İnsan Merkezli Yapay Zeka: Literatür Taraması

Human Centered Artificial Intelligence: Literature Review

[1]Mehmet Kaan Erol
Bilgisayar Mühendisliği Teknoloji
Fakültesi Marmara Üniversitesi
İstanbul, Türkiye
kaanerol@marun.edu.tr

***Özetçe*—Derin öğrenme, yapay zekaya olan ilgiliyi arttıran önemli bir gelişme olmuştur. Yapay zekanın daha yaygın hale gelmesiyle karar süreçlerimize dahil olmuştur. İnsan ihtiyaçlarının, değerlerinin bu süreçte yapay zeka gelişmelerinin merkezinde tutulması gereklidir. Bu çalışmada, gözden geçirme yaklaşımı izlenerek insan merkezli yapay zeka hakkında literatür taraması yapılmıştır. Literatürde yapılan çalışmalar açıklanabilirlik, sosyal etki ve gerçek dünya problemlerine uygulanabilirlik olarak ayrılmış ve incelenmiştir. İnsan merkezli yapay zeka hakkındaki güncel gelişmeler, zorluklar ve fırsatlar sunulmuştur.**

Anahtar Kelimeler—insan merkezli yapay zeka açıklanabilir yapay zeka, yapay zeka bilgisayar etkileşimi, uygulanabilirlik, yorumlanabilirlik

***Abstract*—Deep learning has been an important development that has increased the interest in artificial intelligence. As artificial intelligence has become more common, it has been included in our decision processes. Human needs and values should be kept at the center of artificial intelligence developments in this process. In this study, a literature review on human-centered AI was conducted following the review approach. Studies in the literature are divided and examined as explainability, social impact, and applicability to real world problems. Current developments, challenges and opportunities in human-centered AI are presented.**

Keywords—human-centered artificial intelligence, explainable artificial intelligence, artificial intelligence computer interaction, applicability, interpretability

I. GİRİŞ

İnsan merkezli yapay zeka, bir teknoloji olarak YZ'nın (yapay zeka) geliştirilmesi sırasında varlık olarak insanı ve değerlerini göz önüne alarak geliştirilmesidir. Bazı kavramlar; güvenilirlik, açıklanabilirlik, güvenlik, etik ve uygulanabilirliktir. YZ teknolojisinin gelişmesiyle beraber daha geniş uygulama alanları bulunmuştur [1]. Hayatımıza daha çok girmesi geliştirenlere ve kullanıcılara yeni sorumluluklar yüklemektedir [2]. Teknoloji olarak YZ'nın geliştirilmesi sadece teknik bir problem olarak ele alınmamalı, uygulamada insanlar ile olan etkileşimi de göz önünde bulundurulmalıdır [3]. İnsan merkezli YZ, YZ ve makine öğrenimi üzerine bir perspektiftir ve algoritmaların insanlardan oluşan büyük bir sistemin parçası olduklarının bilinciyle tasarlanması gerekmektedir [4]. Bu çalışma insan merkezli YZ hakkında literatürde yapılan çalışmaları açıklanabilirlik, sosyal etki ve gerçek dünya problemlerine uygulanabilirlik olarak 3 ana başlık altında incelemektedir.

II. LİTERATÜR TARAMA YÖNTEMİ

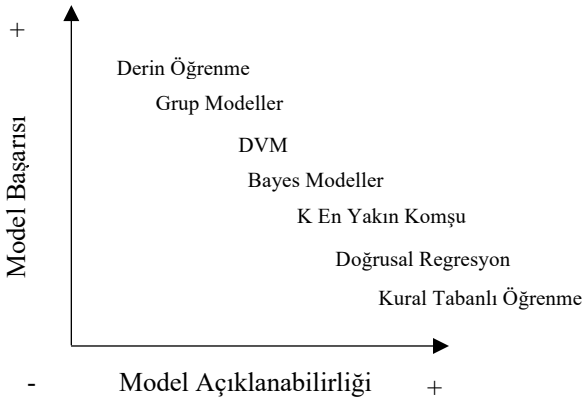
İnsan merkezli YZ yeni bir çalışma alanı olduğundan belli akademik veritabanlarıyla sınırlı kalınmamıştır. Literatür tarama süreci Google'n akademik arama motoru olan Google Scholar ile belli anahtar kelimeler kullanılarak yapılmıştır. Literatür taraması sırasında kullanılan anahtar kelimeler; "insan merkezli yapay zeka", "açıklanabilirlik", "sosyal etki", "insan yapay zeka etkileşimi", "yorumlanabilirlik", "güvenilirlik", "etik yapay zeka", "yapay zeka destekli karar verme" olarak seçilmiştir. Yayın yılı 2009 – 2020 olan güvenilir kaynaklara sahip çalışmalar seçildi. Seçilen kaynaklar özet ve sonuç değerlendirilerek son kez elendi. Seçilen yayınlar açıklanabilirlik, sosyal etki ve gerçek dünya problemlerine uygulanabilirlik olarak üç ana gruba ayrıldı. Her grup kendi içinde sentezlendi ve ayrı başlıklarda bilgi bütün olarak sunuldu ve örnek çalışmalar verildi.

III. İNSAN MERKEZLİ YAPAY ZEKA

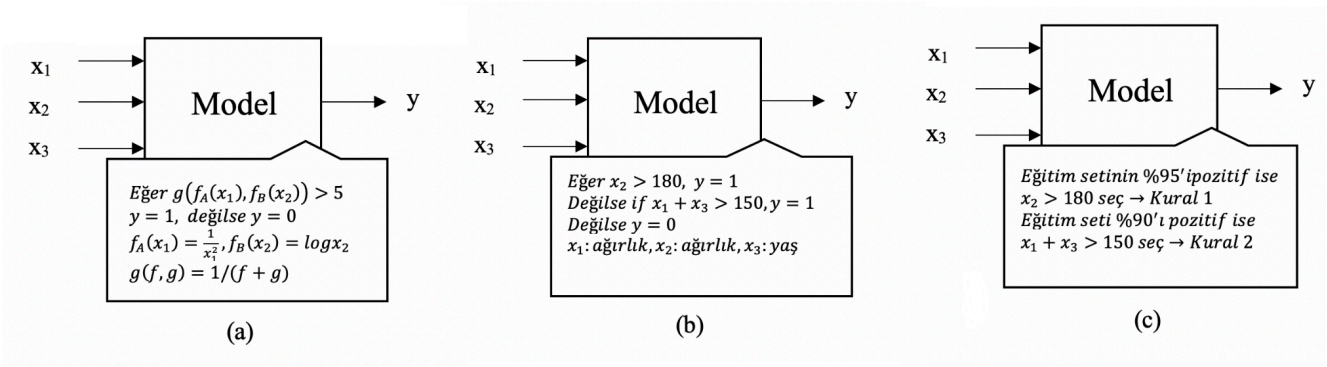
İnsan merkezli YZ, bir tasarımın geliştirilmesinde insanları ve ihtiyaçlarını, duygularını, davranışlarını ve bakış açısını merkeze alan tasarım yaklaşımıdır [5]. İnsanları bir sistemin parçası olarak değil, insanları tasarıma dahil tasarımın her aşamasında "merkezlenmiş" olarak görmektir. YZ, kullanıldığı toplumu ve endüstriyi etkileme potansiyeline sahiptir [5]. Bu potansiyel etik, sosyal etki gibi insanlar için neticelerin farkında olarak tasarlamayı gerektirmektedir.

A. Açıklanabilirlik

Açıklanabilirlik literatürde şeffaflık, güven, adalet, yorumlanabilirlik, ve hesap verebilirlik ile ilgilidir [6, 7, 8]. Doshi vd. tarafından açıklanabilirlik, "bir insana anlaşılır açıklama veya sunma becerisi" olarak tanımlanmaktadır [9].



Şekil 1. Model yorumlanabilirliği ve performans arasındaki denge.



Şekil 2. Farklı açıklanabilirlik düzeylerini gösteren diyagram [7].

Gilpin vd., tersine, açıklanabilirliği yorumlanabilirlikten daha geniş bir konu olarak kabul etmektedir [10]. Rudin, açıklanabilir makine öğrenimini daha sınırlı bir anlamda, "Kara kutu olmayan bir model" olarak tanımlamaktadır [11]. YZ sistemlerinin çoğu kararlarını ve eylemlerinin nedenlerini açıklayamamaktadır. Bazı YZ uygulamaları için açıklamalar gerekli olmasa da savunma, tıp, finans ve hukuk alanındaki kritik uygulamalar için, kullanıcıların süreçleri etkili bir şekilde yönetebilmesi için kararı etkileyen faktörleri anlamaları gerekmektedir [12, 13, 14]. Çoğunlukla, en yüksek tahmin başarısına sahip yöntemler (derin öğrenme yöntemleri) en az açıklanabilir ve en açıklanabilir olan (karar ağaçları, doğrusal regresyon, kural tabanlı yaklaşımlar) yöntemler de en az tahmin başarısına sahiptir (Şekil 1'de gösterilmektedir) [7, 15].

Açıklanabilir bir YZ sistemi, davranışını kullanıcılar için daha anlaşılır hale getirmek, uzman kullanıcılar için makinenin algıladığı örüntüleri değerlendirerek çeşitli araştırmalar da yapılabilir hale getirmektedir [16]. Daha anlaşılabilir bir YZ sistemi için bazı genel ilkeler: yapabilme becerilerini ve anlayışını açıklayabilmesi; ne yaptığını, ve zamana bağlı ne yaptığını açıklayabilmesi; harekete geçiren faktörleri açıklayabilmesidir [17]. Bununla birlikte, her açıklama her kullanıcıya hitap etmemektedir. YZ'nın uygulama alanlarına göre talep edilen açıklanabilirlikte değişmektedir. Şekil 2'de gösterilen diyagram aynı model için farklı seviyelerde yorumlar içermektedir.

Sistemin açıklanabilirlik durumunu değerlendirmek için birkaç yol önerilmiştir; ancak, herhangi bir açıklanabilir YZ sisteminin bir kullanıcı için yeterince anlaşılır olup olmadığını ölçmenin ortak bir yolu bulunmamaktadır [18]. Bunlar genel olarak kullanıcının bakış açısına bağlı öznel değerlendirmelerdir [18, 19, 20]. Bazı çalışmalarda modelin karmaşıklığını sadeleştirme yoluyla yerel açıklanabilirlik sağlamaktadır [21, 22, 23]. Ribeiro vd., yerel olarak tahmini yorumlanabilir bir modeli öğrenerek, tahminlerini yorumlanabilir ve modele sadık bir şekilde açıklayabilen bir algoritma önermektedir [21].

B. Sosyal Etki

YZ "sosyal, kültürel, ekonomik ve politik etkileşimlerimize giderek daha fazla aracılık ediyor", insanların yaşamlarını etkiliyor ve bir bütün olarak toplumları etkilemektedir [24]. Akıllı asistanlar, sürücüsüz arabalar, ürün önermek ve banka kredilerini onaylamak için arka planda çalışan algoritmalar gibi, günlük hayatın her noktasında kullanılmaktadır. Bu sistemleri tasarlarken insanların yaşamlarını etkileme potansiyeli göz önünde bulundurarak tasarlamak önemlidir [5]. İnsan Merkezli YZ, insanları destekleyen, üretkenliği teşvik eden, sorumluluğu netleştiren ve sosyal katılımı kolaylaştıran YZ sistemleri tasarlamak için önemlidir. Ayrıca insanlara özgü mahremiyet, güvenlik, çevrenin korunması, sosyal adalet ve insan haklarının dikkate alınmasını teşvik etmektedir [25].

Chancellor vd., insan merkezli makine öğrenimi tanımına göre, diğer disiplinlere yönelik, "döngüye insanı" getiren veya yalnızca yöntem yenilikleri sunan bir yaklaşımdan daha fazlasıdır [23]. İnsan merkezlilik, insanların, toplulukların ve toplumun ihtiyaçlarına kasıtlı olarak yeniden odaklanmak ve sorunları çözmek için uygun araçları tanımlamaktır [26, 27]. Jarrahi vd. göre, YZ'yı her derde deva olarak görmek ve yaklaşmak ileriyi görememektir [28]. Ancak uzun vadeli olarak bu sistemler bir organizasyonun sosyal dokusuna mantıklı bir şekilde entegre edildiklerinde etkili olabilmektedir [29]. Riedl, insanları sosyokültürel bir bakış açısıyla anlayan ve insanların onları anlamasına yardımcı olan YZ sistemleri olarak insan merkezli YZ'yı iki yöne ayırmıştır [30]. Bazı YZ araştırmacıları, hesap verebilirlik, yorumlanabilirlik ve şeffaflık gibi sosyal sorumluluk göz önünde bulundurularak tasarlanmış akıllı sistemlere atıfta bulunmak için insan merkezli YZ terimini kullanmaya başladı [30]. Yine Riedl'a göre insanlar ile YZ sistemleri arasındaki sağduyu kaynaklı başarısızlık (istenilen sonucun elde edilememesi) YZ sisteminden kaynaklı değil, kullanıcı kaynaklıdır [30]. Çünkü hedefin bir kısmı veya hedefe varan yol yeterince açık tanımlanmamıştır. İnsanlar topluluklar olarak yaşar ve aynı toplumdan, kültürden insanlar tarafından ortak paylaşılan bilgiler mevcuttur. Günlük hayatta bir insan için sıradan bir komut bir YZ sistemi için fazla anlaşılmaz, eksiklerle dolu olabilmektedir. Sağduyu bilgisinin bilgi temellerini oluşturmak için çeşitli çalışmalar bulunsa da, yeterli değildir [30, 31, 32].

İnsan merkezli YZ'dan beklenti insan gibi düşünüp hareket etmesi değildir. Etkileşim kurduğu bireyi ya da birey grubunu sosyal ve kültürel normlar dahilinde ayrımcılık ve ön yargıdan kaçınarak değerlerini, beklentilerini, ihtiyaçlarını anlamasıdır [30].

TABLO 1. Amershi vd. yaptığı çalışmaya göre kullanıcılarla etkileşim sırasında uygulanma ihtimaline göre kategorize edilmiş 18 insan-YZ etkileşimi tasarım kılavuzu [38].

Başlangıç	Etkileşim Sırasında	Hata Sırasında	Zamanla
Sistemin neler yapabileceğini netleştirin.	Bağlama dayalı zaman hizmetlerini kullanın.	Verimli çağrıları destekleyin.	Son etkileşimleri hatırlayın.
		Beklenmedik işlemleri sonlandırın.	Kullanıcı davranışından öğrenin.
	Bağlamsal olarak alakalı bilgileri gösterin.		Dikkatlice güncelleyin ve uyarlayın.
Kullanıcının YZ sisteminin neler yapabileceğini anlamasını sağlayın.		Etkili düzeltmeyi destekleyin.	Ayrıntılı geri bildirimleri teşvik edin.
	İligili sosyal normları eşleştirin.	Belirsizliği ortadan kaldıran hizmetleri destekleyin.	Kullanıcı eylemlerinin sonuçlarını iletin.
			Genel kontroller sağlayın.
	Sosyal önyargıları azaltın.	Sistemin neden yaptığını netleştirin.	Kullanıcıları değişiklikler hakkında bilgilendirin.

C. Uygulanabilirlik

İnsanları merkeze almak, YZ sistemlerinin etik olmasını, benimsenmesini, kullanılabilir olmasını sağlar ve YZ sistemlerinin istenmeyen zararlı sonuçlarından kaçınmaya yardımcı olmaktadır [33]. Xu vd. göre, YZ insanların yerini almamalı, bunun yerine insan yeteneklerini artırmalıdır [34]. İnsanları değiştirmeme konusundaki bu ifade, insanları merkeze yerleştirmek için önemli bir girişimdir [35]. Aynı makale insan merkezli YZ çözümlerinin etik, açıklanabilir, anlaşılır, yararlı ve kullanılabilir olması gerektiğini önermektedir [34]. Ayrıca, üç bileşeni içeren çalışan bir insan merkezli YZ çerçevesi sağlar: etik olarak uyumlu tasarım, insan etkenleri tasarımı ve teknoloji geliştirme [34].

YZ geliştirme aşamasında insan merkezli olamama nedeniyle, istenmeyen algoritmik ön yargı örnekleri ortaya çıkmaktadır. Bu ön yargılı algoritmik ön yargılar geliştirme aşamasından önce veri toplama aşamasından kaynaklanmaktadır [35]. Geliştirme aşamasında bu riskler önceden tanımlanmalı ve potansiyel riskler önlenmelidir. IBM ve Aequitas, veri bilimcilerinin makine öğreniminden önce veri kümelerindeki yanlılığı tespit etmelerine yardımcı olmak için araç setleri geliştirmişlerdir [36, 37]. Uygulanabilirlikteki bir diğer sorun etiktir. Bir insanın etik olmayan bir davranışta bulunması ne kadar yanlışsa bir YZ sisteminin için de aynı derece yanlıştır. Bu gibi problemler insan merkezli YZ sistemlerinin uygulanabilirlikteki başlıca sorunlarıdır.

Raymond vd. göre, HCI (insan bilgisayar etkileşimi) ve UX (kullanıcı deneyimi) araştırmacılarının insan YZ etkileşimleriyle ilgili zorlukları her zamankinden daha fazla ele alması gerektiği yönündedir [35]. Amershi vd. göre, insan YZ etkileşiminin ilkeleri, HCI topluluğunda yirmi yıldan fazla bir süredir tartışılmakta, ancak YZ'deki gelişmeler ve insana yönelik uygulamalarda YZ teknolojilerinin artan kullanımları nedeniyle daha fazla çalışmaya ihtiyaç vardır [38]. HCI'yi YZ geliştirme ile entegre etmek, çok disiplinli bir yaklaşım sağlamaktadır [35]. Amershi vd. yaptıkları çalışmada YZ teknolojilerini kullanan uygulamaların ve özelliklerin tasarımı üzerinde çalışan uygulayıcılar ve insan – YZ etkileşim tasarımı için 18 yönergeden oluşan bir tasarım kılavuzu sunmuşlardır. Tablo 1'de gösterilen yönergeleri dört aşamalı bir süreç kullanarak geliştirmişlerdir. Birinci aşamada, 150'den fazla tasarım önerisi 20 kılavuzluk bir kümede birleştirilmiş, ikinci aşamada, bir sezgisel değerlendirmesi gerçekleştirilerek 18'e indirilmiş [38]. Üçüncü aşama, 49 katılımcının kılavuzların uygunluğunu ve netliğini değerlendirmek için geri bildirimlerine dayanarak, anlaşılırlığı artırmak için bazı yönergeler yeniden ifade edilmiş ve dördüncü aşamada, doğrulamak için gözden geçirmelerin uzman değerlendirilmesi yapılmıştır [38].

IV. SONUÇ VE TARTIŞMA

1980'lerde bilgisayar teknolojisinin gelişmesiyle beraber kişisel bilgisayar kavramı ortaya çıktı. O döneme kadar sistemler yalnızca uzmanlar ve alana özel ilgi duyan teknoloji meraklıları tarafından yine uzman kullanıcı için geliştirilmekteydi. Hedef sadece sistem geliştirmekti ve kullanıcı ön planda bulunmamaktaydı. Bu HCI gibi teknik ve sosyal bilimlerin ortak çalışmasını gerektiren bir çalışma alanının doğmasına sebep olmuştur. HCI, insan bilgisayar arasındaki etkileşimi incelemekte ve geliştirme aşamasında kullanıcıyı ön plana almaktadır.

Bugün YZ sistemlerinde de benzer bir durum görmek mümkündür. Günümüzde ağırlıklı olarak YZ sistemlerini geliştirenler teknik alanda uzman kişilerdir. Geliştirme sürecinde insan ve insanın toplumsal varlığı değerlendirilmemekte, YZ sistemlerinden beklenti daha hızlı ve daha yüksek doğruluk oranlarıdır. Derin öğrenme yöntemlerinin popülerleşmesiyle YZ sistemleri daha da belirsiz hale gelmiştir. Derin öğrenme, büyük veri kümeleri üzerinde çıkarımlar yapmak için kullanılan bir makine öğrenmesi modelidir. Çok katmanlı ve doğrusal olmayan yapıları nedeniyle kara kutu modeli olarak açıklanabilir olmadığından eleştirilmektedir. Özellikle tıp, savunma sanayi, otonom araçlar gibi kritik alanlarda kullanılan bu sistemler için açıklanabilirlik ve hesap verilebilirlik daha da önem kazanmıştır. YZ sistemlerinin hızlı ve yüksek doğrulukta karar vermesi yeterli değildir. Şeffaf ve anlaşılabilir olması da gerekmektedir.

V. İLERİYE YÖNELİK ÇALIŞMALAR

Literatürde insan merkezli YZ üzerine çeşitli alanlardan ve çok disiplinli çalışmalar mevcuttur. Bu çalışmalar YZ sistemlerinin mevcut durumunu incelemiş ve olası etkilerine yönelik yaklaşımlarda bulunmuştur. Tam anlamıyla uygulanabilir bir yöntem ya da açıklanabilirlikte olduğu gibi genel bir ölçek önerilememiştir. Mevcut çalışmalar YZ

sistemlerinin günümüzdeki ve gelecekteki olası etkilerine dayanarak çıkarımlarda bulunmuş ve bazı çalışmalarda bunlara çeşitli çözüm önerileri getirilmiştir. Bu çözüm önerileri problemlerin çeşitliliğinde olduğu kadar çeşitli disiplinlerden gelmiştir. İnsan merkezli YZ ve alt çalışma alanları olan açıklanabilirlik, etik, güvenilirlik, güvenlik vd. konularında yeterince teknik çalışma bulunmamaktadır. Önde gelen üniversitelerde bu alanda çalışmalar yapmak üzere büyük fonlar toplanmış ve araştırma merkezleri kurulmuştur.

KAYNAKÇA

[1] Kantarjian, H., Yu, P. P., "Artificial intelligence, big data, and cancer," *JAMA oncology*, vol. 1, no. 5, pp. 573-574, 2015.

[2] Xu, W., "Toward human-centered AI: a perspective from human-computer interaction," vol. 26, no. 4, pp. 42-46, 2019.

[3] Kowert, W., "The foreseeability of human-artificial intelligence interactions," *Tex. L. Rev*., vol. 96, pp. 181, 2017.

[4] Riedl, M. O., "Human-centered artificial intelligence and machine learning," *Human Behavior and Emerging Technologies*, vol.1, no. 1, pp. 33-36, 2019.

[5] Auernhammer, J., "Human-centered AI: The role of Human-centered Design Research in the development of AI," 2020.

[6] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, Mohan K. "Trends and Trajectories for Explainable, Account- able and Intelligible Systems: An HCI Research Agenda," *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems Association for Computing Machinery*, pp. 1–18, 2018.

[7] Hoffman, R. R., Mueller, S. T., Klein, G., Litman, J., "Metrics for explainable AI: Challenges and prospects," 2018.

[8] Shneiderman, B. "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction,* vol. 36, no. 6, pp. 495-504, 2020.

[9] F. Doshi-velez, Been Kim. "A Roadmap for a Rigorous Science of Interpretability," pp. 1050, 2017.

[10] Leilani H. Gilpin, D. Bau, Ben Z. Yuan, A. Bajwa, M. Specter, Lalana Kagal. 2018. "Explaining Explanations: An Ap- proach to Evaluating Interpretability of Machine Learning," 2018.

[11] C. Rudin. "Please stop doing 'explainable'," *ML*, 2018.

[12] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K. R. Muller, "Explainable AI: Interpreting, Explaining and Visualizing Deep Learning," *Springer Nature*, 2019.

[13] [13] H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven, "Explainable and Interpretable Models in Computer Vision and Machine Learning," *Springer*, 2018.

[14] O. Biran, C. Cotton, "Explanation and justification in machine learning: A survey," *The IJCAI-17 Workshop on Explainable AI (XAI)*, vol. 8, no. 1, pp. 8-13, 2017.

[15] Defense Advanced Research Projects Agency. "Broad Agency Announcement, Explainable Artifcial Intelligence (XAI)," *DARPA-BAA-16-53*, 2016.

[16] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Chatila, R. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82-115, 2020.

[17] V. Bellotti, K. Edwards, "Intelligibility and accountability: Human considerations in context-aware systems," *Hum. Comput. Interact*. vol. 16, pp. 193–212, 2009.

[18] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G. Z. "XAI—Explainable artificial intelligence," *Science Robotics,* vol. 4, no. 37, 2019.

[19] Ribeiro, M.T., Singh, S., Guestrin, C. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," *CHI 2016 Workshop on Human Centered Machine Learning*, 2016.

[20] Ribera, M., Lapedriza, A. "Can we do better explanations? A proposal of user-centered explainable AI," *In IUI Workshops*, 2019.

[21] Si, Z. Zhu, S. "Learning AND-OR Templates for Object Recognition and Detection," *IEEE Transactions On Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2189-2205, 2013.

[22] Lake, B.H., Salakhutdinov, R., Tenenbaum, J.B. "Human-level concept learning through probabilistic program induction," Science, vol. 350, pp. 1332-1338, 2015.

[23] Chancellor, S., Baumer, E. P., De Choudhury, M. "Who is the 'Human' in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media," *Proceedings of the ACM on Human-Computer Interaction*, pp. 1-32, 2019.

[24] Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Jennings, N. R. "Machine behaviour," *Nature*, vol. 568, no. 7753, pp. 477-486, 2019.

[25] Shneiderman, B. "Human-Centered Artificial Intelligence: Three Fresh Ideas," *AIS Transactions on Human-Computer Interaction*, vol. 12, no. 3, pp. 109-124, 2020.

[26] L. Bannon. "Reimagining HCI: toward a more human-centered perspective," *interactions*, vol. 18, no. 4, pp. 50–57, 2011.

[27] R. Kling, S. L. Star. "Human centered systems in the perspective of organizational and social informatics," *ACM SIGCAS Computers and Society*, vol. 28, no. 1, pp. 22–29, 1998.

[28] Jarrahi, M. H. "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," *Business Horizons*, vol. 61, no. 4, pp. 577-586, 2018.

[29] Sawyer, S., Jarrahi, M. H. "Sociotechnical approaches to the study of Information Systems," *In Computing handbook, third edition: Information systems and information technology*, pp. 5-1, 2014.

[30] Riedl, M. O. "Human-centered artificial intelligence and machine learning," *Human Behavior and Emerging Technologies*, vol. 1, no. 1, pp. 33-36, 2019.

[31] Lenat, D. "Cyc: A large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, pp. 33–38, 1995.

[32] Liu, H., Singh, P. "ConceptNet—A practical commonsense reasoning tool-kit," *BT Technology Journal*, vol. 22, no. 4, pp. 211–226, 2014.

[33] Villaronga, E. F., Kieseberg, P., Li, T. "Humans forget, machines remember: Artificial intelligence and the right to be forgotten," *Computer Law & Security Review*, vol. 34, no. 2, pp. 304-313, 2018.

[34] Xu,W. "TowardHuman-CenteredAI:APerspective from Human-Computer Interaction," *Interactions* vol. 26, no. 4, pp. 42-6, 2019.

[35] Bond, R. R., Mulvenna, M., Finlay, D., Wong, A., Koene, A., Brisk, R., Adel, T. "Human Centered Artificial Intelligence: Weaving UX into Algorithmic Decision Making," *In RoCHI 2019: International Conference on Human-Computer Interaction*, 2019.

[36] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Zhang, Y. "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," 2018.

[37] Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Ghani, R. "Aequitas: A bias and fairness audit toolkit," 2018.

[38] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Horvitz, E. "Guidelines for human-AI interaction," *In Proceedings of the 2019 chi conference on human factors in computing systems,* pp. 1-13, 2019.