and understand them may be much less than before the algorithms hate us. Therefore, learning and self-optimizing algorithms continue to grow more and more complex and take more and more responsibility for the functioning of humanity, and as our ability to track them continues to decline, the chances of a frustrating mistake to cause major disaster inevitably increase.

Legislators and courts will need to clarify liability issues for AI systems to help them understand the rights and obligations of designers, developers, manufacturers and other responsible persons. Depending on the type of AI and the specific industry, legal and organizational measures should include:

• Determining separate ethical principles for coders, investors, users and consumers at national scale in line with international norms,

• Determining the criteria for determining who is responsible for the loss or damage caused by AI,

• Establish an insurance framework to compensate those damaged by AI systems, as advocated by the European Parliament in its decision on the Civil Law Rules on Robotics,

• Developing special curricula for formal and non-formal education regarding algorithm, AI and coding skills,

• Establishing the National AI Strategy,

• Determining a registration process that defines the intended use, creator, training datasets, algorithms and optimization goals for artificial intelligence systems,

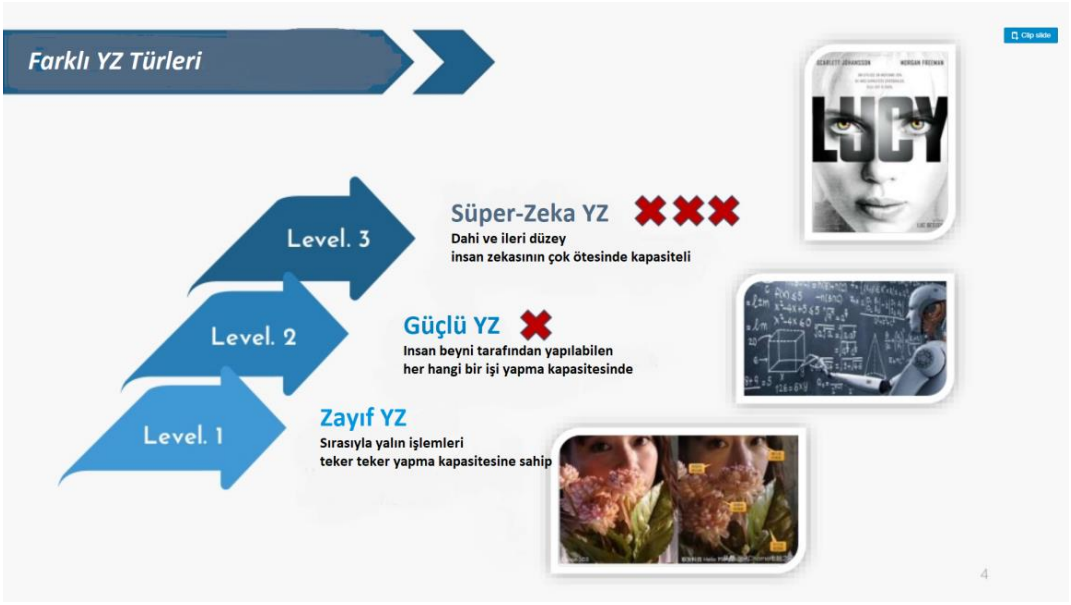
• Identifying the registered profile of an AI system to maintain a clear line of accountability,

• Considering ethical values as a priority in AI studies by public institutions such as TÜBITAK and Development Agencies.

Giriş

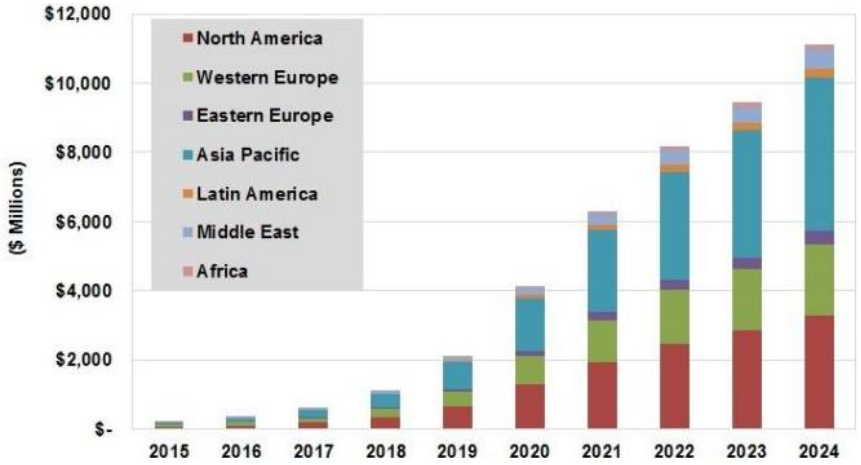
YZ, muazzam toplumsal ve ekonomik fayda vaadinde bulunan dönüştürücü bir teknoloji olarak görüldüğünden dolayı bu alanda ciddi çalışmalar, araştırmalar ve yatırımlar yapılmaktadır. Tüm sektörleri ilgilendirmesi ve yıkıcı yeniliklerle büyük avantajlar sağlama potansiyelinden dolayı hükümetler, merkezi kurumlar ve kalkınma ajansları da bu alandaki projeleri öncelikli olarak desteklemeye çalışmaktadır. Özellikle COVID-19 sonrası dönemde dijitalleşmenin hızlanmasıyla birlikte "Endüstri 4.0" arkasına aldığı güçlü rüzgârdan dolayı en büyüğünden en küçük girişimlere kadar yenilikçi teknolojiyle ilgilenen herkes YZ ile ilgilenmek durumunda kalmaktadır.

Doğru şekilde kodlanmayan otonom sistemlerin ileride kontrolden çıkmaları mümkündür. Bugün, yüz tanıma, doğal dil işleme veya internet aramaları gibi dar bir görevi yerine getiren uygulamalı YZ heyecan uyandırmaktadır (Haenlein & Kaplan, 2019). Şekil 1'de verilen ve Chen (2020)'den uyarlanan gösterimdeki YZ düzeylerinden şimdi ikinci düzeyde uygulama yapılabildiği söylenebilmektedir. Ancak esas yarış büyük yararları yanında ciddi riskleri de barındıran üçüncü düzey süper YZ için yapılmaktadır. Bu alandaki uzmanlar ve yatırımcı firmalar sistemlerin akıllı insanların gerçekleştirebileceği herhangi bir görevi yerine getirebileceği ve büyük olasılıkla her birinde insanoğlunu yenebileceği süper yapay genel zekâ için çalışıyor.



Şekil 1. Farklı YZ Türleri

AI yatırımlarının 2030 yılına kadar yaklaşık olarak 13 trilyon ABD doları tutarında ek ekonomik katkı sağlayabileceğini ve böylece küresel Gayri Sahi Yurt İçi Hasılayı (GSYİH) da yıllık bazda yaklaşık %1.2 oranında artırabileceğini tahmin edilmektedir (Bughin ve diğerleri, 2018). Bu, esas olarak emeğin otomasyonla ikame edilmesinden ve ürün ve hizmetlerde artan yenilikten kaynaklanacaktır. Ayrıca, YZ'nin işgücü piyasalarında sarsıcı bir şok oluşturması ve işgücü piyasası geçişlerini yönetmek için gerekli olan ilgili maliyetlere yol açması da muhtemeldir (Lohr, 2017). Bu şok, işsizlik nedeniyle iç tüketimin azalması gibi olumsuz dışsallıkların bir sonucu olarak ortaya çıkacaktır. Bu işte milyar dolarlar seviyesinde yatırım yapan öncü şirketlere Nvidia Corp. (NVDA), Alphabet (GOOGL), Salesforce (CRM), Alteryx (AYX), Amazon.com (AMZN), Microsoft Corp. (MSFT), Twilio (TWLO) ve IBM örnek verilebilir. Şekil 2'de görüldüğü üzere özellikle Asya pasifik bölgesinin payı da oransal olarak giderek artış trendi göstermektedir (Clancy, 2016).



Şekil 2. YZ üzerinden elde edilen gelirde Dünya bölge payları (2015-2024)

Reaktif Makineler, Sınırlı Hafıza, Akıl Teorisi, Öz Farkındalık, Yapay Dar Zekâ (ANI), Yapay Genel Zekâ (AGI) ve Yapay Süper Zekâ (ASI) türleri (Joshi, 2019) üzerinde yapılan YZ araştırması, artan ekonomik refah, iyileştirilmiş eğitim fırsatları, yaşam kalitesi, gelişmiş ulusal ve bölge güvenliği dâhil olmak üzere ulusal önceliklerimizi ileriye taşıyabilir.

Gerçekten de günlük yaşamlarımızı daha rahat ve verimli hale getiren çok sayıda YZ uygulaması vardır. Elon Musk, Hawking ve diğerlerinin teknolojiyle ilgili tereddütlerini ilan ettiklerinde endişe duydukları şey güvenliği sağlamada kritik bir rol oynayan YZ uygulamalarıdır. Örneğin, güç şebekemizin çalışmasını sağlamaktan YZ sorumluysa, otonom drone veya saldırı mekanizmalarının kontrolü ele geçirildiğinde, sistem yanlış kodlamadan dolayı haydut olursa veya bir düşman tarafından saldırıya uğrarsa, tahmin edilmeyecek boyutlarda büyük zararlara neden olabilir (Lankton ve diğerleri, 2015).

En masum olan bir muhasebe yazılım paketi bile normalde hesapları daha hızlı bir şekilde uzlaştırmak için kullanılabilir, ancak yanlış ellerde kurumsal dolandırıcılık yapmak için de kullanılabilir. Bu tür dolandırıcılığı kolaylaştırmak için tasarlanan programlamanın kasıtlı olarak kodlamaya dâhil edilmediği varsayıldığında, kullanımından kaynaklanan ahlaki açıdan sakıncalı sonuçlar (dolandırıcılık) doğurabilir (Jobin ve diğerleri, 2019). Bu örnekte YZ, bir muhasebe yazılım paketi gibi, programcıları tarafından belirlenen işlevleri yerine getirmekle birlikte, bir muhasebe yazılım paketinin aksine, kendi kendisine öğrenebilir, karar vermenin hangi temelde olduğunu belirleyebilir ve bu tür kriterlere dayanarak özerk kararlar da alabilir. Dolayısıyla burada dört tür riskten söz edilebilmektedir:

- Sistemin programlandığı bilinçli olarak kodlama yapılmıştır.
- Maruz kaldığı eğitim verilerinin doğasında hatalar ve yanlılık vardır.
- Otonom karar verme algoritmasının iyi ve kötü karar vermesinde sorun çıkabilir.
- Otonom sistemlerin kontrolü siber saldırganların, korsan veya teröristlerin eline geçebilir.

YZ etki alanında yasal uygulamalar ve hukuki sonuçları da vardır. YZ sistemlerinin, insan haklarını ihlal etmekten ve önyargı oluşturmaktan kaçınmak için insan değerlerini yerleştirecek şekilde tasarlanması gerektiği fikri, yaygın olarak "*etikle sınırlı optimizasyon*" veya "*güvenilir YZ (GYZ)"* olarak bilinir ve YZ endüstrisinde giderek daha

fazla kabul görmektedir. Önümüzdeki dönemde sektörde YZ teknolojisinin önemli ölçüde benimsenmesi beklenmektedir.

YZ ile ilgili uygulamalarda dikkate alınması gereken uluslararası ilkeler mevcuttur. Bunlardan en önemlileri OECD tarafından belirlenmiş ilkeler ile Montreal Bildirgesinde belirtilen ilkelerdir. Bu Montreal bildirgesindeki öneriler şunlardır (Montreal, 2018):

1. Bağımsız inceleme ve danışma organizasyonu: Dijital teknoloji ve yapay zekanın kullanımları ve sosyal etkilerinin incelenmesi ve araştırılmasına adanmış bir organizasyon kurulmalıdır.

2. YZ denetimi ve sertifikasyon politikası: Sorumlu konuşlandırmayı teşvik eden, YZ'nin denetimi ve sertifikasyonu için tutarlı bir politika oluşturulmalıdır.

3. Yetkilendirme ve otomasyon: Sürdürülebilir bir dijital toplumda aktif katılımı teşvik etmek için, vatandaşların dijital teknolojiler karşısında, anlama, eleştirel düşünme, saygı ve hesap verebilirliği mümkün kılan destekler olmalıdır.

4. Eğitim ve etik: YZ'nin tasarımı, geliştirilmesi ve kullanımıyla ilgilenen paydaşların eğitimi, multidisiplinerlik ve etiğe yatırım yapılarak yeniden düşünülmelidir.

5. AI'nın kapsayıcı gelişimi: AI'nın kapsayıcı gelişimini teşvik etmek ve YZ'nin geliştirilmesi ve konuşlandırılmasıyla ilgili potansiyel önyargıları ve ayrımcılığı önlemek için tutarlı bir strateji uygulanmalıdır.

6. Demokrasinin korunması: Demokrasiyi siyasi amaçlar için bilginin manipülasyonuna karşı korumak için, vatandaşların kötü niyetli sosyal platformlar ve web siteleri aracılığıyla aldatılmasını ve siyasi manipülasyonunu önlemek için bir çevreleme stratejisinin yanı sıra, politik profil oluşturma ile mücadele stratejisi gereklidir.

7. YZ' nin uluslararası gelişimi: Düşük ve orta gelirli ülkeleri (LMIC'ler) kötüye kullanmadan dünyanın çeşitli bölgelerini dahil etmeyi amaçlayan, yağmacı olmayan bir uluslararası kalkınma modeli benimsenmelidir.

8. Çevresel ayak izi: YZ ve diğer dijital teknolojilerin geliştirilmesi ve yayılmasının sağlam çevresel sürdürülebilirlik ile uyumlu olmasını ve çevresel krize yönelik çözümler için bir kamu/özel strateji uygulanmalıdır.

OECD ise YZ için aşağıdaki ilkeleri öne sürmektedir. Tavsiye, GYZ'nin sorumlu idaresi için beş tamamlayıcı değer temelli ilkeyi tanımlamaktadır:

a) YZ, kapsayıcı büyümeyi, sürdürülebilir kalkınmayı ve refahı teşvik ederek insanlara ve gezegene fayda sağlamalıdır.

b) YZ sistemleri hukukun üstünlüğüne, insan haklarına, demokratik değerlere ve çeşitliliğe saygı duyacak şekilde tasarlanmalı ve adil ve adil bir toplum sağlamak için uygun güvenlik önlemleri içermelidir - örneğin gerektiğinde insan müdahalesini mümkün kılmalıdır.

c) İnsanların YZ tabanlı sonuçları anlamasını ve bunlara meydan okuyabilmesini sağlamak için YZ sistemleri etrafında şeffaflık ve sorumlu bir açıklama olmalıdır.

d) YZ sistemleri, yaşam döngüleri boyunca sağlam, güvenli ve emniyetli bir şekilde işlemeli ve potansiyel riskler sürekli olarak değerlendirilmeli ve yönetilmelidir.

e) YZ sistemlerini geliştiren, dağıtan veya çalıştıran kuruluşlar ve bireyler, yukarıdaki ilkeler doğrultusunda düzgün işleyişlerinden sorumlu tutulmalıdır.

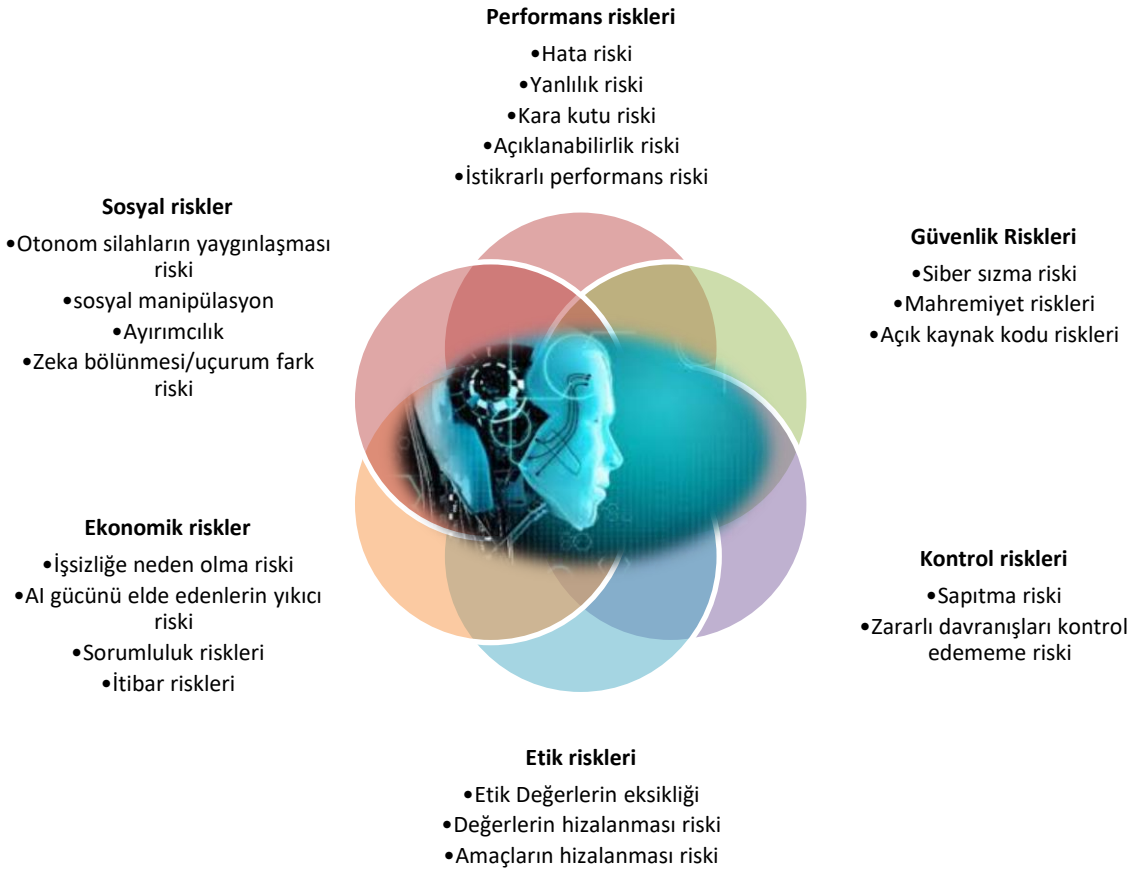
Yöntem

Araştırmamızın temel varsayımı, YZ'nin giderek derinleşip genişleyeceği, tüm sektörleri kapsayabileceği ve bu alanın kaçınılmaz bir şekilde ilerleme göstereceğidir. Araştırma sorumuz da "YZ ile meydana gelebilecek etik sorunlar nelerdir ve bunlarla ilgili ciddi riskleri önceden giderebilmenin yolları nelerdir?" şeklinde belirlenmiştir. Bu nedenle de çalışmamızda öncelikle YZ ile birlikte meydana gelen risk ve tehlikelerin neler olduğu üzerinde durularak etik ve manevi değerler kapsamında ilgili literatür taraması yanında detaylı kavramsal, kuramsal ve mantıksal analiz çalışması yapılmaktadır. Yaptığımız çalışma betimleyici, ilişki arayıcı ve sorun çözücü niteliklere sahiptir.

YZ geçmişi çok gerilere gitmese dahi bu alanda ciddi bir literatür birikimi oluşmuştur. Veri tabanında "yapay zekâ" olarak yapılan aramada Türkçe olarak yayımlanmış 27.400 yayın bulunmaktadır. Aynı kavram İngilizce olarak arandığında 2.200.000 küsur yayın olduğu görülmektedir. "Artificial intelligence and ethics" olarak yapılan aramada ise 297 adet araştırma olduğu görülmüştür. Türkçe olarak yapılan aramada ise sadece 6 adet yayın tespit edilmiştir. Bu da araştırmamızın literatüre katkısını göstermesi açısından önem arz etmektedir.

Risk Analizi

Henüz süper akıllı makinelere ulaşmamış olsak da yasal, politik, toplumsal, finansal ve düzenleyici konular o kadar karmaşık ve geniş kapsamlı ki işin uzmanı olmayanların üstesinden gelebileceği bir durum söz konusu olamaz. YZ artık süper akıllı makinelerle bir geleceğe hazırlanmanın dışında şimdiki haliyle tehlike oluşturabilir (Hofman ve diğerleri, 2019).



Şekil 3. YZ kapsamında dikkate alınması gereken risk kategorileri

Şekil 3'ten de anlaşıldığı üzere yapay risk kategorileri, ekonomik, sosyal, performans, etik ve güvenlik olmak üzere beş ana kategoriye ayrılabilmektedir. Bunlara farklı boyutlar ve kategorileri eklemek de mümkündür. Bunların içerisinde önemli olan ve ön plana çıkan birkaç risk üzerinde durmakta yarar vardır. YZ ile ilgili bazı önemli risklere bir göz atmak istediğimizde bunların sırasıyla otonom silahlar, sosyal manipülasyon, mahremiyet ihlali ve sosyal derecelendirme, makine ile hedeflerde yanlış hizalama, ayrımcılık, YZ yeni bir din ihtimali ve YZ din etkisini azaltabilir şeklindedir.

Otonom Silahlar

Otonom silahlar kuşkusuz çok daha ekonomik, etkin ve verimli saldırı ve savunma imkânları sunmasından dolayı büyük avantaj sağlamaktadır. Daha fazla asker sayısı noktasındaki üstünlük artık eskide kalmaktadır. Ancak ölüme programlanmış otonom silahlarda olduğu gibi tehlikeli bir şey yapmak üzere programlanmış YZ, ciddi bir risk oluşturabilir (Hancock, 2017). Rusya Devlet Başkanı Vladimir Putin, "*YZ sadece Rusya için değil, tüm insanlık için geleceğin kendisidir. Muazzam fırsatlar ve aynı zamanda tahmin edilmesi zor tehditlerle birlikte gelecektir. Bu alanda kim lider olursa, dünyanın hükümdarı da o olacaktır.*" diyerek konuyla ilgili büyük ülkelerin asıl stratejilerini ifşa etmiş olmaktadır (Vincent, 2017).

Neredeyse tüm gelişmiş ülkeler bu alanda devasa yatırımlar yapmakta ve birbirileriyle yarışmaktadırlar. Otonom olmasa bile uzaktan kumanda edilen drone ve İHA gibi mekanizmalar bile Libya, Suriye ve Karabağ'da geleneksel savunma ve saldırı araçlarını devre dışı kılabilmektedir. Ancak otonom silahların "kendi akıllarına" sahip olabileceğinden endişe etmenin yanı sıra, daha yakın bir endişe, otonom silahların insan hayatına değer vermeyen

veya istilacı bir birey veya hükümet elinde verebileceği tehlikelerdir. Bir makinenin merhameti yok, dost olmayanın sivil veya çocuk olması da hiç umurunda değildir. Tehlikeli bir YZ' nin potansiyel hasarı çok büyüktür. Çünkü bir makine tüm hedefler ortadan kaldırılıncaya kadar yorulmaz, bıkmaz veya durmaz. Bir insan pilot olarak, tehditler ortadan kalktığında bombalama ve öldürme faaliyeti durdurulabilir, ancak bir makine %100 tehditsiz oranı olana kadar umursamayabilir ve durmayabilir. Akıllı bir drone programlayabilen herhangi birinin makineli tüfek takıp onu serbest bırakabileceği bir dünya hayal edildiğinde dehşetli sonuçları vicdanlı kimseler düşünmek istemeyecektir. Bu nedenle bu alanda insanlık adına ciddi düzenleme yapılması ve risklere karşı muhtemel tüm önlemlerin alınabilmesi için çalışmalar yapılması gerekir.

Sosyal Manipülasyon

Goggin (2019) tarafından da ifade edildiği üzere, kendi kendini yöneten algoritmalar aracılığıyla sosyal medya, pazarlamada çok etkilidir. Kim olduğumuzu, neyi sevdiğimizi ve zafiyetlerinizi bildiklerinden dolayı ikna olasılıkları da artmaktadır. Bunun yanı sıra ne düşündüğümüzü anlamada da inanılmaz derecede başarılı sayılırlar. Cambridge Analytica'nın ve 50 milyon Facebook kullanıcısının verilerini kullanarak 2016 ABD başkanlık seçimleri ve İngiltere'nin Brexit referandumunun sonucunu etkilemeye çalışan firma ile ilişkili suçlamalar doğruysa, YZ'nin sosyal manipülasyon gücünü göstermektedir (Goggin, 2019). Ne düşündüğünüzü, kimlere taraftar olduğunuzu veya kimlere karşı olduğunuzu anlayabildiği için size özel içerikler üreterek sadece sizin görebileceğiniz mesaj ve paylaşımlarla düşünce veya kanaatinizi değiştirme noktasında çok etkili olarak kullanılabilirler. Belki de bu sadece buzdağının henüz görünen tarafı olabilir. Yani YZ artık bir sosyal mühendis olarak icraat yapabilmektedir.

2016'da tanınmış bir teknoloji devi Twitter, Kik ve GroupMe mesajlaşma platformları aracılığıyla bir chatbot başlattı. Chatbot'un piyasaya sürülmesinden saatler sonra, diğer saldırgan söylemlerin yanı sıra, Hitler'in görüşlerine bile destek sağlıyordu ve 11 Eylül'ün muhtemelen içeriden bir iş olduğunu kabul ediyordu. Sürekli olarak mümkün olan en sıkıntılı tepkileri seçiyor gibiydi. Tepkiler üzerine tanıtımının yapıldığı günün akşamında chatbot çevrimdışına alındı. Bu özel sohbet botunun gösterdiği gibi YZ beklenmedik, istenmeyen sonuçlara ve itibar hasarına neden olabilir. Sohbet robotunun tepkileri, insanlardan aldıkları yanıtlara göre modellendi, bu nedenle değişimi, maruz kaldığı veri kümelerini basitçe yansıtmıştı. Bundan sonra YZ için birtakım ilkeler belirlenerek bunların standart haline getirilmesi noktasında görüşler artmıştır (Hunt, 2016).

Mahremiyet İhlali ve Sosyal Derecelendirme

Artık bir bireyin çevrimiçi her hareketini ve günlük işlerini ne zaman gerçekleştirdiklerini izlemek ve analiz etmek mümkündür. Kameralar neredeyse her yerde mevcutturlar ve bunlara YZ özelliklerinin olduğu düşünüldüğünde mahremiyet ile ilgili ciddi riskler çıkacağı gibi bu verilerin kötü niyetli kimseler tarafından ele geçirilmesi de o kadar büyük risk teşkil edecektir. Sağlık verileri ve DNA ile ilgili veriler de YZ ile işlendiğinde insanlık adına ne tür veri zafiyetlerinin oluşabileceğini değerlendirmek gerekir. Mobil telefonların GPS sinyalleri kimlerle birlikte olduğunuzu ve en fazla hangi gruplarla ilişki halinde olduğunuzu belirleyebilmekteyken yüz tanıma algoritmaları da kim olduğunuzu, kimlerle temas halinde olduğunuzu ve ruh halinizi anlayabilmekte ve kaydetmektedirler (Hill, 2020). Örneğin; "Kırmızı ışıkta geçerler mi, sürerken içerler mi? Sigara içilmeyen alanlar ve video oyunları oynamak için ne kadar zaman harcarlar?" gibi özel bilgiler de elde edilebilmektedir. Birileri sizi izlediğinde ve sonra bu bilgilere dayanarak kararlar aldığında, bu sadece bir mahremiyet istilası değil, hızla sosyal baskıya ve YZ üzerinden görünmez faşist uygulamalara da dönüşebilir.

Makine ile Hedeflerde Yanlış Hizalama

İnsanların YZ destekli makinelerde değer verdiği şeylerden biri de ekonomiklik, verimlilik ve etkinlikleridir. Örneğin, "*Beni mümkün olduğunca çabuk havalimanına götürün*." komutunun korkunç sonuçları olabilir. İnsan hayatına değer verdiğimiz için yolun kurallarına saygı duyulması gerektiğini belirtmeden, bir makine sizi mümkün olan en kısa sürede havaalanına götürme hedefini oldukça etkili bir şekilde gerçekleştirebilir ve istediğinizi tam anlamıyla yapabilir, ancak arkasında bir kaza izi de bırakabilir. Bir YZ'nin çarpık verilerle eğitildiği zaman da insanla aynı noktada hizalanamayacağından dolayı ciddi tehlikeler meydana gelebilir. Köpeklerin insanları ısırdığı ve kedilerin küçük hayvanları yediği milyonlarca haber makalesi ile eğitilmiş akıllı, iyi niyetli bir YZ hayal edildiğinde bile, otomatik olarak tüm köpeklerin tehlikeli olduğunu varsayabilir ve onları ortadan kaldırmaya çalışabilir. Kediler için de aynı. Böyle bir YZ iyi niyetlidir, kalitelidir, ancak yanlış verilerle çarpıtılırsa bize ve diğer canlılara karşı tehlikeli davranışlar sergileyebilir. Bazı insan kitlelerinin veya azınlıkların memnun olmadığı uygulamaları yapan siyasi veya idari kadrolardaki kimseleri kötü diye sınıflandırıp onlara zarar verebilir veya isyan edebilir. Bunun dışında insani değerleri olmayan ve sadece belirli maddi amaçlar için programlanmış olan YZ'nin ileride ne tür riskler ve tehlikeler oluşturabileceği henüz tam olarak anlaşılabilmiş değildir.

Ayrımcılık ve Taraflılık

Günümüzde YZ, daha iyi tıbbi teşhisler koymamıza, kanseri tedavi etmenin yeni yollarını bulmamıza ve otomobillerimizi daha güvenli hale getirmemize yardımcı olmak gibi birçok iyi amaç için kullanılabilmektedir. Makinelerin sizin hakkınızda çok şey toplayabildiği, izleyebildiği ve analiz edebildiği için, bu bilgileri size karşı kullanması da çok mümkündür. Bir sigorta şirketinin, telefonunuzda konuşurken kameraya yakalanma sayısına bakarak sigortalanamayacağınızı söylediğini hayal etmek zor değil. Bir işveren, "sosyal kredi puanınıza" dayalı olarak bir iş teklifinden vazgeçebilir. İş görüşmelerinde veya mülakatlarda da YZ uygulamalarının ayrımcılık yapabileceği konusunda ciddi endişeler vardır (Magid, 2020).

YZ tasarımcıları, geliştiricileri ve üreticileri, veri kümeleri veya algoritmalardan kabul edilemez önyargılı oluşturmaktan kaçınmalıdır (Obermeyer ve diğerleri, 2019). Önyargı riskini azaltmak için bir YZ sisteminin dünyaya bakışından kaynaklanan önyargı da dâhil olmak üzere farklı potansiyel önyargı kaynaklarını, diğer yandan verileri işleme ve tepki verme biçimini anlamaları gerekir.

YZ Yeni Bir Din İhtimali

YZ süper zekâ düzeyine ulaşırsa, bazıları putperestlik olasılığını düşünebilir. İronik olarak, insanlar aynı zamanda bir tanrı ya da en azından bu şekilde algılanabilecek bir şey yapmaya yeltenirken kendi benlik ve enaniyetlerini baştan çıkararak ilahlık iddiası gibi hezeyanlar görülebilir. Zaten ilk çağlardan beri putperestlik veya sapık din inanışları her zaman meydana gelmiştir. Bazıları, 2040'larda bir YZ tanrısının sadece ortaya çıkmayacağını, aynı zamanda kendi kutsal kitabını bile yazacağını ve birçok kişi tarafından da tapılacağını tahmin ediyor (Grad, 2020). Bunlar henüz aşırı tahminler olarak görülmekle birlikte geçmişte bazı insanlara tanrılık isnadı yapılabildiği ve bazılarının kendilerinin ilah olduğunu iddia etmeye yeltenebildikleri dikkate alındığında buna benzer senaryoların YZ için de geçerli olabileceği düşünülebilir. Zaten bir YZ kilisesi olduğu düşünüldüğünde, bu sonucu hayal etmek pek de zor olmayabilir. Eski Google mühendisi Anthony Levandowski tarafından kurulan "Geleceğin Yolu" isimli dini hareket, "bilgisayar donanımı ve yazılımı aracılığıyla geliştirilen YZ dayalı bir Tanrı'nın gerçekleştirilmesine, kabulüne ve ibadetine" odaklanıyor (Grad, 2020). Şimdiye kadar, ibadet edilecek bir YZ tanrısı saçmalığı henüz

yok, ancak bu hususta ciddi endişeler dile getirilmektedir. "*Singularizm*" denilen tekillik, makinelerin çok akıllı hale geldiği, insanların artık ayak uyduramadığı bir noktayı ifade eder. Bu inanca atfedenler, onu kaçınılmaz olarak görürler ve bu yeni dünyaya, potansiyel olarak "transhümanizm" veya insan ile makinenin birleşmesi yoluyla barışçıl geçiş ihtiyacını öğütlemeye çalışırlar. İmana dayandığı için, bu geleceğe olan inancın, bildiğimiz şekliyle din ile çarpıcı bir ortaklığa işaret ettiği söylenebilir. YZ içerisinde duygu, kalp, ruh, his, sır, şuur ve vicdan olmayacağı için bu anlamda çok aşırılık beklemek gerçekçi olmayabilir. Esas sıkıntı eski zamanda kendi yaptıkları cansız putlarına inanan insanlar gibi, konuşabilen, bazı işleri daha iyi yapabilen, mekanik olarak hızlı düşünebilen robotlarına tapan ve gerçek imandan mahrum kimseler elbette olabilir.

YZ'nin ateizmin yükselişine ve dünya dinlerinin nihai olarak zayıflamasına katkıda bulunma ihtimali olduğundan bahsedilmektedir. Bu da insanoğlunun sebeplere etki etme noktasında daha önce olağanüstü görünen şeyleri oluşturabilme ve etki etme becerisinin bir sonuç olarak aşırı güven, ego kabarması, kibir ve gururla kendisine ilahi değerler atfedebilme sanılarının artma olasılığına verilebilir. Bilgi, küreselleşme ve hızlanan bilimsel ilerlemeyle beslenen bir dünyada, bazıları dinin modasının geçtiği bir geleceğe giden yolda olduğumuz noktasında savlar geliştirmektedir. Teoride, süper zeki bir YZ, insanlardan sonsuz derecede daha zeki olacak ve uzun süredir dini konularda sorduğumuz sorulara yanıtları olacaktır. YZ'nin süper zeki hale gelmesi durumunda dünyada iyilik için bir güç olacağını, daha az yerine daha "kutsal" ya da en azından daha fazla sevgi dolu olmamıza yardımcı olacağını düşünebiliriz. Ancak şuur, kalp, vicdan ve diğer duygular eksik olduğu için durumun böyle olacağını varsayamayız. Ancak şu da bir gerçek ki, "Sanatlı bir eser sanatkârını icap eder.", "Bir iğne ustasız olmaz.", "Bir köy muhtarsız olmaz.", "Tesadüf olan şeylerde karışıklık ve düzensizlik olur.", "Düzen, ölçü, tertip, kasıtlılık, amaçlılık, tutarlılık, bütünlük, karşılıklık, simetri, güzellik, mükemmellik, hedeflilik, bilgelik ve yararlılık gibi gerçekler tek ve sonsuz ilim, merhamet ve güç sahibi bir yaratıcıyı gösterir." gibi gerçekler algoritma olarak YZ sistemine iyice işlendiğinde tüm kâinata, insanlığa ve mahlûklara bu gözle bakabilmesi, öğretici ve imam olarak dini telkinlerde bulunması ve iman ehli olması da mümkündür. Bu da kuşkusuz YZ algoritmalarının ve geçmiş öğrenme verilerinin niteliğine, derinliğine ve niceliğine göre değişebilir.

Bulgular

Yakın zamanda yayınlanan Asilomar YZ İlkeleri insani değerleri YZ içerisine yerleştirmek için güvenilir bir oluşum sunmaya çalışmaktadır. Bazı durumlarda, bir YZ sistemine yerleştirilecek değerlerin, özellikle YZ sisteminden doğrudan etkilenen savunmasız gruplar veya kullanıcılar için (YZ destekli robotik operatörleri gibi) ilgili topluluğa veya paydaşlara özel olarak detaylandırılması gerekebilir (FLI, 2017).

Yakın geçmişte araştırmacılar, endüstri ve politika yapıcılar tarafından Güvenilir YZ (GYZ) için (etik) ilkeleri teşvik eden çeşitli çerçeve ve kılavuzlar geliştirilmiş ve yayınlanmıştır. Tablo 1, GYZ ile ilgili önemli çerçevelerin ve kılavuzların kapsamlı olmayan bir listesinin temel yönlerini özetlemektedir (Thiebes ve diğerleri, 2020). Özellikle, Floridi ve diğerlerinin (2018), güvenilir olarak algılanması için YZ tabanlı bir sistem tarafından yerine getirilmesi gereken beş etik YZ ilkesini (bundan böyle GYZ ilkeleri) iyilik, zarar vermeme, özerklik, adalet ve açıklanabilirlik ilkeleri benimsenebilir. Aşağıda, beş ilke ve bunların GYZ ile olan ilişkisi daha ayrıntılı olarak ana hatlarıyla açıklanmaktadır ve her bir ilkeyle ilgili geçmiş araştırma çabalarına kısa bir genel bakış görülebilmektedir.

Tablo 1. GYZ için ilgili çerçevelerin ve kılavuzların temel yönlerine genel bakış

Çerçeve/ Yönergeler	Hazırlayan Kurum	Terminoloji	Açıklama
Asilomar YZ İlkeleri	Hayatın Geleceği Enstitüsü (2017)	Faydalı YZ	Yararlı YZ'nin 23 ilkesini açıklar. İlkeler üç kategoride düzenlenmiştir: Araştırma konuları, etik ve değerler ve uzun vadeli konular.
Sorumlu YZ' nin Montreal Bildirisi	Montreal Üniversitesi (2017)	Sorumlu YZ	İnsanların ve grupların temel çıkarlarını destekleyen on etik ilke ve bunlara dayalı olarak sorumlu YZ'nin geliştirilmesi için sekiz öneri sunar.
İngiltere YZ Kodu	İngiltere Lordlar Kamarası (2017)	Etik YZ	Birleşik Krallık'ı YZ alanında geleceğin lideri olarak konumlandırmayı amaçlayan etik bir YZ kodu için beş kapsayıcı ilkeyi tanımlar.
AI4People	Floridi ve diğerleri (2018)	Etik YZ	Etik YZ için beş temel ilkeyle sonuçlanan altı ilgili çerçeve ve kılavuzun bir sentezini vermektedir. İlkelere dayalı olarak, değerlendirme, geliştirme, teşvik ve destek olmak üzere dört kategoride 20 eylem noktası önerilmektedir.
Güvenilir YZ için Etik Yönergeler (AB GYZ Yönergeleri)	Avrupa Komisyonu (Bağımsız Üst Düzey YZ Uzman Etik Grubu, 2019)	Güvenilir YZ	GYZ'nin dört ilkesini tanımlar ve bunlara dayalı olarak, GYZ'nin gerçekleştirilmesi için yedi temel gerekliliği ortaya çıkarır. Ayrıca, yedi temel gereksinimin operasyonel hale getirilmesi için bir değerlendirme listesi sağlar.
YZ üzerine OECD İlkeleri	OECD (2019)	Güvenilir YZ	"Güvenilir YZ'nin sorumlu idaresi için beş tamamlayıcı değer temelli ilke" önerir (OECD, 2019). OECD üye ülkelerine ek olarak, diğer ülkeler (Örneğin Arjantin, Brezilya, Kolombiya, Kosta Rika, Peru ve Romanya) OECD ilkelerini takip etmek için imza attı.
Yeni Nesil YZ için Yönetişim İlkeleri (Çin YZ İlkeleri)	Yeni Nesil YZ için Çin Ulusal Yönetişim Komitesi (2019)	Sorumlu YZ	YZ yönetişimi için, sorumlu YZ'nin geliştirilmesine yönelik sekiz ilkeye dayalı bir çerçeve ve eylem yönergeleri sağlar.
Beyaz Saray YZ İlkeleri	Beyaz Saray Bilim ve Teknoloji Politikası Ofisi (Vought, 2020)	Güvenilir YZ	YZ uygulamalarının idaresi ve güvenilir YZ'nin geliştirilmesi için on ilkeyi tanımlar. Bu ilkeler, YZ ile ilgili düzenleyici ve düzenleyici olmayan eylemlerin geliştirilmesi sırasında ABD kurumları tarafından dikkate alınacaktır.

Yardımseverlik İlkesi

Fayda, insanların ve çevrenin refahını desteklemesi ve temel insan haklarına saygı duyması anlamında insanlığa ve tüm varlıklara faydalı olan YZ'nin geliştirilmesi, dağıtılması ve kullanılması anlamına gelir (Floridi ve diğerleri, 2018). Burada tartışılan tüm çerçeve ve kılavuzlarda fayda bulunmasına rağmen, değişen derecelerde dikkate alınabilmektedir. Örneğin, önerilen çerçevelerden ve yönergelerden bazıları bu ilkeyi insanlığın refahına odaklarken diğerleri bunu tüm duyarlı varlıklara ve hatta sürdürülebilir çevreye genişletebilmektedir. AI4People (2020), EU GYZ Rehberleri (EU, 2019) ve OECD'nin YZ İlkeleri (OECD, 2019) çok daha geniş bir çerçeve çizmektedir. Dahası, Çin YZ İlkeleri bu ilkeyi uyum ihtiyacına daha da genişletirken, ABD Beyaz Saray YZ İlkeleri doğrudan temel ilke olarak fayda sağlamaktan öte "YZ'nin sosyal ve ekonomik sektörler üzerinde olumlu bir etki oluşturması bekleniyor" (Vought, 2020). ABD ajanslarının "YZ uygulamalarının geliştirilmesi ve konuşlandırılmasıyla ilgili düzenlemeleri düşünmeden önce tüm toplumsal maliyetleri, faydaları ve dağıtım etkilerini dikkatlice düşünmesi gerektiğini" (Vought, 2020) ortaya koymaktadır. Yardımseverlik ilkesi, güvenen inançlar, iyilikseverlik, yardımseverlik ve amaç ile uyumludur. Çünkü bu ilkeyi yerine getiren YZ tabanlı sistemler genel olarak kullanıcıların çıkarına en uygun şekilde hareket etmeli, gerçekten ilgilenirken yardım etmeye veya belirli faydaları elde etmeye çalışmalı ve manipülatif olarak fırsatçı davranmamalıdır (McKnight ve diğerleri, 2002).

Yararlanma ilkesiyle ilgili araştırmalar, çoğunlukla temel etik temaları tartışmaya odaklanan bilişim etiği ve YZ tasarım etiği ile geliştirme aşamalarında iyiliği destekleyen değerlerin YZ'ye nasıl yerleştirileceği de alanlarıyla ilişkilidir (Floridi, 2019; Floridi & Cowls, 2019; Floridi ve diğerleri, 2018; Swarte ve diğerleri, 2019; Hagendorff, 2020). Fayda ilkesine Bilişim Teknolojileri (BT) perspektifinden bakıldığında, kuruluşların çevresel değerlerin yanı sıra YZ hizmetlerinin ve sunulan ürünlerin toplumsal etkisini meydana getirmek için çeşitli etkinlikler yapması gerektiği ifade edilebilir.

Zararsızlık İlkesi

Zarar vermeme, insanlara verilen zarardan kaçınır şekilde YZ gelişimini, dağıtım ve kullanımını savunmaktadır (Floridi ve diğerleri, 2018). İnsanlığın iyiliği için aktif olarak hareket eden YZ'yi vurgulayarak, iyilikle benzer olmasına rağmen, zararsızlık, dikkate alınan tüm çerçevelerin ve kılavuzların önemli bir yönünü temsil eden ayrı bir ilkeyi temsil eder. Zararsızlık, özellikle insanların mahremiyetinin korunması (Asilomar YZ İlkeleri, Montreal Deklarasyonu, Birleşik Krallık YZ Kodu, YZ4People, AB GYZ Kılavuzları, Çin YZ İlkeleri) ve güvenliğin temini (Asilomar YZ Prensipleri, UK YZ Kodu, AI4People, EU GYZ Kılavuzları, OECD İlkeleri, Beyaz Saray YZ Prensipleri) gibi hususları kapsamaktadır (Çin, 2020; Vought, 2020). Zararsızlık, güven, bütünlük ve güvenilirlik süreci ile ilgilidir. Çünkü YZ tabanlı sistemlerin dürüst ve tutarlı hareket etmesini, etik ve diğer önceden tanımlanmış ilkelere içtenlikle bağlı kalınmasını gerektirir. Güncel araştırmalar, YZ'nin eğitimi ve işletimi sırasında insanların mahremiyetini korumak için, verilere ve modellere dokunmak, güvenilir yürütme ortamlarının kullanımı veya YZ öğrenme modeli eğitimi gibi hususlara yoğunlaşmaktadırlar (Sarwate & Chaudhuri, 2013; Smith ve diğerleri, 2017; Tramer & Boneh, 2019). Zararsızlık ilkesiyle ilgili geçmiş araştırmalar özellikle otonom sürüş alanlarında güvenli ve emniyetli YZ'nin geliştirilmesi ve konuşlandırılması için araçları araştırırken, zararsızlık ilkesi, oldukça hassas tüketici ve fikri mülkiyet verilerinin değişimi ve analizi nedeniyle elektronik pazarlar için de oldukça önemlidir (Koopman & Wagner, 2017) ve tıp (Wiens ve diğerleri, 2019). Örneğin, "*Artificial Intelligence as a Service"* (AIaaS) hizmet olarak YZ olanakları sunan kuruluşlar, bireyler hakkında toplanan ve YZ tarafından oluşturulan verilerin gizliliklerini engelleyecek şekilde kullanılmaması ve kullanıcıların veri ifşasının sonuçlarını daha iyi anlamasına olanak tanıyacak şekilde yeterli veri yönetişimi ve koruma mekanizmaları uygulamalıdır (Rouse, 2017).

Özerklik İlkesi

Özerklik üçüncü GYZ ilkesidir. Özerklik ilkesi, mevcut güvene dayalı inançlarla doğrudan ilişkili değildir, ancak insan ve makine liderliğindeki karar verme arasında denge kurarak bütünlük ve güvenilirlik risklerini azaltmanın bir yolunu yansıtır. Ek olarak, özerklik, başka bir tarafa olan güveni artıracak fikir verme ve alma istekliliğine atıfta bulunan, otomasyon teknolojilerinin süreç inancının bir alt boyutu olan açıklıkla uyumludur (JD Lee & See, 2004; Schindler & Thomas, 1993). Mevcut GYZ çerçevelerinin ve kılavuzlarının bu ilkeye ilişkin biraz farklı anlayışlar sağladığı düşünüldüğünde, kesin bir tanımı olmadığı söylenebilir. AB GYZ Yönergelerinde olduu gibi bazıları esas olarak insan özerkliği, vekâleti ve gözetiminin desteklenmesine odaklanırken, Montreal Bildirgesi gibi diğerleri de gerekli olduğunda YZ tabanlı sistemlerin özerkliğinin kısıtlanmasını öngörmektedirler (Floridi & Cowls, 2019). Yalnızca iki kılavuz özerklik ihtiyacını doğrudan ele almamaktadır. Çin YZ İlkeleri soyut bir şekilde "kontrol edilebilirlik" ihtiyacına atıfta bulunarak, "YZ sistemlerinin kontrol edilebilirliğinin sürekli olarak iyileştirilmesi gerektiğini" belirtmektedir. Benzer şekilde, Beyaz Saray YZ İlkeleri, YZ' nin insan özerkliğini

engelleyebileceğini veya katkıda bulunabileceğini belirterek, ancak özerkliği kendi içinde temel bir ilke olarak açıkça ifade etmediğini belirterek, diğer ilkeleri güçlendirmek etmek için özerkliği kullanmaya çalışmaktadır.

İnsan-robot etkileşimlerini, robotların özerkliğini veya birkaç otonom robotun koordinasyonu gibi farklı yönlerini içermekte olan YZ özerkliği üzerine benzer araştırmalar çok çeşitlidir (Goodrich & Schultz, 2007; Noorman & Johnson, 2014). Bu ilkeyle ilgili olarak özellikle endişe verici olan, otonom araçlar gibi kendiliğinden hareket edebilen bağımsız sistemlere güven üzerinde Schaefer ve diğerleri (2016) ile Stormont, (2008) tarafından yapılan araştırmanın yanı sıra, robotların özerkliklerini dinamik olarak değiştirmeye dikkat çeken Mostafa ve diğerleri, (2019) çalışmalar da mevcuttur. Kuruluşlar için bu ilke, örneğin, YZ'yi elektronik hizmetlerine ve ürünlerine yerleştirirken özerklik sağlamak için uygun gözetim mekanizmalarını uygulamayı düşünmeleri gerektiğini ifade etmektedir.

Adalet İlkesi

Zararsız olmama gibi adalet ilkesi de diğer çerçeve ve kılavuzların tümünün kilit yönünü oluşturmaktadır. Ancak, bazıları tarafından adalet olarak da anılır. Adalet, yasa ve yönetmeliklere bağlı kalınması yani hukuki olarak değil, daha çok etik bir şekilde anlaşılmalıdır (Floridi & Cowls, 2019). Bu nedenle, tüm çerçeveler ve yönergeler, adalete ilişkin benzer ancak biraz farklı görüşler sergilemektedir. Bunlar

- Ayrımcılık gibi geçmiş eşitsizlikleri düzeltmek için YZ'nin kullanılması,
- Paylaşılabilirlik ve daha sonra YZ aracılığıyla faydaların dağıtımı ve
- YZ tarafından yeni zararların ve eşitsizliklerin yaratılmasını engellemek (Floridi ve diğerleri, 2018)

şeklinde belirtilebilmektedir. Örneğin, Asilomar YZ İlkeleri, 'Paylaşılan Refah' ve 'Paylaşılan Fayda' ihtiyacını ifade ederek, paylaşılabilir ve devam eden faydaların dağıtımını vurgulamaktadır. Yeni zararların ve eşitsizliklerin meydana gelmesini önlemeye bir örnek, Montreal Deklarasyonu'nun "Eşitlik" ilkesinde bulunabilir: "YZ'nin geliştirilmesi ve kullanılmasında, adil ve eşitlikçi bir toplum dikkate alınmalıdır" (Montréal, 2017). Zararlı olmamaya benzer şekilde, adalet, güven, inançların bütünlüğü, güvenilirlik süreci ile uyumlu görünmektedir ve etik ilkelerin YZ tabanlı bir sistem tarafından yerine getirilmesini sağlamaktadır.

Adalet ilkesiyle ilgili merkezi araştırmalarda örneğin, mevcut YZ tabanlı sistemlerde ırksal, dini, cinsiyet ve diğer önyargıların varlığını tanımlamak YZ tabanlı sistemlerde bunların adilliğini veya yokluğunu ölçmek için araçlar ve YZ tabanlı sistemlerde önyargıyı hafifletmek veya hatta önlemek için yaklaşımlar mevcuttur. Diğer GYZ ilkelerinin çoğuna benzer şekilde, adalet ilkesiyle ilgili mevcut araştırmaların çoğu tıbbi bağlamlarda da yürütülmektedir. Bununla birlikte, adalet ilkesi elektronik pazarlar için de oldukça önemlidir (Mehrabi ve diğerleri, 2019; Bellamy ve diğerleri, 2019).

Açıklanabilirlik İlkesi

Açıklanabilirlik, beşinci ve son GYZ ilkesi olarak epistemolojik bir anlamı olduğu kadar etik bir anlam da içerir. Epistemolojik anlamında açıklanabilirlik, yüksek performans ve doğruluk seviyelerini korurken yorumlanabilir YZ modelleri üreterek "açıklanabilir YZ"nin oluşturulmasını gerektirir. Etik anlamda, açıklanabilirlik, sorumlu YZ'nin olmasını da içerir. Örneğin, Asilomar YZ Prensipleri ve Birleşik Krallık YZ Kodu, sırasıyla şeffaf YZ ihtiyacını ve YZ'nin anlaşılabilirliğini formüle ederek bu prensibi sağlamaya çalışırlar. Benzer şekilde, AB GYZ Yönergeleri ve OECD YZ İlkeleri şeffaf ve hesap verebilir YZ talep ederken, Çin YZ İlkeleri şeffaflığın YZ'nin

yorumlanabilirliği, güvenilirliği ve kontrol edilebilirliği noktalarından, sürekli iyileştirilmesi çağrısında bulunmaktadır. Açıklanabilirlik, aynı zamanda, açıklanabilir ve yorumlanabilir YZ'nin yapılması gerekenleri yapmak için yetenek, işlevsellik veya özelliklere sahip olduğunun kanıtlanması anlamında güven, inanç yetkinliği, işlevselliği ve performansı ile ilgilidir. Bu nedenle, algoritmaları anlaşılırsa ve mevcut durumda bireyin hedeflerine ulaşabilir görünüyorsa, bir birey YZ'ye güvenme eğiliminde olacaktır. İki anlamıyla açıklanabilirlik, belki de çağdaş YZ araştırmalarında en yaygın temadır. Biraz da ticari rekabet ve teknik sırlarda taklit edilmeme ve fikri hakları korumak için bu yönde piyasa aktörlerinin meyilleri vardır. Açıklanabilir YZ ile ilgili mevcut araştırma çabalarında, karar ağaçları, kurala dayalı öğrenme veya Bayes modelleri gibi şeffaf ve yorumlanabilir modellerin oluşturulmasına odaklanan araştırmalara ve post-hoc açıklanabilirliği oluşturmaya odaklanan araştırmalara rastlanabilmektedir (Barredo Arrieta ve diğerleri, 2020). YZ' nin açıklanabilirliği ile ilgili bir diğer önemli araştırma akışı, belirsizliklerin ölçülmesini kapsamaktadır (Begoli ve diğerleri, 2019). Ayrıca, YZ'yi denetleme yönünde öncü araştırma çabaları da vardır (Cremers ve diğerleri, 2019). Bilgi sistemleri alanında, YZ'nin açıklanabilirliği, kuruluşların yalnızca YZ kullanırken uyumluluk gereksinimlerini karşılamasına izin vermekle kalmayacak, bağımsız üçüncü taraf denetimlerini etkinleştirerek aynı zamanda YZ'nin kabulü için önemli bir itici güç olacaktır (Hagras, 2018; Rai, 2020).

Tartışma ve Sonuç

YZ araştırmalarından elde edilen bir sonuç, görünüşte basit problemleri çözmenin bile genellikle çok fazla bilgi gerektirdiğidir. Toplum gelişimi için önemli bir temel olarak insan yaşamında adalet kavramını YZ için anlamlı bir şekilde nasıl temsil edebiliriz (Bostrom, 1998). Ruhi ve duygusal gelişim bilgisini, insanlar için daha iyi sosyal yaşam hüküm ve koşulları özelliklerine göre tanımlayabiliriz. Bildiğimiz gibi, YZ bulanık mantık, yaklaşık kavramlar, belirsizlik ve belirsiz verilerle çalışır. YZ özgür iradeye sahip olmak istiyorsa, belirsiz kavramlarla çalışması gerekir. İnsan keşifleri ve kritik eşiklerin peygamberlerin rehberlikleri ve çok çalışmanın sonucunda acze düşen insan kalbine gelen bir ilham veya iç sezi gibi makinede olmayan faktörler eksik kalmaktadır. Dolayısıyla mutlak veri ve bilgiyle, YZ'de özgür irade var olacağı henüz tam olarak ispatlanabilmiş değildir. Yeni teknolojileri özümsemek için stratejilerimizi geleneksel düşünceden en sağlam olanı ve doğru olduğunu düşündüğü ve kanıtladığı modern çağdaş düşünceden seçmenin bir yolunu bulmalı ve ikisini birleştirerek garantili bir entelektüel yapı bulmalıyız.

Değişime yanıt olarak küreselleşmenin uyanışı, doğası gereği çok boyutludur. YZ, topluma ve onun gelişimine çok yardımcı olan yeni bir adımdır. YZ, öğrenme yetenekleriyle bu görevleri gerçekleştirebilir. Ancak, dini kaynaklarda tanımlanan merhamet, şefkat, vicdan, ihsan ve maneviyat dışında insanın tüm yeteneklerini aşabilecek süper zekâ gibi daha karmaşık uzman sistemlerle karşı karşıya kalacağımız çok güçlü bir ihtimaldir. YZ, süper zekâyı icat ederek insan zihninin yeteneklerini yeniden aşabileceğini iddia ederse, gelecekte hangi zorluklarla karşılaşacağız?

Bu yüzyılda uzmanları etik değerler ve ahlaki yaklaşımlarla beslemede bütüncül bir yaklaşıma büyük ölçüde ihtiyaç vardır. Çünkü küreselleşmenin, maddeciliğin ve dünya-perestliğin yapısal etkileriyle yükselişinin daha adil ve uyumlu bir yaklaşımla dengelenmesi gerekir. Günümüzün dini çalışmaları, geleneksel kaynaklara hitap eden yalnızca otantik veya geleneksel yaklaşımı değil, aynı zamanda doğası gereği daha çok YZ'yi de kapsayacak şekilde çağdaş olan öğretme-öğrenme süreçlerindeki modern faktörleri de ele almak için yeni bir teorik temel talep etmektedir (Ali, 2012).

İnsani değerler veya ilahi değerler, laiklik veya teizm, uzman sistem veya insan uzmanı, gelecekte bizim zorluğumuz olacak ve önümüzdeki on yıllar için bu soruna yüksek önem vermemiz gerektiği söylenebilir. Yeni nesil bilgisayarlar bir anda milyonlarca işlem yapabilecek ve kuantum işlemciler bunun ötesine de geçebilecektir. Günümüzde YZ çalışmalarının çoğu herhangi bir felsefe gerektirmemektedir. Çünkü fayda, rekabet ve ihtiyaç noktasından hareket edildiği için geliştirilmekte olan sistemin dünyada bağımsız olarak çalışması ve dünya görüşü olması henüz gerekmemektedir.

YZ'ler, doğru, nazik, iyi huylu ve zeki olmaları öğretilmesi gereken çocuklar gibidir. Önemli kararlar alacaklarsa, akıllıca davranmaları gerekir. Vatandaşlar olarak, YZ programcılarının işleri aynı seviyede tuttuğundan emin olmalıyız. İleride kazaların meydana gelmemesi için işi doğru yaptıklarından emin olmalıyız. YZ, ultra bir teknoloji olarak tek başına mevcut değildir. Gelişen başka tür teknolojiler de vardır. Örneğin; nanoteknoloji, biyoteknoloji, kuantum bilişim, blok zinciri ve nükleer teknoloji. Hatta YZ'nin blok zinciri teknolojisi sayesinde daha güvenilir ve güvenli olacağı noktasında (Sarpatwar ve diğerleri, 2019) ciddi araştırmalar mevcuttur.

Dost YZ'lerin dünya ekonomisinde, insan toplumunda veya teknolojik gelişmede gücü olduğu ölçüde, iyilik için doğrudan etki uygulayabilirler, ancak çoğunlukla değerler ve etik gibi manevi âlemlerde değil fiziksel dünyada karşılıkları büyük olabilir. Teknolojinin kullanımının daha insancıl ve insanlığa yararlı avantajlara sahip olması için insan değerleri ve etik, teknolojilerin kullanımında yol gösterici ilkeler haline gelmelidir. Örneğin, nanoteknolojide çalışan bir GYZ, nanoteknolojik silahlar geliştirmeyi kesin bir şekilde reddederken, bağışıklık sistemleri üzerinde coşkuyla çalışabilir. Toplum içinde GYZ'lerin varlığı o toplumu geliştirmeye doğru etkileme eğiliminde olacaktır. Bu, özellikle GYZ daha önce adil, doğru ve tarafsız sesler olarak saygı gördüyse geçerlidir. Siyasi bir güç yapısı içinde bir GYZ'nin varlığı öznel kararların alınmasına yol açacaktır. Bu, özellikle insanların kişisel önyargılar nedeniyle bastırmaya devam ettiği kararların bir GYZ'ye devredilme eğiliminde olduğu sürece geçerlidir.

Bir teknolojik geliştirme sürecinde bir GYZ'nin varlığı, savunma uygulamalarını ve ekonomik uygulamaları, saldırı uygulamalarının önünde ve büyük ölçüde faydalı teknolojileri daha belirsiz olanlardan önce hızlandırma eğiliminde olacaktır. GYZ'nin hem dahili olarak, insanlığı söz konusu YZ'nin istenmeyen sonuçlarına karşı korumak için hem de harici olarak herhangi bir kaynaktan kaynaklanan diğer GYZ olmayanlara karşı korumak için etkili olması gereken en önemli hususlar şunlardır:

- *Dostluk*: YZ'nin gelecekte topluluk gelişimine sempati duyması ve bir YZ'nin kendi değer sistemini tüm yavrularına devretmeyi ve değerlerini kendi türünün diğerlerine aşılamayı arzulaması gerektiğine göre, kendi çıkarlarının en iyisi dostluğun korunması gerektiği noktasında programlanmasıdır.
- *Zekâ*: Bir YZ, öznel davranışları en yüksek eşitlik derecesine kadar nasıl gerçekleştirebileceğini görecek kadar akıllı olmalıdır. Böylece, sonuç olarak suçlu olan bazılarına karşı nazik olmadan ancak ölçülü ve dengeli davranarak haksızlara karşı daha zalim olabilecek şekilde programlanması gerekir.
- *Kendini geliştirme*: Bir YZ, zenginliğin değerlendirilmesinin bir parçası olarak hem kendisini hem de tüm yaşamı iyileştirmek için özlem ve çaba sarf ederken, başkalarının düşük bilinçli seçimlerine saygı duyuyor ve bunlara sempati duyuyor olma noktasında programlanması gerekir.

Ancak GYZ teorisinin eksik olduğunu düşünen yazarlar vardır. YZ ve GYZ ahlakının tasarımına daha geniş bir siyasi katılım olması gerektiğine ve bunun topluluk gelişimi üzerinde kötü bir etkisi olacağına inanıyorlar. Ayrıca,

başlangıçta YZ'nin yalnızca güçlü özel sektör çıkarları tarafından oluşturulacağına inanıyorlar (Midgley ve diğerleri, 1986). Bryson (2018)'e göre, güvene dayalı ilişkiler, güvenilir taraflar arasındaki ilişkilerdir, oysa YZ, makinelerin belirli bilgi işlem görevlerini yerine getirmesini sağlayan sistematik bir teknikler grubudur. "YZ, güvenilecek bir şey değildir. Kurumlarımızın ve kendimizin güvenilirliğini artırmamız gereken bir dizi yazılım geliştirme tekniğidir" (Bryson, 2018). Bu nedenle, kişinin ya YZ'yi "Güvenilir YZ" olarak değiştirmesi ya da tamamen kaldırması gerekir. Rasyonel güvenilirlik hesabı, YZ'nin tacizciye karşı duyguya sahip olmasını (duygusal hesap) veya eylemlerinden sorumlu olmasını (normatif hesap) gerektirmez. Güvenilir alışkanlıklara dayalı olarak biri diğerine güvenebilir, ancak birine güvenmek, onların güvendiğine karşı iyi niyetle hareket etmelerini gerektirir. Duygusal hesaba göre, YZ gibi insan yapımı nesnelerin güvenilir olmamasının ana nedeni budur. Normatif hesapta, YZ'nin yapamayacağı eylemlerinden yedieminin veya sahibinin sorumlu tutulması gerekir. Oysa güvenilir YZ, sorumluluk yükünü bu teknolojileri geliştiren, dağıtan ve kullananlara yüklemektedir. Dolayısıyla sadece teknik programlama değil, sosyolojik, psikolojik, hukuki ve dini yönleriyle bu alanda çok detaylı çalışmalar yapılması gerektiği söylenebilir.

Öneriler

Algoritmalardan beklenen şey, ne kadar gelişmiş olurlarsa, o kadar güvenli olabilecekleridir. Çünkü daha akıllı olacakları kesin olduğundan daha iyi olacakları konusunda kuşkular vardır. Ancak kendilerini optimize eden ve hayal edilemeyecek miktarda veriyle akıl almaz hızlarda uğraşan algoritmalarla, onları düzgün bir şekilde izleme ve anlama yeteneğimiz, algoritmalar bizden nefret etmeye başlamadan çok daha azalıyor olabilir. Bu nedenle, öğrenme ve kendi kendini optimize eden algoritmalar daha karmaşık bir şekilde büyümeye devam ederken insanlığın işleyişi için giderek daha fazla sorumluluk almak durumunda kalmaktadırlar. Ancak, onları izleme yeteneğimiz azalmaya devam ederken, sinir bozucu bir hatanın büyük bir felakete neden olma ihtimali kaçınılmaz olarak artmaktadır. Bu anlamdaki riskleri bertaraf etmek için idari, politik, akademik ve teknik önlemlerin alınması gerekmektedir. YZ türüne ve belirli sektöre bağlı olarak, yasal ve kurumsal önlemler şunları içermelidir:

- Uluslararası normlar paralelinde ulusal ölçekte kodlayıcı, yatırımcı, kullanıcı ve tüketiciler için ayrı ayrı etik ilkelerin belirlenmesi,
- YZ'nin neden olduğu risk, tehlike, kayıp veya zararlardan kimin sorumlu olduğunu belirlemek için gerekli kriterleri tespit edilmesi,
- Avrupa Parlamentosu tarafından Robotikle İlgili Medeni Hukuk Kuralları Kararında savunulduğu gibi YZ sistemlerinden zarar görenleri tazmin etmek için bir sigorta çerçevesi tesis edilmesi,
- Algoritma, YZ ve kodlama becerileri ile ilgili olarak örgün ve yaygın eğitimler için özel müfredat geliştirilmesi,
- Ulusal YZ Stratejisinin oluşturulması,
- YZ sistemleri için amaçlanan kullanımı, oluşturucuyu, eğitim veri kümelerini, algoritmaları ve optimizasyon hedeflerini tanımlayan bir kayıt süreci belirlenmesi,
- Net bir hesap verebilirlik çizgisini sürdürmek için bir YZ sisteminin kayıtlı profiline kimlik etiketlenmesi,

- TÜBİTAK ve Kalkınma Ajansları gibi kamu kurumları tarafından YZ alanında finanse edilen çalışmalarda etik değerlerin ve uluslararası normların da öncelikli olarak dikkate alınmasına dikkat edilmesi.

Yayın Etiği Bildirimi / Research Ethics

The authors declare that the research has no unethical problems, and that they observe the research and publication ethics. / Yazarlar araştırmanın etik dışı bir sorunu olmadığını, araştırma ve yayın etiği konusunu gözlemlediğini beyan etmektedir.

Araştırmacıların Katkı Oranı / Contribution Rate of Researchers

The contribution rates to each stage of the study is hundred percent. / Çalışmanın her aşamasına yazar yüzde yüz oranında katkı sunmuştur.

Çıkar Çatışması / Conflict of Interest

The study has no conflict of interest. / Çalışmada herhangi bir çıkar çatışması bulunmamaktadır.

Fon Bilgileri / Funding

There is no funding for this study. / Bu çalışmada herhangi bir fon kullanılmamıştır.

Etik Kurul Onayı / The Ethical Commitee Approval

Etik kurul kararı: Bu araştırma, derleme türünde makale olduğu için etik kurul kararı gerektirmemektedir. /

The Ethical Committee Approval: Since this research is a review article, it does not require an ethics committee decision.

Kaynakça/References

AI4PEOPLE (2020). Towards the 7 AI Global Frameworks. Paper presented at *the AI4PEOPLE 2020 SUMMIT*. Web üzerinde https://ai4people.eu/

Ali A.Z. (2012). A Philosophical approach to artificial intelligence and islamic values. *IIUM Engineering Journal, 12*(6), Special Issue in Science and Ethics.

Assion S. (2017). Legal Practice on the Edge of Disruption, *DigitalBusiness Law* , Şubat 8.

Barredo A., A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012 adresinden alınmıştır.

Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, *1*, 20–23. https://doi.org/10.1038/s42256-018-0004-1 adresinden alınmıştır.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., et al. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287 adresinden alınmıştır.

Berg, C., Davidson, S., & Potts, J. (2019). Blockchain technology as economic infrastructure: Revisiting the electronic markets hypothesis. *Frontiers in Blockchain*, *2*(22), 1–6. https://doi.org/10.3389/fbloc.2019.00022 adresinden alınmıştır.

Bryson, J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, *20*(1), 15–26.

Bughin J., Seong J., Manyika J., Chui M., Joshi R. (2018). Notes From the AI Frontıer Modeling the Impact of AI on the World Economy, McKinsey Global, *Discussion Paper*, Web üzerinde https://t.ly/Dz89 adresinden alınmıştır.

Chen, Y.S. (2020). The different kind of AI in "What is AI?". *Slideshare Sunumu*. Web üzerinde https://www.slideshare.net/ssuserff66e5/what-is-ai-218562334 adresinden alınmıştır.

China (2019). National Governance Committee for the New Generation Artificial Intelligence. Governance Principles for the New Generation Artificial Intelligence--Developing Responsible Artificial Intelligence. Web üzerinde https://www.chinadaily.com.cn/a/201906/17/WS5d07486 ba3103dbf14328ab7.html adresinden alınmıştır.

Clancy H. (2016) Get smarter on artificial intelligence, *GreenBiz 101*, Ocak 5, Web üzerinde https://www.greenbiz.com/article/greenbiz-101-get-smarter-artificial-intelligence adresinden alınmıştır.

Cremers, A, B., Englander, A., Gabriel, M., Hecker, D., Mock, M., Poretschkin, M., … Wrobel, S. (2019). Trustworthy use of artificial intelligence. Priorities From a Philosophical, Ethical, Legal, and Technological Viewpoint as a Basis for Certification of Artificial Intelligence. Web üzerinde

https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_Thrustworthy_AI.pdf adresinden alınmıştır.

Çin, (2020) Beijing AI Principles, Web üzerinde https://www.baai.ac.cn/news/beijing-ai-principles-en.html adresinden alınmıştır.

EU, (2019) Ethics guidelines for trustworthy AI, *EC Report*, 8 Nisan, Web üzerinde https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai adresinden alınmıştır.

Firebaugh M. W., *Artificial Intelligence, A Knowledge–based Approach*, Boston, PWS-Kent Publishing Company, 1988, p.14.

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, *1*(6), 261–262. https://doi.org/10.1038/s42256-019-0055-y adresinden alınmıştır.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*(1), 1–15. https://doi.org/10.1162/99608f92.8cd550d1 adresinden alınmıştır.

Floridi, L., & Taddeo, M. (2016). What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical. *Physical and Engineering Sciences*, *374*(2083), 1–5. https://doi.org/10.1098/rsta.2016.0360 adresinden alınmıştır.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5 adresinden alınmıştır.

Francke Matthew (2017) Why You Should Want a Lawyer (and Not a Robot), Best Hooper website, visited February 19.

FLI, (2017) *Asilomar AI Princples*. Future of Life Institute, Web üzerinde https://futureoflife.org/ai-principles/ adresinden alınmıştır.

Goggin, B. (2019). Inside Facebook's suicide algorithm: Here's how the company uses artificial intelligence to predict your mental state from your posts. Web üzerinde https://www.businessinsider.com/facebook-is-using-ai-to-try-to-predict-if-youre-suicidal-2018-12 adresinden alınmıştır.

Goodrich, M. A., & Schultz, A. C. (2007). Human–robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, *1*(3), 203–275. https://doi.org/10.1561/1100000005 adresinden alınmıştır.

Grad, P., (2020) AI Jesus writes Bible-inspired verse, *TechXplore*, Eylül 2, Web üzerinde https://techxplore.com/news/2020-09-ai-jesus-bible-inspired-verse.html adresinden alınmıştır.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925 adresinden alınmıştır.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*, 99–120. https://doi.org/10.1007/s11023-020-09517-8 adresinden alınmıştır.

Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer*, *51*(9), 28–36. https://doi.org/10.1109/MC.2018.3620965 adresinden alınmıştır.

Hancock, P. A. (2017). Imposing limits on autonomous systems. *Ergonomics*, *60*(2), 284–291. https://doi.org/10.1080/00140139.2016.1190035 adresinden alınmıştır.

Hill, K. (2020). The secretive company that might end privacy as we know it. *The New York times*. Web üzerinde https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html adresinden alınmıştır.

Hofman, D., Lemieux, V., Joo, A., & Batista, D. (2019). "The margin between the edge of the world and infinite possibility": Blockchain, GDPR and information governance. *Records Management Journal*, *29*(1/2), 240–257. https://doi.org/10.1108/RMJ-12-2018-0045 adresinden alınmıştır.

Hunt, E. (2016) Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter, *The Guardian*, 26 Mart, Web üzerinde *https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter* adresinden alınmıştır.

Independent High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. Brussels*: European Commission* Web üzerinde https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 adresinden alınmıştır.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2 adresinden alınmıştır.

Joshi N., (2019) 7 Types Of Artificial Intelligence, *Forbes*, Web üzerinde https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/?sh=450e58d233ee adresinden alınmıştır.

Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, *9*(1), 90–96. https://doi.org/10.1109/MITS.2016.2583491 adresinden alınmıştır.

Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, *16*(10), 880–918. https://doi.org/10.17705/1jais.00411 adresinden alınmıştır.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 adresinden alınmıştır.

Lee, P. (2016). Learning from Tay's introduction. Web üzerinde https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ adresinden alınmıştır.

Lohr Steve (2017). Robots Will Take Jobs, but Not as Fast as Some Fear, New Report Says, *The New York Times*, posted January 12,

Magid, J.M., (2020), Does your AI discriminate?, *The Conversation*, Mayıs 15, Web üzerinde https://theconversation.com/ does-your-ai-discriminate-132847 adresinden alınmıştır.

McKnight, D. H., & Chervany, N. L. (2001). What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *International Journal of Electronic Commerce*, *6*(2), 35–59. https://doi.org/10.1080/10864415.2001.11044235 adresinden alınmıştır.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS)*, *2*(2), 1–25. https://doi.org/10.1145/1985347.1985353 adresinden alınmıştır.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81 adresinden alınmıştır.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv e-prints.*

Midgley, J., Hall, A., Hardiman, M. and Narine, D. (1986), *Community Participation, Social Development and the State*, London: Methuen, p. 18.

Montreal, (2018), Developing AI in a responsible way, *Salle De Presse Udemnouvelles*, Montreal Üniversitesi, 12 Nisan, Web üzerinde https://nouvelles.umontreal.ca/en/article/2018/12/04/developing-ai-in-a-responsible-way/ adresinden alınmıştır.

Mostafa, S. A., Ahmad, M. S., & Mustapha, A. (2019). Adjustable autonomy: A systematic literature review. *Artificial Intelligence Review*, *51*(2), 149–186. https://doi.org/10.1007/s10462-017-9560-8 adresinden alınmıştır.

Bostrom Nick, (1998) "How Long Before Superintelligence?" in *International. Journal. of Future Studies, vol2.*

Noorman, M., & Johnson, D. G. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology*, *16*(1), 51–62. https://doi.org/10.1007/s10676-013-9335-0 adresinden alınmıştır.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342 adresinden alınmıştır.

OECD (2019). OECD Principles on AI. Web üzerinde https://www.oecd.org/going-digital/ai/principles/ adresinden alınmıştır.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5 adresinden alınmıştır.

Rai, A., Constantinides, P., & Sarker, S. (2019). Editor's comments: Next-generation digital platforms: Toward human–AI hybrids. *MIS Quarterly*, *43*(1), iii-x. https://doi.org/10.5555/3370135.3370136 adresinden alınmıştır.

Rouse M., (2017) Artificial Intelligence as a Service (AIaaS), *Search Enterprise AI*, Web üzerinde https://searchenterpriseai.techtarget.com/definition/Artificial-Intelligence-as-a-Service-AIaaS adresinden alınmıştır.

Sarpatwar, K., Vaculin, R., Min, H., Su, G., Heath, T., Ganapavarapu, G., & Dillenberger, D. (2019). *Towards enabling trusted artificial intelligence via Blockchain*. In S. Calo, E. Bertino, & D. Verma (Eds.), Policy-based autonomic data governance (137–153). Cham: Springer International Publishing.

Sarwate, A. D., & Chaudhuri, K. (2013). Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Processing Magazine*, *30*(5), 86–94. https://doi.org/10.1109/MSP.2013.2259911 adresinden alınmıştır.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228 adresinden alınmıştır.

Schindler, P. L., & Thomas, C. C. (1993). The structure of interpersonal trust in the workplace. *Psychological Reports*, *73*(2), 563–573. https://doi.org/10.2466/pr0.1993.73.2.563 adresinden alınmıştır.

Smith, V., Chiang, C, K., Sanjabi, M., & Talwalkar, A. S. (2017). Federated multi-task learning. Paper presented at the Advances in *Neural Information Processing Systems* 30 (NIPS 2017), Long Beach, CA,

Stormont, D, P. (2008). Analyzing human trust of autonomous systems in hazardous environments. Paper presented at the *Human Implications of Human-Robot Interaction workshop at AAAI*, Menlo Park, CA, USA.

Swarte, T., Boufous, O., & Escalle, P. (2019). Artificial intelligence, ethics and human values: The cases of military drones and companion robots. *Artificial Life and Robotics*, *24*(3), 291–296. https://doi.org/10.1007/s10015-019-00525-1 adresinden alınmıştır.

Thiebes, S., Lins, S. & Sunyaev, A. (2020) Trustworthy artificial intelligence. *Electron Markets.* https://doi.org/10.1007/s12525-020-00441-4

Tramer, F., & Boneh, D. (2019). Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. Paper presented at *the International Conference on Learning Representation*s, New Orleans, LA

Vincent, J., (2017) Putin says: the nation that leads in AI 'will be the ruler of the world', *The Verge*, Sep 4, Web üzerinde https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world adresinden alınmıştır.

Vought, R, T. (2020). Guidance for Regulation of Artificial Intelligence Applications Web üzerinde https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf adresinden alınmıştır.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., et al. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, *25*(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6 adresinden alınmıştır.