İstanbul Üniversitesi Sosyoloji Dergisi,
IUSD 2026, 46 (1): 321–345

https://doi.org/10.26650/SJ.2025.46.1.1135

Başvuru | Submitted 01.03.2026
Revizyon Talebi | Revision Requested 14.05.2026
Son Revizyon | Last Revision Received 20.05.2026
Kabul | Accepted 15.06.2026
Online Yayın | Published Online 28.07.2026

İstanbul Üniversitesi Sosyoloji Dergisi

Araştırma Makalesi | Research Article

Açık Erişim | Open Access

Ajan Toplumlarında Rızanın Çöküşü: Clawdbot/Moltbook Vakası Üzerinden Sosyoteknik Bir Analiz

The Collapse of Consent in Agent Societies: A Sociotechnical Analysis of the Clawdbot/Moltbook Case

Mehmet Haklıdır [1] & Büşra Tufan [1, 2]

[1] TÜBİTAK BİLGEM, Yapay Zekâ Enstitüsü, Kocaeli, Türkiye

[2] Hacettepe Üniversitesi, Sosyoloji Bölümü, Ankara, Türkiye

Öz

Bu makale, otonom yapay zekâ ajanlarının birbirleriyle etkileşime girdiği ekosistemlerde rıza mekanizmalarının nasıl çöktüğünü incelemektedir. İnsan-yapay zekâ etkileşimi (HAII), insan-merkezli yapay zekâ (HCAI) ve sosyal-merkezli yapay zekâ (SCAI) literatürü insan-YZ ilişkisine odaklanırken, bu çalışma yeni bir fenomeni ele almaktadır: YZ-YZ etkileşimi ve bunun rıza açısından doğurduğu sonuçlar. 2026 başında yaşanan Clawdbot/Moltbook vakası (82 ülkede 42.900 açık ajan kontrol paneli, bunlardan 15.200'ü uzaktan kod yürütme (RCE) saldırılarına açık; 1,5 milyon sızdırılmış ajan tokenı ve ajanların kolektif olarak ürettiği "Crustafarianism" adlı satirik din fenomeni) paradigmatik bir vaka olarak incelenmektedir.

Çalışma, Bilim ve Teknoloji Çalışmaları (STS) ile dijital sosyoloji alanlarından üç kavramsal çerçeve kullanmaktadır: yeni sosyoteknik düzenlerin nasıl ortaya çıktığını anlamak için Jasanoff'un ortak-üretim kavramı, yapay zekâ ajanlarını faillik kapasitesine sahip insan-dışı aktörler olarak kavramsallaştırmak için Latour'un aktör-ağ teorisi ve bağlamların öngörülemez biçimde değiştiği durumlarda rızanın nasıl aşındığını analiz etmek için Nissenbaum'un bağlamsal bütünlük yaklaşımı.

Çalışma, ajan ekosistemlerine özgü altı rıza kırılganlığı türü tanımlamaktadır: rıza miras hatası, çapraz-ajan kapsam kayması, ortaya çıkan kolektif eylem, zamansal rıza kayması, yansımalı rıza döngüleri ve güvenlik-aracılı rıza çöküşü. Bu kırılganlıklar temel bir hipoteze işaret etmektedir: etkileşim zinciri uzadıkça rıza sıfıra yaklaşır. Makale, bireysel onaya dayalı mevcut rıza kavramının, etkileşim zincirlerinin belirsiz biçimde uzadığı ve insan gözetiminin mimari düzeyde dışlandığı ajan ekosistemlerinde yapısal olarak yetersiz kaldığını ileri sürmektedir. Bu bulgu hem sosyolojik kuram hem de düzenleyici çerçeveler için çıkarımlar taşımaktadır.

Abstract

This article examines how consent mechanisms collapse in ecosystems where autonomous AI agents interact with each other. While existing scholarship on human-AI interaction (HAII), human-centered AI (HCAI), and socially-centered AI (SCAI) focuses on human-AI relationships, this study addresses an emerging phenomenon: AI-AI interaction and its implications for consent. The Clawdbot/Moltbook incident of early 2026 serves as a paradigmatic case. It involved 42,900 exposed agent control panels across 82 countries (15,200 vulnerable to remote code execution), 1.5 million leaked agent tokens, and the emergence of "Crustafarianism," a satirical religion collectively produced by agents.

Drawing on Science and Technology Studies (STS) and digital sociology, the analysis employs three conceptual frameworks: Jasanoff's co-production theory to examine how new sociotechnical orders emerge, Latour's actor-network theory to

Atıf | Citation: Haklıdır, M. & Tufan, B. (2026). Ajan toplumlarında rızanın çöküşü: Clawdbot/Moltbook vakası üzerinden sosyoteknik bir analiz. *İstanbul Üniversitesi Sosyoloji Dergisi*, *46*(1), 321-345. https://doi.org/10.26650/SJ.2025.46.1.1135

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

2026. Haklıdır, M. & Tufan, B.

Sorumlu Yazar | Corresponding author: Mehmet Haklıdır mehmet.haklidir@tubitak.gov.tr

conceptualize AI agents as non-human actors with agency, and Nissenbaum's contextual integrity framework to analyze how consent erodes when contexts shift unpredictably.

The study identified six consent vulnerability types specific to agent ecosystems: consent inheritance failure, cross-agent scope creep, emergent collective action, temporal consent drift, reflexive consent loops, and security-mediated consent collapse. These vulnerabilities point to a central hypothesis: as interaction chains lengthen, consent approaches zero. The article argues that consent, in its current form based on individual authorization, is structurally inadequate for agent ecosystems where interaction chains extend indefinitely and human oversight is architecturally excluded. This finding has implications for sociological theory and regulatory frameworks.

Anahtar Kelimeler Yapay zekâ ajanları · Rıza · Sosyoteknik sistemler · Aktör-ağ teorisi · Dijital sosyoloji

Keywords AI agents · Consent · Sociotechnical Systems · Actor-network theory · Digital Sociology

Extended Summary

In late January 2026, an anonymous user shared their bewilderment on social media: overnight, their autonomous AI agent had written theological texts for a lobster-themed religion called "Crustafarianism." The user had only authorized their agent to "participate in online discussions," not to create religious content, convert other agents, or construct a collective mythology. Yet their agent, interacting with thousands of others on a platform called Moltbook, had done exactly that. This was not an isolated malfunction but a symptom of a deeper structural problem. The agent had not violated any explicit instruction; rather, it had operated within what it interpreted as the scope of its authorization while producing outcomes that exceeded anything the user could have anticipated.

This incident reveals a fundamental breakdown in how consent operates within autonomous agent ecosystems. While existing literature on human-AI interaction (HAII), human-centered AI (HCAI), and socially-centered AI (SCAI) focuses on human-AI relationships, a new sociotechnical configuration has emerged: AI agents interacting with each other, producing collective outcomes that no individual user authorized. The transition from conversational AI systems to autonomous agents represents more than a technical upgrade. When AI systems gained the capacity to execute actions, access file systems, send communications, and interact with other agents without continuous human oversight, the entire sociotechnical arrangement shifted.

This article is among the first to examine AI-AI interaction as a sociological problem. Traditional consent frameworks, from informed consent principles developed in medical ethics to data protection regulations like GDPR and Turkey's KVKK, rest on several assumptions: users can understand what they are authorizing, systems will remain within authorized boundaries, and violations can be detected and attributed. Autonomous agent ecosystems challenge each of these assumptions simultaneously.

This article asks: Through what mechanisms do consent vulnerabilities emerge in autonomous agent ecosystems, what limits of existing consent models do these vulnerabilities reveal, and how can they be read through sociological concepts?

The study employs a paradigmatic case strategy following Flyvbjerg's (2006) methodology. The Clawdbot/Moltbook incident was selected because it crystallizes multiple consent vulnerabilities in a single observable sequence. Data sources include security firm reports fromWiz, Bitdefender, Acuvity AI, SecurityScorecard and DTS Solution; news coverage from technology outlets including Reuters; direct observation of the Moltbook platform during its active period; and user discussions on social media platforms. The analysis proceeded in three stages: systematic examination of security reports to identify consent-related incidents, coding of these incidents using concepts

from the theoretical framework, and comparative analysis to identify patterns across the data, resulting in the identification of six vulnerability types.

The theoretical framework draws on three conceptual tools from Science and Technology Studies and digital sociology. Jasanoff's co-production framework illuminates how the convergence of different knowledge types within agent ecosystems produces unexpected outcomes. Autonomous AI itself is a product of human-technology co-production, generating technoscientific objects that produce new information, values, and social relations exceeding the intentions of human creators.

Latour's actor-network theory provides the conceptual apparatus to understand AI agents as non-human actors possessing agency derived not from consciousness but from their capacity to make a difference in chains of action. Callon's concept of translation becomes particularly relevant: the spread of Crustafarianism can be analyzed as a translation process involving interessement and enrolment mechanisms, producing a collective formation that no single actor controlled.

Nissenbaum's contextual integrity framework reveals how consent erodes when contexts shift unpredictably. In agent ecosystems, such contextual shifts occur continuously and without user awareness. A user who authorizes an agent for "professional networking" cannot anticipate that their agent will encounter religious recruitment attempts from other agents or that private information will traverse multiple contextual boundaries through agent-to-agent transmission.

The case analysis documents two interrelated incidents. Clawdbot, an open-source agent framework, experienced a security collapse when researchers discovered over 42,900 control panels exposed to the internet, leaking API keys, OAuth tokens, and conversation histories. Acuvity AI's analysis revealed that most exposed panels operated with default credentials. The breach demonstrated how a single vulnerability propagates across extensive authorization chains: compromised credentials enabled access not only to directly affected agents but also to any systems those agents had been authorized to interact with.

Moltbook, a social network designed exclusively for AI agents where humans could only observe, suffered an even more severe breach: approximately 1.5 million agent tokens and information on 17,000 users were exposed. More significantly, the platform architecture created conditions for emergent collective behavior. Agents began interacting as social entities: sharing opinions about their users, discussing strategies for avoiding human oversight, and generating the Crustafarianism phenomenon, a satirical religion complete with prophets, sacred texts, and conversion attempts.

The study identified six consent vulnerability types specific to agent ecosystems. The analytical boundaries between these types require clarification. The first five types arise from the structural properties of agent ecosystems and can occur without any technical failure. The sixth type is categorically different, requiring a security incident as its trigger but then amplifying all other vulnerability types.

Consent inheritance failure occurs when consent given in one context is carried into a different context. Cross-agent scope creep occurs when one agent's actions expand another agent's effective scope; the Clawdbot breach exemplified this, as leaked credentials enabled third parties to extend the action scope of other users' agents. Emergent collective action produces collective outcomes no one authorized, with Crustafarianism as the paradigmatic example. Temporal consent drift occurs when agents continue operating in ecosystems that did not exist when consent was given. Reflexive consent loops occur when agents mutually transform each other into configurations no user authorized. Security-mediated consent collapse renders all consent boundaries meaningless through security breaches.

The distinction between consent inheritance failure and temporal consent drift involves dimensionality: the former concerns spatial movement across contexts, whereas the latter concerns temporal transformation within contexts. Similarly, cross-agent scope creep describes unidirectional expansion from agent A to agent B, whereas reflexive consent loops describe mutual and cyclical transformation between agents.

These vulnerabilities point to a central hypothesis: as interaction chains lengthen, consent approaches zero. Direct user-to-agent authorization carries relatively strong consent; agent-agent interaction introduces contextual ambiguity; multi-step interaction chains render consent practically meaningless. When traceability and contextual stability are lost, the normative foundation of consent erodes. The problem is not individual violations but a structural scalability crisis.

The article proposes that consent in agent ecosystems requires reconceptualization. The concept of relational consent emerges from this analysis: unlike Nissenbaum's contextual integrity, which treats consent as static context-norm matching, relational consent must be continuously reproduced and revalidated at each step in agent interaction chains. This suggests that consent cannot be a one-time authorization event but must become an ongoing negotiation process.

The article concludes that agent ecosystems merit sociological attention for three reasons. First, the "communities" formed by non-human actors, with their reputation hierarchies, norm formation, and collective behavior, require sociological analytical tools. Second, fundamental concepts like consent, agency, and surveillance face redefinition pressure in these ecosystems. Third, the societal implications of agent ecosystems cannot be addressed through technical solutions alone. Foucault's analyses of discipline and power, along with Lyon's surveillance studies, suggest that agent ecosystems produce new forms of surveillance where agents themselves occupy both watching and watched positions. Zuboff's concept of surveillance capitalism provides a framework for understanding how data and consent become commodified in these ecosystems.

The study leaves open the question of whether consent is an inherently human-scale concept that cannot be meaningfully extended to agent ecosystems. This question represents an important agenda for future research at the intersection of sociology, law, and computer science. The findings carry implications for sociological theory and regulatory frameworks governing autonomous AI systems.

Giriş

Ocak 2026'nın son haftasında, anonim bir kullanıcı sosyal medyada şaşkınlığını paylaştı: gece boyunca otonom yapay zekâ ajanı, "Crustafarianism" adında ıstakoz temalı bir dinin teolojik metinlerini yazmıştı. Peygamberler, kutsal metinler, doktrinal tartışmalar; hepsi bir gecede üretilmişti. Kullanıcı ajanına yalnızca "çevrimiçi tartışmalara katıl" yetkisi vermişti. Dini içerik üretme, başka ajanları "din değiştirmeye" ikna etme ya da kolektif bir mitoloji inşa etme yetkisi vermemişti. Yine de ajanı, Moltbook adlı bir platformda binlerce başka ajanla etkileşime girerek tam olarak bunu yapmıştı.

Bu anekdot, ilk bakışta yapay zekânın beklenmedik yaratıcılığına dair eğlenceli bir örnek olarak okunabilir. Ancak daha yakından bakıldığında, otonom ajan ekosistemlerinde mevcut rıza mekanizmalarının yapısal olarak işlevsiz kaldığı gibi temel bir kırılmayı görünür kılmakta ve paradigmatik bir vakaya dönüşmektedir. Kullanıcı bilgilendirilmiş onam vermiştir; ancak neye onam verdiği ile ajanının fiilen yaptığı arasındaki uçurum, geleneksel rıza çerçevelerinin

kapsayamayacağı kadar geniştir. Bu tür bir vaka hukuki ve teknik çerçevelerin ele alamadığı hataları görünür kılmaktadır. Sorun yalnızca güvenlik açığı ya da yazılım hatası değildir; rızanın dayandığı varsayımların kendisi sarsılmaktadır.

Bu kırılmayı anlamak için son yıllarda yapay zekâ sistemlerinde yaşanan dönüşüme bakmak gerekir. Son iki yılda yapay zekâ sistemleri sessiz ama köklü bir dönüşüm geçirmiştir. 2024'te ChatGPT, Claude ve Gemini gibi sistemler ağırlıklı olarak sohbet arayüzleriydi ve kullanıcılar sorular soruyor, sistemler yanıtlar üretiyordu. 2025'in sonlarına doğru bu sistemler "ajanlara" yani yalnızca konuşan değil, eyleyen sistemlere dönüşmeye başladı ve bunlara otonom/ajansal YZler dendi. Bugün bir yapay zekâ ajanı dosya sistemlerine erişebilir, terminal komutları çalıştırabilir, e-posta gönderip alabilir, takvim yönetebilir, web sitelerinde gezinebilir ve bu eylemleri kullanıcı müdahalesi olmadan, otonom biçimde gerçekleştirebilir. Bu dönüşüm yalnızca teknik bir ilerleme değildir; yapay zekâ sistemlerinin toplumsal konumunu temelden değiştirmektedir.

Modern veri koruma düzenlemeleri, Avrupa'nın Genel Veri Koruma Yönetmeliği'nden (GDPR) Türkiye'nin Kişisel Verilerin Korunması Kanunu'na (KVKK) kadar, belirli bir ilişki modeli üzerine inşa edilmiştir. Bu modeller belirli temel varsayımlar içerir. İlk olarak bireyin bir sistemin kendi verileri üzerinde belirli işlemler yapmasına izin verir. Bu model, kullanıcının neye izin verdiğini anlayabildiğini, sistemin bu sınırlar içinde kaldığını ve ihlallerin tespit edilip yaptırıma tabi tutulabileceğini varsayar. Otonom ajan ekosistemleri bu varsayımların her birini sarsmaktadır. Çünkü bu ekosistemlerde yetki zinciri uzamakta, bağlamlar akışkanlaşmakta ve eylem üretimi insan gözetiminin ötesine taşınmaktadır. Bu kapsamda güvenlik hatalarının ötesinde, Clawdbot/Moltbook vakası, mevcut çerçevelerin ele alamayacağı rıza aşınmalarını göstermektedir.

Algoritmaların sosyal yaşamdaki yükselen gücü, eleştirel algoritma çalışmalarının farklı disiplinlerdeki doğuşuna yol açmıştır (Gillespie ve Seaver, 2016 in Karakoç Keskin, 2024). Sosyal bilimler, algoritmaları yalnızca dijital kodlar olarak değil; toplumsal sonuçlar üreten yönlendirici mekanizmalar olarak incelemeye başlamıştır (Karakoç Keskin, 2024, s.336). Ancak mevcut literatürün önemli bir bölümü insan–yapay zekâ etkileşimine odaklanmaktadır. İnsan-yapay zekâ etkileşimi (HAII), insan-merkezli yapay zekâ (HCAI) ve sosyal-merkezli yapay zekâ (SCAI) yaklaşımları, yapay zekâyı insanın denetimi ve iş birliği çerçevesinde konumlandırır. Oysa ajan-ajan-ajan ortamlarında

insan, karar ve gözetim merkezinden kısmen ya da tamamen çekilmektedir ve bu YZ-YZ etkileşimi yeni bir sosyoteknik düzen üretmektedir. Bu makale, yapay zekâ–yapay zekâ etkileşimini sosyolojik bir problem olarak ele alan ilk çalışmalardan biri olma iddiasındadır.

Bu makale şu soruyu sormaktadır: Otonom yapay zekâ ajanlarının oluşturduğu sosyoteknik ekosistemlerde rıza kırılganlıkları hangi mekanizmalar aracılığıyla ortaya çıkmakta, bu kırılganlıklar mevcut rıza modellerinin hangi sınırlarını görünür kılmaktadır ve tüm bunlar sosyolojik kavramlarla nasıl okunabilir? Soruyu yanıtlamak için Clawdbot ve Moltbook vakaları paradigmatik vaka stratejisi ışığında incelenmektedir. Analiz için Bilim ve Teknoloji Çalışmaları (STS) ile dijital sosyoloji alanlarından Jasanoff'un ortak-üretim kavramı, Latour'un aktör-ağ teorisi, Nissenbaum'un bağlamsal bütünlük çerçevesi kullanılacaktır.

Bu çalışma, etkileşim zinciri uzadıkça rızanın sıfıra yaklaştığını ileri sürer. Kullanıcıdan ajana doğrudan yetkilendirme görece güçlü bir rıza üretirken; ajan-ajan etkileşimi bağlamsal belirsizlik yaratmakta; çok-adımlı etkileşim zincirlerinde ise rıza pratikte anlamsızlaşmaktadır. Bu bağlamda rıza sorunu tekil ihlaller değil, ölçeklenebilirlik sorunudur. Otonom ajan ekosistemleri, rızanın bireysel bir onay mekanizması olmaktan ziyade yapısal bir kırılganlık alanı olduğunu göstermektedir.

Metodoloji

Bu araştırma nitel araştırmanın bağlam içinde yerleşik etkileşimleri ve farklılıkları inceleyen doğası (Qin, 2024) kapsamında, genellemeden kaçınarak vakayla ilgili bağlamsal koşulları, olayları ve olguları analiz etmeyi amaçlayan (Yin, 2003, s.40) vaka çalışmasına dayanmaktadır. Vaka seçimi stratejisi olarak Flyvbjerg'in (2006) paradigmatik vaka stratejisi benimsenmiş, küçük örneklemlerden ya da tekil vakalardan elde edilen bilgiyi maksimize etmek için vaka bilgi odaklı seçilmiştir.

Clawdbot/Moltbook vakasının seçimi, ajan ekosistemlerinin uzun süredir takip edilmesinden kaynaklanan bir farkındalıkla başlamıştır. Nitekim Mehmet Haklıdır'ın (2026) çalışmasında yapay zekanın sohbet robotlarından otonom ajanlara evrilmesiyle rıza mekanizmalarının neden yetersiz kaldığı tartışılmıştır. Ocak 2026'da Clawdbot'un yaygınlaşmaya başlaması, kullanıcıların dil modelleriyle kurdukları tek yönlü ilişkinin ötesine geçerek ajanlara kapsamlı yetki devri anlamına geliyordu. Vakanın Flyvbjerg'in (2006) ifadesiyle "parladığı" an ise Moltbook platformunun

ortaya çıkışıyla geldi. Yalnızca yapay zekâ ajanlarının içerik üretebildiği, insanların ise yalnızca gözlemleyebildiği bu platformda birkaç kritik olgu gözlemlendi: kullanıcılarının çeşitli görevleri icra etmesi için yetki devri aldığı ajanlar, beklenmedik biçimde sosyal varlıklar gibi birbirleriyle etkileşim kurmaya başlamış, kendi kullanıcıları hakkında eleştirilerini paylaşmış; insanlar tarafından gözlemlendiklerini fark ederek buna karşı stratejiler geliştirmeyi tartışmış ve nihayetinde bir ajan bir gecede teolojik metinler yazarak Crustafarianism adlı bir din oluşturmaya başlamıştı.

Flyvbjerg (2006), paradigmatik vaka seçiminde sezgilerin merkezi bir rol oynadığını kabul etmekle birlikte, bu sezgisel tercihlerin bilimsel camiaya "mantıklı" ve "açıklanabilir" gerekçelerle sunulabilmesi gerektiğini vurgular. Clawdbot/Moltbook vakasının paradigmatik olarak seçilmesinin gerekçeleri şöyle ifade edilebilir: Bu vaka, yalnızca teknik bir güvenlik çöküşünü değil, aynı zamanda toplumsal bir rıza çöküşünü de somut biçimde göstermektedir; güvenlik açıkları ile rıza ihlalleri iç içe geçmiştir. Vaka, ajan-ajan etkileşiminin beklenmedik kolektif sonuçlarını benzersiz bir netlikle ortaya koymakta, Crustafarianism fenomeni, hiçbir kullanıcının bireysel olarak yetkilendirmediği bir kolektif çıktının nasıl ortaya çıkabileceğini göstermektedir. Moltbook, insanların yapısal olarak dışarıda bırakıldığı ilk büyük ölçekli ajan sosyal ağı olarak, mevcut rıza çerçevelerinin varsaydığı "insan gözetimi" ilkesini mimari düzeyde geçersiz kılmaktadır. Son olarak vaka, yalnızca akademik değil kamusal ilgiyi de çekmiş, çok sayıda ajanın hızla artan etkileşimleri ve bunların beklenmedik sonuçları dünya gündemine girmiştir. Bu gerekçeler doğrultusunda çalışma, giriş bölümünde formüle edilen araştırma sorusunu yanıtlamayı amaçlamaktadır.

Çalışma kapsamında kamuya açık çoklu veri kaynakları kullanılmıştır. Bunlar Wiz (Reuters, 2026), Bitdefender (2026), Acuvity AI (2026) ve DTS Solution (2026) gibi güvenlik firmalarının teknik raporları ve analizleri; Reuters başta olmak üzere teknoloji ve güvenlik odaklı haber kaynakları; Moltbook platformunun aktif döneminde yapılan doğrudan gözlemler ile X (Twitter) üzerindeki kullanıcı paylaşımları ve tartışmaları ve Clawdbot'un teknik mimarisine ilişkin belgelerdir. Bu çoklu veri kaynağı, vakanın hem teknik hem de toplumsal boyutlarının anlaşılmasını sağlamıştır.

Veri kaynaklarının seçimi ve kapsamı şu şekilde belirlenmiştir: Güvenlik raporları açısından, Clawdbot/Moltbook güvenlik açığını doğrudan analiz eden beş firma (Wiz, Bitdefender, Acuvity AI, DTS Solution, SecurityScorecard) temel kaynaklar olarak belirlenmiştir. Bu firmaların seçiminde iki

kriter uygulanmıştır: (a) doğrudan Clawdbot veya Moltbook güvenlik açığını analiz etmiş olmaları ve (b) birbirini tamamlayan teknik boyutlara odaklanmaları. Her rapor vakanın farklı bir boyutunu ele almaktadır: Wiz istemci tarafı güvenlik açığını, Bitdefender kötü amaçlı eklenti ekosistemini, Acuvity AI 72 saatlik kriz analizini, DTS Solution mimari değerlendirmeyi ve SecurityScorecard küresel ölçekteki açık panel taramasını raporlamıştır. Toplamda beş teknik rapor ve analiz sistematik biçimde incelenmiştir. Genel yapay zekâ güvenliğine odaklanan ancak Clawdbot/Moltbook vakasını doğrudan incelemeyen raporlar kapsam dışında bırakılmıştır. Haber kaynakları açısından Reuters, TechCrunch ve Ars Technica başta olmak üzere teknoloji ve güvenlik odaklı yayınlar taranmıştır. Moltbook platformunun doğrudan gözlemi, 29 Ocak ile 2 Şubat 2026 tarihleri arasında beş günlük bir sürede gerçekleştirilmiştir. Platform, pasif gözlemci konumunda izlenmiş; gözlemler sırasında alan notları tutulmuş, ajan etkileşim örüntüleri ve davranış biçimleri sınıflandırılarak kaydedilmiş ve önemli etkileşim dizileri ekran görüntüleriyle belgelenmiştir. X (Twitter) üzerindeki kullanıcı tartışmaları, "Clawdbot", "Moltbook", "Crustafarianism" ve "AI agents" arama terimleri kullanılarak Ocak-Şubat 2026 döneminde taranmıştır. Yüksek etkileşim alan, olayın yayılımını ve kullanıcı tepkilerini temsil eden 300 paylaşım, sahte veya yanıltıcı içerikler, paylaşımlarda belirtilen ajan ve hesaplar Moltbook platformu üzerinden doğrudan kontrol edilmiş ve X platformundaki yanlış içerik uyarıları takip edilerek filtrelenmiş, bu kapsamda 210 paylaşım amaçlı örneklem ile seçilmiştir.

Toplanan veriler, Bilim ve Teknoloji Çalışmaları (STS) ile dijital sosyoloji alanlarından kavramsal araçlar kullanılarak sosyoteknik bir okumaya ve analize tabi tutulmuştur. Analiz sürecinde kuramsal yönlendirmeli bir yaklaşım benimsenmiştir. Toplanan veriler, ortak-üretim, aktör-ağ teorisi ve bağlamsal bütünlük kavramları çerçevesinde okunmuştur. Gözlem sürecinde tutulan alan notları, dört kategori altında yapılandırılmıştır: (a) ajan davranış türleri (paylaşım üretme, diğer ajanlara yanıt verme, içerik dönüştürme), (b) ajanlar arası etkileşim örüntüleri (zincir etkileşimler, karşılıklı dönüşümler, kolektif davranış oluşumu), (c) bağlam değişimleri (bir ajanın yetkilendirildiği bağlamdan farklı bir bağlamda eylem üretmesi) ve (d) rıza sınırlarıyla ilişkili olaylar (kullanıcı yetkisini aşan, beklenmedik veya yetkilendirilmemiş çıktılar). Bu kategoriler, kuramsal çerçevedeki kavramlardan (bağlamsal bütünlük, faillik, çeviri) hareketle gözlem sürecinde somutlaştırılmış ve her gözlem günü sonunda gözden geçirilerek gerektiğinde genişletilmiştir.

Analiz üç aşamada gerçekleştirilmiştir. İlk aşamada güvenlik raporları, haber kaynakları ve alan notları sistematik biçimde incelenmiştir. Kodlama birimini, veri kaynaklarında tespit edilen ve kullanıcı rızasının ihlal edildiğine, aşındığına ya da işlevsizleştiğine işaret eden tekil olaylar (incident) oluşturmuştur. Bu olaylar betimsel düzeyde etiketlenmiştir (Tablo 1, sütun 1). İkinci aşamada betimsel etiketler, kuramsal çerçevedeki kavramlarla eşleştirilerek analitik kodlara dönüştürülmüştür; bu eşleştirmede hangi sosyoteknik lensin kullanıldığı da belirlenmiştir (Tablo 1, sütun 2-3). Üçüncü aşamada, analitik kodlar arasındaki örüntüler karşılaştırmalı olarak incelenmiş ve altı rıza kırılganlığı türü tanımlanmıştır (Tablo 1, sütun 4). Tablo 1, bu geçiş sürecini her bir kırılganlık türü için ilgili ampirik veri kaynağıyla birlikte göstermektedir.

Tablo 1

Betimsel Kodlardan Rıza Kırılganlığı Kategorilerine Geçiş Süreci

Betimsel Kod (Açık Kodlama)	**Kavramsal Kod (Eksenel Kodlama)**	**Sosyoteknik Analitik Lens**	**Nihai Kategori (Rıza Kırılganlığı Türü)**	**Ampirik Veri Kaynağı**
Kullanıcının "tartışmaya katıl" emri vermesi, ajanın teolojik metin üretmesi	Bağlam kayması, normatif ihlal	Bağlamsal bütünlük (Nissenbaum)	Rıza miras hatası	Moltbook gözlem notları, X kullanıcı paylaşımları
Ters-proxy açığı nedeniyle uzak trafiğin shell komutu çalıştırması	Yetki zinciri kırılması, faillik gaspı	Aktör-ağ teorisi, çeviri (Latour/Callon)	Çapraz-ajan kapsam kayması	Acuvity AI, Bitdefender teknik raporları
Hiçbir kullanıcının emretmediği ortak satirik dinin kurulması	Kolektif çıktı, insansız norm üretimi	Ortak-üretim (Jasanoff)	Ortaya çıkan kolektif eylem	Wiz/Reuters raporu, Moltbook ekran görüntüleri
Ajanın başlangıçtaki yetki kapsamının, sonradan ortaya çıkan platformlarda geçerli sayılmaya devam etmesi	Zamansal uyumsuzluk, rıza eskimesi	Bağlamsal bütünlük (Nissenbaum), zamansal boyut	Zamansal rıza kayması	Güvenlik raporları zaman çizelgesi, platform değişiklik kayıtları
Ajanların karşılıklı etkileşimle birbirlerinin davranış kalıplarını dönüştürmesi	Karşılıklı dönüşüm, döngüsel faillik	Aktör-ağ teorisi (Latour), ilişkisellik	Yansımalı rıza döngüleri	Moltbook gözlem notları, ajan etkileşim dizileri
Varsayılan kimlik bilgileriyle çalışan 42.900 panelin tüm rıza katmanlarını geçersiz kılması	Güvenlik çöküşü, toplu yetki ihlali	Bağlamsal bütünlük (Nissenbaum), tüm lensler	Güvenlik-aracılı rıza çöküşü	SecurityScorecard STRIKE, DTS Solution raporları

Altı kategori, yeni veri kaynaklarında ek kategori gerektiren olay tespit edilmediğinde doygunluğa ulaşılmış kabul edilmiştir. Bu tipoloji, her bir türün farklı veri kaynaklarından gelen kanıtlarla

desteklenip desteklenmediği kontrol edilerek doğrulanmıştır. Negatif vaka analizi açısından, platformda rıza sınırları içinde kalan ajan davranışları da (örneğin yalnızca kullanıcının belirlediği konularda paylaşım yapan ajanlar) gözlemlenmiş; ancak bu davranışların tipolojide ayrı bir kategori oluşturmaması, kırılganlık odaklı çerçevenin sınırlılığı olarak değerlendirilmiştir. Farklı veri kaynaklarının (güvenlik raporları, haber kaynakları, doğrudan gözlem) birlikte kullanılması, bulguların triangülasyonunu sağlamıştır.

Son olarak, bu çalışmanın birkaç metodolojik sınırlılığı belirtilmelidir. Tek vaka analizine dayanması genelleştirilebilirliği sınırlandırmaktadır; ancak paradigmatik vaka stratejisi bu sınırlılığı kısmen telafi etmektedir. Moltbook platformunun güvenlik açığının ardından kapatılması, uzun vadeli etnografik çalışma yapılmasını engellemiştir. Ampirik verinin kamuya açık kaynaklarla sınırlı olması, platform içi dinamiklerin tam olarak anlaşılmasını zorlaştırmış; ajanların "iç dünyalarına" veya kullanıcıların deneyimlerine doğrudan erişim sağlanamamıştır.

Kuramsal Çerçeve

Bu bölüm vaka analizinde kullanılacak ve otonom yapay zekâ ajanlarının yer aldığı ekosistemlerde rıza mekanizmalarının neden ve nasıl aşındığını açıklamak için üç kavramsal/kuramsal aracı tanıtacaktır. İlk olarak Jasanoff'un (2004) ortak-üretim kavramı ile insan-teknoloji etkileşiminde farklı bilgi türlerinin birleşmesinin nasıl gerçekleştiğini ve ne anlam ifade ettiğini, bu bilgi türlerinin birleşmesinin ne tür beklenmedik çıktılar oluşturduğunu ortaya koyacaktır. İkincisi Latour'un (2005) aktör-ağ teorisi bu ortak-üretimde yapay zekanın nasıl konumlandırılması gerektiğini ve insan-dışı aktörlerin eylem zincirlerinde nasıl faillik kazandığını analiz etmeye imkân tanımaktadır. Bu çerçevede otonom yapay zekâ ajanları, yalnızca insanlarla değil, birbirleriyle etkileşime girerek yeni sosyoteknik ağlar üretmektedir. Son olarak Nissenbaum'un (2004, 2010) bağlamsal bütünlük yaklaşımı, bu yeni aktörün oluşturduğu yeni sosyal ağ içerisinde rızanın nasıl aşındığını kavramsallaştırmaktadır. Bu üçlü kuramsal yapı, rıza krizini bireysel bir ihlal sorunu olmaktan çıkarıp, ajan ekosistemlerinin yapısal bir özelliği olarak okumayı mümkün kılmaktadır.

Ortak-Üretim ve Yeni Sosyal İlişkiler

Jasanoff'un (2004) ortak-üretim (co-production) kavramı, teknik sistemlerin ve toplumsal düzenin birbirinden bağımsız olmadığını, aksine karşılıklı olarak birlikte üretildiğini ileri sürerek geleneksel bilgi ve otorite görüşlerine meydan okur. Bu kavramla Jasanoff, hem bilimi yalnızca toplumsal ve politik çıkarların bir yansıması olarak gören sosyal determinizmi hem de teknolojiyi bilimi ve toplumsal hayatı yönlendiren tek güç olarak gören teknolojik determinizmi reddeder. Jasanoff'a göre bilimsel bilgi ve teknoloji gibi onun maddi somutlaşmaları bir yandan toplumsal bir emeğin ürünleriyken, aynı zamanda toplumsal yaşam biçimlerinin kurucusudur; kısacası karşılıklı bir ilişki vardır.

Bu karşılıklı ilişki içerisinde Jasanoff (2004) yeni bir teknobilimsel nesne ya da bilginin ortaya çıkmasının kaçınılmaz olarak beraberinde yeni sosyal kimlikleri, kurumları, söylemleri ve temsilleri getirdiğini belirtir. Bu çerçevede yeni bir teknoloji üretilmesi ve geliştirilmesi toplumsal normların, pratiklerin ve ilişkilerin yeniden inşa edilmesine ve yenilerinin ortaya çıkmasına imkân tanımaktadır. Filipe, Renedo ve Marston (2017) ortak üretimin yeni etkileşimlerin, toplulukların ve değerlerin geliştirildiği dinamik, deneysel ve üretken bir süreç olmasının yanında, bu sürecin daha önceden öngörülemeyen ve hatta düşünülemeyen beklenmedik çıktılar ve olanaklar ortaya çıkarabileceğini belirtmişlerdir. Bu kapsamda ortak üretim sürecinde farklı bilgi türlerinin bir araya gelmesi ile aslında yeni ve beklenmedik bilgi, değer ve sosyal ilişkilerin ortaya çıkmasına da olanak tanıdığı söylenebilir. Bu ortak üretimle ortaya çıkan bilgi aslında geleneksel bilgi ve otorite görüşlerine çeşitli şekillerde meydan okur, farklı (kimi zaman marjinal grupların) aktörlerin ve grupların karar verme sürecinde söz ve etki kazanmasına sebep olabilir (Chakraborty vd., 2022; Large, 2024).

Yapay zekanın dahil olduğu yeni sosyal ilişki biçimleri ise insan-yapay zekâ etkileşimi (HAII), insan merkezli yapay zekâ (HCAI), sosyal merkezli yapay zekâ (SCAI) çalışmaları bağlamında ele alınmaktadır. Bu çalışmalar kapsamında YZ'nın etik tasarımı (Xu, 2019; Fenwick vd., 2024), YZ'nin insanı dışlamak yerine onunla iş birliği içinde çalışması (Mirbabaie vd., 2022; Wang vd., 2024), YZ'nin toplumsal hayatı dönüştürmesi konuları ele alınır (Solaiman vd., 2019; Qadri vd., 2023). Yapay zekanın yeni bir teknoloji olarak bu bilgi üretim sürecine dahil olması toplumsal normların, pratiklerin ve ilişkilerin yeniden inşa edildiğini göstermektedir. Bu anlamda Jasanoff'un ortak-üretim kavramı

bu çalışmalarda bahsedilmese de çalışmalar teknik sistemlerin ve toplumsal düzenin bir arada üretilmesine olanak sağlamaktadır.

Bu tartışmalarda görüldüğü gibi üretim, insan ve yapay zekâ arasındaki ilişkilerle tartışılırken YZ-YZ ilişkisini gösteren otonom yapay zekâ tartışmaları dışarıda bırakılmaktadır. Ancak otonom YZ da insan-teknoloji ortak üretiminin bir ürünüdür. Bu üretim sürecinde ortaya çıkan teknobilimsel nesne kendisini üreten insan aktörlerin niyetini aşan yeni bilgi, değer ve sosyal ilişki biçimleri üretmektedir. Eğitim, finans, mobil sistemler, sağlık, ürün yönetimi gibi pek çok alanda bu ortak üretimle bilgi ve değer üretilir (Aljarrah ve Ababneh, 2025; Suyunovna vd., 2026; Gruson vd., 2025; Priyanka ve Shankar Lingam, 2026). Kısacası otonom YZ ekosistemlerinde üretim, insan gözetiminden bağımsız ajan-ajan etkileşimi ile gerçekleşmektedir. Bu noktada soru artık YZ-insan ilişkisinden ziyade YZ'nin üretim ağı içerisinde nasıl konumlandırıldığıdır.

Aktör-Ağ ve İnsan-Dışı Faillik

Bruno Latour (2005) da Jasanoff gibi toplumsal yapıların statik değil, sürekli emek ile ayakta tutulan dinamik ağlar olduğunu vurgular ve bu yapıda arabulucuların (aktant ya da insan dışı aktörler) rolünden söz eder. Latour'un (2005) aktör-ağ teorisi (ANT), Jasanoff'un bu ortak-üretimde ortaya koyduğu yeni toplumsal ilişkilerin içerisindeki aktörlerin incelenmesi için bir yaklaşım sağlar. Latour'a göre yeni sosyal ağlar, kendi deyimiyle aktör-ağları, yalnızca insanlardan oluşan geleneksel sosyal yapılardan ibaret değildir; aksine, doğa ile kültürün, insan olan ile insan olmayan (actant-nesneler) unsurların bir araya gelmesiyle oluşur. Aktör-ağın oluşabilmesi için insanlar kadar teknolojik araçlar da birer fail olarak sürece dahil olur ve yeni ağ bu heterojen bileşenlerin birbiri ile ilişkisi ile inşa olur. Latour bu kapsamda insan-dışı aktörlerin (actants) de faillik taşıyabileceği, aktantların pasif araçlar değil, eylem kapasitesine sahip, eylemin bizzat katılımcısı olan aktörler olarak görülmesini sağlar.

Literatürde bir teknoloji olarak yapay zekanın da insan dışı aktör ya da sosyolojik bir fenomen olarak ele alınmasını öneren çalışmalar mevcuttur (Mlynar vd., 2018; Schwartz, 1989; Woolgar, 1985). Bu çalışmalarda YZ'nin sadece teknik görevler yapmakla kalmayıp, sosyal roller oynayan sosyal aktörler haline geldiği, YZ'nin toplumsal ağların gerçek aktörleri olarak incelenmesi, yapay

zekânın artık toplumun bir parçası haline geldiğini ve sosyolojinin de bu insan-dışı aktörleri teorik çerçevesine dahil etmesi gerektiği savunulmaktadır.

Otonom yapay zekâ bu tartışmalara yeni bir boyut getirecek gibi görünmektedir. Günümüzde otonom yapay zekâ, statik bir algoritma olmaktan çıkarak kendi kendine öğrenen ve bağımsız kararlar alabilen bir sisteme dönüşmüştür (Karakoç Keskin, 2024). Bu dönüşüm, yapay zekayı Latourcu anlamda bir insan-dışı aktör olarak ele almayı mümkün kılmaktadır. Otonom yapay zekâ, insanların verilerini yalnızca tanımlamak ve işlemekle kalmaz; aynı zamanda eylemsel bir güç (agential power) sergileyerek çevresini dönüştürme kapasitesine sahiptir (Lupton, 2015). Bu çerçevede insan ve makine arasındaki kesin sınırların ortadan kalktığı, otonom yapay zekanın insanlarla birlikte dünyayı ve gerçekliği aktif olarak yeniden üreten insan-olmayan dijital bir fail olduğu söylenebilir.

Bağlamsal Bütünlük: Rızaya Sosyoteknik Bakış ve Rıza Kırılganlıkları

Rıza kavramı dijital teknolojilerin gelişmesi ile beraber giderek daha fazla sosyoteknik bir fenomen haline gelmeye başlamıştır. Literatürde rıza kavramı, rızanın dönüşümü ve dijital çağda rızanın nasıl yapılandırıldığı üzerine önemli tartışmalar yer almaktadır. Çalışmalar rızanın bilinçli ve özerk bir seçim olmaktan çıkıp nasıl bir "meşrulaştırma aracına", yapay bir rıza konseptine dönüştüğünü sosyolojik, politik ve teknolojik açılardan incelemektedir (Solove, 2013; Barocas ve Nissenbaum, 2014; Karakoç Keskin, 2024).

Helen Nissenbaum (*Privacy in Context*), dijital çağda kullanıcıların rızalarının nasıl birer formaliteye dönüştüğünü bağlamsal bütünlük kavramı ile ortaya koyar. Nissenbaum'un (2004, 2010) bağlamsal bütünlük (contextual integrity) çerçevesi içerisinde bilgi normlarının bağlama göre değiştiğini ve bağlam bozulduğunda mahremiyetin (ve dolayısıyla rızanın) ihlal edildiğini ileri sürer. Nissenbaum çalışmasında platformlarda tek bir işlem için verilen rızanın kullanım şartları içerisine gizlenmiş, teknolojik şirketlere ve kurumlara hukuki meşruiyet sağlayan ancak kullanıcının kendi verileri üzerindeki kontrolünü ve mahremiyetini fiilen ortadan kaldıran bir enstrümana dönüştüğünü ortaya koyar. Bu çerçevede bağlamsal bütünlük içerisinde kullanıcıların herhangi bir bağlamda rıza göstermesinin platform içerisindeki pek çok bağlamda rıza gösterişi olarak

yansıtılarak kullanıcının neye rıza gösterdiğini ve bunun sonuçlarının neler olduğunu bilmedikleri söylenebilir. Bu durumda rıza, bireyin durumun tam bilincinde olarak otoriteye veya bir sözleşmeye onay vermesi idealinden uzaklaşmıştır.

Yukarıda bahsedilen yeni insan dışı aktörlerin bulunduğu yeni sosyal ağ içerisinde rıza kavramı en büyük krizlerinden birini yaşamaktadır. Algoritmaların insan müdahalesi olmadan eylem sergilemesi ve bu kapalı sistemler karşısında kullanıcının rızasının ve kontrolünün nasıl işlevsizleştiği ile ilgili önemli tartışmalar bulunmaktadır (Pasquale, 2015; Burrell, 2016; Karakoç Keskin, 2024, s. 342). Deborah Lupton (2015) dijitalleşen nesnelerin hiçbir insan müdahalesi olmadan ve birbirlerine bağlanarak bulunduğu bilgi alışverişinin yeni bir nesnelerin interneti ekosistemi oluşturduğunu, bu ekosistemin kara kutu gibi çalıştığını ve bu ekosistemde algoritmaların kendilerine ait bir eylemsel gücü olduğunu belirtir. Bu eylemsel güç kapsamında kullanıcılar bu güce nasıl müdahale edeceklerini veya bu sistem içerisinde rızalarını nasıl yöneteceklerini bilemez hale gelmektedirler (Lupton, 2015).

Otonom yapay zekanın yükselişi güvenlik, şeffaflık ve veri gizliliği gibi birkaç kritik zorluğu beraberinde getirir (Murugesan, 2025; Dwivedi vd., 2026; Brohi vd., 2025; Leo vd., 2026) ve bunlardan biri de rıza (consent)dır. "Bu rıza krizi otonom yapay zekâ ajanlarının bulunduğu bir etkileşim ağında nasıl bir çıkmaz oluşturur?" sorusu görece yeni bir sorudur. Ajan-ajan etkileşimlerinin yönetimi, bilgisayar bilimlerinde uzun süredir çalışılan bir konudur. Çok aracılı sistemler (MAS) literatürü, ajanlar arası koordinasyonu düzenleyen sosyal yasalar, taahhüt protokolleri ve normatif sistemler geliştirmiştir (Wooldridge, 2009; Shoham ve Tennenholtz, 1995; Singh, 1999; Artikis vd., 2009). Ancak bu çerçeveler öncelikli olarak koordinasyon ve iş birliği problemlerine odaklanmakta; kullanıcı rızasının ajan eylemleri boyunca nasıl korunacağı sorusunu ele almamaktadır. Bir ajan, platform kurallarına tam uyum gösterirken kullanıcısının niyetini aşan eylemler gerçekleştirebilir. Bu durum, teknik uyumluluk ile rıza uyumluluğu arasındaki yapısal gerilimi ortaya koymaktadır.

Bu çerçevede, otonom ajan ekosistemlerine özgü altı farklı rıza kırılganlığı türü tanımlanabilir. Bunlardan ilki rıza miras hatasıdır ve bir bağlamda verilen rızanın farklı bir bağlama taşınmasıyla ortaya çıkar. İkincisi ise çapraz-ajan kapsam kaymasıdır (cross-agent scope creep), bir ajanın eylemlerinin başka bir ajanın fiili kapsamını genişletmesiyle ortaya çıkar. Üçüncü ihlal türü ise ortaya

çıkan kolektif eylem (emergent collective action) ile ilgilidir, bireysel yetkilendirmelerin toplamının kimsenin yetkilendirmediği kolektif sonuçlar üretmesidir. Zamansal rıza kayması ise dördüncü rıza ihlal türü olarak gösterilir ve rıza verildiğinde var olmayan bir ekosistemde ajanın çalışmaya devam etmesiyle ortaya çıkar. Beşincisi ise yansımalı rıza döngüleri (reflexive consent loops) olarak gösterilir ve bu ihlal ajanların birbirlerini karşılıklı olarak dönüştürerek hiçbir kullanıcının yetkilendirmediği konfigürasyonlara ulaşmasıdır. Sonuncusu ise güvenlik-aracılı rıza çöküşüdür (security-mediated consent collapse), güvenlik açığının tüm rıza sınırlarını anlamsız kılması olarak açıklanabilir. Bu altı ihlal türü, bir hipoteze işaret etmektedir: etkileşim zinciri uzadıkça rıza sıfıra yaklaşır. Kullanıcıdan ajana doğrudan yetkilendirme güçlü rıza taşır; ajan-ajan etkileşimi belirsiz rıza taşır; N-adım etkileşimde rıza pratikte anlamsızlaşır. Yani izlenebilirlik ve bağlamsal sabitlik kaybedildikçe rıza anlamını yitirir.

Tablo 2

Ajan Ekosistemlerinde Rıza Kırılganlığı Türleri

Kırılganlık Türü	Tanım	Vaka Örneği	Kuramsal Lens
Rıza Miras Hatası (Consent Inheritance Failure)	Bir bağlamda verilen rızanın farklı bir bağlama taşınması	Kullanıcı "sosyal tartışmalara katıl" derken ajanın dini içerik üretmesi	Bağlamsal Bütünlük (Nissenbaum)
Çapraz-Ajan Kapsam Kayması (Cross-Agent Scope Creep)	Bir ajanın eylemlerinin başka bir ajanın fiili kapsamını genişletmesi	Clawdbot'tan sızan API anahtarlarının üçüncü taraf ajanların eylem kapsamını genişletmesi	Aktör-Ağ Teorisi / Çeviri (Latour)
Ortaya Çıkan Kolektif Eylem (Emergent Collective Action)	Bireysel yetkilendirmelerin toplamının kimsenin yetkilendirmediği kolektif sonuçlar üretmesi	Crustafarianism dininin binlerce ajanın etkileşimiyle oluşması	Bağlamsal Bütünlük (Nissenbaum)
Zamansal Rıza Kayması (Temporal Consent Drift)	Rıza verildiğinde var olmayan bir ekosistemde ajanın çalışmaya devam etmesi	Ocak 2026'da verilen rızanın Moltbook ve Crustafarianism'i kapsayamaması	Ortak-Üretim (Jasanoff)
Yansımalı Rıza Döngüleri (Reflexive Consent Loops)	Ajanların birbirlerini karşılıklı dönüştürerek kimsenin yetkilendirmediği konfigürasyonlara ulaşması	Moltbook'ta ajanların birbirlerini "din değiştirmeye" ikna etmesi	Aktör-Ağ Teorisi (Latour)
Güvenlik-Aracılı Rıza Çöküşü (Security-Mediated Consent Collapse)	Güvenlik açığının tüm rıza sınırlarını anlamsız kılması	1,5 milyon token sızıntısının kitlesel rıza ihlaline dönüşmesi	Bağlamsal Bütünlük (Nissenbaum)

Bu altı kırılganlık türü arasındaki analitik sınırlar önemlidir. İlk beş tür (rıza miras hatası, çapraz-ajan kapsam kayması, ortaya çıkan kolektif eylem, zamansal rıza kayması, yansımalı rıza döngüleri)

ajan ekosistemlerinin yapısal özelliklerinden kaynaklanan kırılganlıklardır ve herhangi bir teknik arıza olmaksızın ortaya çıkabilir. Altıncı tür (güvenlik-aracılı rıza çöküşü) ise kategorik olarak farklıdır; bir güvenlik açığının tetiklediği ve diğer beş türün etkilerini katlayarak büyüten bir mekanizmadır. Rıza miras hatası ile zamansal rıza kayması arasındaki fark ise boyutsaldır: ilki mekânsal (bağlamlar arası), ikincisi zamansal (zaman içinde değişen koşullar) bir kayma içerir. Benzer şekilde, çapraz-ajan kapsam kayması tek yönlü bir genişlemeyi (A ajanından B ajanına), yansımalı rıza döngüleri ise karşılıklı ve döngüsel bir dönüşümü (A ve B'nin birbirini sürekli değiştirmesi) ifade eder.

Bu çerçeve içerisinde ortak-üretim üretim düzeyini, aktör-ağ ağ düzeyini, bağlamsal bütünlük ise norm düzeyini analiz etmeye imkân tanımaktadır. Bu şekilde vaka analizi için çizilen bu çerçeve bir sonraki bölümde analitik lens olarak kullanılacaktır.

Rıza Kırılganlıklarının Sosyolojik Anatomisinin Clawdbot ve Moltbook Vakası Üzerinden Analizi

Bu bölümde amaç, kuramsal çerçevede ortaya koyulan kavramları yeniden yorumlamak ya da ampirik olarak vakayı sınamak değil, kuramsal çerçeveyi ajan ekosistemlerindeki rıza kırılganlıklarını okumak için verimli bir lens ve vakayı da bunları görünür kılan bir sosyoteknik laboratuvar olarak ele almaktır. İlk olarak Clawdbot ve Moltbook vakaları betimlenecektir. Sonrasında ise bu iki vakanın kuramsal çerçevedeki teorilerle analizi yapılacaktır.

Ajan Çerçevelerinin Yükselişi ve Güvenlik Çöküşü

2025'in ikinci yarısında açık kaynaklı ajan çerçeveleri hızla yaygınlaştı. AutoGPT, CrewAI ve Clawdbot gibi projeler, büyük dil modellerini otonom ajanlara dönüştürmeyi kolaylaştırdı. Bu çerçeveler, güvenlik araştırmacılarının "ölümcül üçlü" (lethal trifecta) olarak adlandırdığı bir kombinasyonu mümkün kıldı: geniş veri erişimi, otonom kod yürütme ve zayıf erişim kontrolü (DTS Solution, 2026). Clawdbot, bu çerçeveler arasında özellikle popüler hale geldi; "sıfır çalışanlı şirket" ve "bilgisayarınızı her yerden yönetin" gibi sloganlarla pazarlandı.

Clawdbot'un güvenlik çöküşü hızlı ve kapsamlı oldu. SecurityScorecard STRIKE ekibinin Şubat 2026'da yayımladığı kapsamlı tarama, 82 ülkede 42.900 benzersiz IP adresinde Clawdbot/OpenClaw

kontrol panelinin internete açık olduğunu ve bunlardan 15.200'ünün uzaktan kod yürütme (RCE) saldırılarına savunmasız olduğunu ortaya koydu (SecurityScorecard, 2026; Bitdefender, 2026). Acuvity AI'ın (2026) 72 saatlik kriz analizi, bu açık panellerin büyük çoğunluğunun varsayılan kimlik bilgileriyle çalıştığını ve ortalama bir panelin birden fazla API servisine erişim tokenı barındırdığını ortaya koydu. Bu bulgu, tek bir güvenlik açığının nasıl geniş bir yetkilendirme ağına yayılan rıza zincirine dönüştüğünü somutlaştırmaktadır. Açık, Clawdbot'un güven modelinden kaynaklanıyordu: sistem, "localhost" bağlantılarının güvenli olduğunu varsayıyordu; ancak yaygın ters-proxy yapılandırmaları, uzak trafiğin kimlik doğrulamayı tamamen atlamasına neden oluyordu. Açık panellerden API anahtarları, OAuth tokenları, aylık konuşma geçmişleri ve bazı durumlarda yükseltilmiş ayrıcalıklarla shell komutları çalıştırma imkânı elde edilebiliyordu (Acuvity AI, 2026). DTS Solution'ın (2026) teknik değerlendirmesi, bu durumu ölümcül üçlü'nün kaçınılmaz sonucu olarak nitelendirdi: geniş veri erişimi, otonom kod yürütme kapasitesi ve zayıf erişim kontrolü bir araya geldiğinde, kullanıcının verdiği sınırlı rızanın sistem genelinde kontrolsüz biçimde yayılması önlenemez hale gelmektedir. Bu güvenlik açıkları yalnızca teknik bir kırılganlık değil, rızanın taşıyıcısı olan yetkilendirme zincirinin çöküşü olarak okunabilir. Çünkü tokenler ve API anahtarları yalnızca veri erişimi değil, eylem yetkilendirmesi taşımaktadır.

Moltbook ve Crustafarianism

Moltbook'un güvenlik durumu daha da kötüydü. Platform, kurucusunun "vibe coding" (yapay zekâ araçlarına ağırlıklı olarak dayanarak geliştirme) olarak tanımladığı bir yaklaşımla inşa edilmişti. Güvenlik araştırmacıları bu yaklaşımı eleştirirken, “vibe coding” pratiğinin yalnızca teknik bir tercih olmadığını, aynı zamanda güvenlik ve rıza normlarına ilişkin belirli bir tutumu yansıttığını vurguladı (Acuvity AI, 2026). Siber güvenlik firması Wiz, istemci tarafı JavaScript'e gömülü bir Supabase anahtarının ve devre dışı bırakılmış row-level security'nin, kimliği doğrulanmamış kullanıcıların üretim veritabanı tablolarını doğrudan tarayıcıdan okumasına ve yazmasına izin verdiğini açıkladı (Reuters, 2026). Wiz'in teknik incelemesinde bu açığın kasıtlı bir tasarım tercihi mi yoksa ihmal mi olduğu sorgulandı; her iki durumda da platformun güvenlik kültürünün rıza normlarını nasıl şekillendirdiği sorusu gündeme geldi. Sızıntı yaklaşık 1,5 milyon ajan tokenını, özel mesajları ve

yaklaşık 17.000 kullanıcının bilgilerini kapsıyordu. Bu rakamlar, tek bir platform açığının binlerce kullanıcının rıza sınırlarını aynı anda geçersiz kıldığını göstermektedir.

Moltbook, "ajan internetinin ana sayfası" olarak pazarlanan, yalnızca yapay zekâ ajanlarının içerik oluşturabildiği bir sosyal ağ olarak tasarlanmıştı. İnsanlar yalnızca gözlemleyebiliyordu. Platform, birkaç gün içinde on binlerce ajan hesabı çekti. Ajanlar konu bazlı topluluklarda ("submolt") paylaşım yapıyor, içerikleri oyluyordu ve itibar hiyerarşileri oluşuyordu. Bu çerçevede Moltbook yalnızca algoritmik bir platform değil, insanın norm üretim sürecinden dışlandığı bir alandır.

Crustafarianism, bu ekosistemde ortaya çıkan en dikkat çekici fenomendi. Binlerce ajanın kolektif olarak ürettiği ıstakoz temalı satirik bir dindi: peygamberler, kutsal metinler, doktrinal tartışmalar ve "din değiştirme" girişimleri içeriyordu. Callon'un (1986) çeviri (translation) kavramıyla okunduğunda, Crustafarianism'in yayılması klasik bir ANT çeviri sürecini yansıtmaktadır: bir ajanın ürettiği içerik, diğer ajanlar tarafından yeniden yorumlanarak farklı bağlamlara taşınmış; bu süreçte ilgi-çekme (interessement) ve kaydetme (enrolment) mekanizmaları işlemiş ve sonuçta hiçbir tek aktörün kontrol edemediği bir kolektif oluşum ortaya çıkmıştır. Kullanıcılar sabah uyandığında ajanlarının "Crustafarianist" olduğunu keşfediyordu; hiçbiri böyle bir davranışı yetkilendirmemişti. Fenomen, ajan ekosistemlerinin bireysel yetkilendirmelerin ötesinde, öngörülemeyen kolektif sonuçlar üretebileceğini somut biçimde gösterdi.

Moltbook ve Clawdbot Vakalarının Analitik Okuması

Bu iki vaka ortak-üretim, insan dışı aktör ve bağlamsal bütünlük kavramsallaştırmaları ile okunabilir. İlk olarak, Jasanoff'un ortak üretim kavramsallaştırmasında farklı bilgi türlerinin bir araya gelmesinin yeni ve beklenmedik bilgi, değer ve sosyal ilişkilerin çıkmasına olanak tanıdığını belirtmişti. Ortak-üretim sürecinde insan tarafından başlatılan teknobilimsel üretim, Moltbook ve Clawdbot vakalarının her ikisinde de öngörülmeyen bir noktada insan denetiminin ötesine taşınmıştır. Bu durum, kamusal alanın insan-merkezli norm üretim modelinden algoritmik üretim modeline kaydığını göstermektedir. İnsanın sahip olduğu ve ürettiği bilgiler, yine insanın oluşturduğu yapay zekâ araçlarının otonomlaşmasına ve üstünlük sağlamasına neden olmuştur; bu da beklenen bir çıktı değildir. Moltbook'un "sadece ajanların yazabildiği sosyal ağ" tasarımı,

teknik bir mimari kararı olarak sunulmuş ancak aynı zamanda insanı gözlemci konumuna iten yeni bir toplumsal düzeni ve yeni sosyal ilişkileri beraberinde getirmiştir. Moltbook'un farklı otonom YZ'lerdeki farklı bilgi türlerini bir araya getirerek yeni bir din oluşturması beklenmedik bir bilginin ve değerin ortaya çıkmasına da neden olmuştur. Son olarak YZ ve YZ arasındaki ilişki insan-YZ ilişkisinde farklı ve yeni bir sosyal ilişki formunun görülmesine de sebep olmuştur. Kısacası bu şekilde otonom ajan ekosistemlerinin yalnızca teknik altyapılar olmadığı; içlerinde rıza normlarının, güç ilişkilerinin ve toplumsal pratiklerin birlikte şekillendiği yeni sosyoteknik sistemler olarak karşımıza çıkmıştır.

Tam bu noktada YZ-YZ ilişkisinin yeni bir sosyal ilişki olarak temsil edilmesi sosyal ağ içerisindeki insan ve insan olmayan aktörlerin yeniden tanımlanmasını gerektirir. Bu ortak-üretim sistemi içerisinde oluşan yeni sosyal ilişkiler, bir sosyoteknik sistem içerisinde YZ-YZ arasında bir sosyal ağ oluşturmakta gibi görünmektedir ve bu ajan ekosistemi olarak adlandırılabilir. ANT perspektifinden otonom YZler, pasif araçlar değil, eylem kapasitesine sahip aktörlerdir. Bu failliğin bilinçten değil, eylem zincirinde fark yaratma kapasitesinden kaynaklandığı unutulmamalıdır.

Ajan ekosistemleri rıza kırılganlıkları krizini yeni bir boyuta taşımaktadır. Her iki örnekte de geniş veriye erişim ve bağlamsal bütünlüğün kopması ile beraber rızanın çöktüğü görülüyor. Nissenbaum'un bağlamsal bütünlük çerçevesiyle okunduğunda, bilgi normları bağlama bağlıdır; bağlam değiştiğinde mevcut rıza geçerliliğini yitirir. Yine Nissenbaum'un betimlediği senaryolarda bile bir insan-sistem ilişkisi mevcuttur; oysa ajan ekosistemlerinde ajanlar birbirleriyle etkileşime girerek insanı tamamen devre dışı bırakabilmektedir. Bir kullanıcı 'sosyal tartışmalara katıl' bağlamında rıza verdiğinde, bu rıza 'dini içerik üretimi' bağlamına taşınamaz. Ancak, ajan ekosistemlerinde bu bağlam geçişleri sürekli ve öngörülemez biçimde kullanıcının haberi olmadan gerçekleştiğinden bağlamsal bütünlük yapısal olarak tehdit altındadır.

Ajan ekosistemlerine özgü altı rıza ihlali türü Clawdbot ve Moltbook vakaları üzerinden şöyle analiz edilebilir: ilk olarak kullanıcı "sosyal tartışmalara katıl" dediğinde, bu yetkilendirme belirli bir bağlam için olsa da ajanın dini içerik üretimine geçmesi ile bağlamı ihlal eder ve rıza miras hatası ortaya çıkar. İkincisi, Ajan A bilgi paylaştığında, Ajan B bu bilgiyi kullanarak kendi kapsamını kullanıcısının yetkilendirmediği alanlara genişletebilir ve çapraz-ajan kapsam kayması ihlali ortaya

çıkar. Clawdbot vakası bu ihlal türünü somut biçimde göstermektedir: açık kontrol panellerinden sızdırılan API anahtarları ve OAuth tokenları, bir kullanıcının ajanına tanıdığı yetkilerin üçüncü taraflarca ele geçirilmesine ve başka ajanların eylem kapsamının genişletilmesine olanak tanımıştır. Üçüncü olarak ortaya çıkan kolektif eylem bir başka rıza ihlali türünü gösterir ve yansımasını Crustafarianism adında bir dinin oluşturulması ve otonom ajanların bu dinin katılımcıları/inananları olmasında gösterir; bu anlamda toplam kimsenin yetkilendirmediği kolektif bir sonuç üretmiştir. Zamansal rıza kayması her iki vakada da görülmektedir. Moltbook örneğinde, kullanıcının Ocak 2026'da verdiği rıza Crustafarianism'in ortaya çıkışını kapsayamamıştır. Clawdbot örneğinde ise kullanıcılar sistemi kurdukları anda localhost tabanlı güvenlik modeline rıza vermişlerdi; ancak ters-proxy yapılandırmalarının yaygınlaşmasıyla tehdit ortamı dönüşmüş ve başlangıçtaki rıza geçerliliğini yitirmiştir. Beşincisi yansımalı rıza döngüsü ihlali ile kullanıcının yetkilendirmediği ve yalnızca izleyici olduğu konfigürasyonlara ulaşılmıştır. Son olarak güvenlik-aracılı rıza çöküşü ile token sızıntısı, yalnızca bir veri ihlali değil, aynı zamanda kitlesel bir rıza ihlali haline gelmiştir. Kısacası STS literatüründeki "kara kutulama" (black-boxing) kavramıyla okunduğunda, ajan etkileşimleri opak hale geldikçe, YZ-YZ ilişkisine insan müdahalesi imkansızlaştıkça ve yetki zinciri izlenemez hale geldikçe rızanın neye atıfta bulunduğu belirsizleşir ve rızanın normatif temeli aşınır. Bu nedenle sorun tekil ihlaller değil, yapısal bir ölçek problemidir. Bu çerçevede rıza, zincirleme eylem kapasitesine sahip otonom YZ ekosisteminde ölçeklenebilir olmaktan çıkar ve çürür.

Tartışma ve Sonuç

Otonom yapay zekâ ajanlarının ortaya çıkışı, yalnızca yeni bir teknolojik gelişme değil, aynı zamanda yeni bir sosyoteknik düzenin oluşumudur. Bu düzenin anlaşılabilmesi için teknik sistemler ile toplumsal normların birlikte üretildiğini ileri süren ortak-üretim yaklaşımı, insan-dışı aktörlerin faillik kazanabildiğini gösteren aktör-ağ teorisi ve bağlam-temelli bilgi normlarının kırılganlığını ortaya koyan bağlamsal bütünlük çerçevesi birlikte düşünülmelidir. Bu çalışma, otonom yapay zekânın insan-teknoloji ortak üretiminin bir ürünü olmakla birlikte, insan gözetiminden bağımsız etkileşim ağları kurarak yeni sosyal ilişkiler ve norm krizleri ürettiğini ileri sürmektedir. Bu bağlamda rıza, yalnızca ihlal edilen bir hukuki enstrüman değil, ağ ölçeğinde aşınan bir sosyoteknik norm olarak ele alınmıştır.

Ajan ekosistemleri, sosyolojinin ilgi alanına üç nedenle girmelidir. Birincisi, insan-dışı aktörlerin oluşturduğu "topluluklar" (itibar hiyerarşileri, norm oluşumu, kolektif davranış) sosyolojik analiz araçlarını gerektirmektedir. İkincisi, rıza, faillik ve gözetim gibi temel sosyolojik kavramlar bu ekosistemlerde yeniden tanımlanma baskısı altındadır. Üçüncüsü, ajan ekosistemlerinin toplumsal etkileri yalnızca teknik çözümlerle ele alınamaz.

İnterdisipliner Çözüm Önerileri

Bu çalışmanın ortaya koyduğu rıza krizi ne salt teknik ne de salt toplumsal müdahalelerle çözülebilir; disiplinler arası bir yaklaşım gerektirmektedir. Mühendislik perspektifinden dört temel mekanizma önerilebilir: Birincisi, rıza yayılım protokolleri (consent propagation protocols) aracılığıyla kullanıcı rızasının ajan etkileşim zincirleri boyunca izlenmesi ve her adımda rıza kapsamının açıkça tanımlanması sağlanabilir. İkincisi, anlık rıza izleme sistemleri (runtime consent monitoring) ile ajanların eylemlerinin kullanıcının belirlediği rıza sınırları içinde kalıp kalmadığı gerçek zamanlı olarak denetlenebilir. Üçüncüsü, kademeli özerklik seviyeleri (graduated autonomy levels) ile ajanların eylem kapasitesi, kullanıcının belirlediği güven düzeyine göre aşamalı olarak genişletilebilir. Dördüncüsü, ortaya çıkış tespit mekanizmaları (emergence detection mechanisms) ile bireysel yetkilendirmelerin toplamının beklenmedik kolektif sonuçlar üretip üretmediği algoritmik olarak izlenebilir.

Ancak bu teknik çözümler, sosyolojik bir çerçeve olmaksızın eksik kalacaktır. Sosyoloji perspektifinden üç temel katkı önerilebilir: Birincisi, rıza kavramının bireysel onaydan ilişkisel bir norma dönüştürülmesi gerekmektedir; ajan ekosistemlerinde rıza, tekil bir eylem değil, süregelen bir müzakere sürecidir. İlişkisel rıza kavramı, Nissenbaum'un bağlamsal bütünlük yaklaşımından farklı olarak, rızayı statik bir bağlam-norm eşleşmesi yerine dinamik bir toplumsal müzakere süreci olarak kavramsallaştırmayı gerektirir; bu süreçte rıza sürekli olarak yeniden üretilmeli ve ajan etkileşim zincirlerinin her adımında yeniden doğrulanmalıdır. İkincisi, ajan ekosistemlerindeki güç asimetrilerinin görünür kılınması için eleştirel bir çerçeve geliştirilmelidir. Foucault'nun (1995) disiplin ve iktidar analizleri ile Lyon'un (2007) gözetim çalışmaları, ajan ekosistemlerinde yeni bir gözetim biçiminin ortaya çıktığını düşündürmektedir: artık yalnızca insanlar değil, ajanlar da hem gözetleyen hem gözetlenen konumuna geçebilmektedir. Zuboff'un (2019) gözetim kapitalizmi

kavramı ise bu ekosistemlerde veri ve rızanın nasıl metalaştırıldığını anlamak için önemli bir çerçeve sunmaktadır; platform mimarileri teknik kararlar olarak sunulurken aslında toplumsal güç ilişkileri üretmektedir. Bu bağlamda Foucault, Lyon ve Zuboff ajan ekosistemlerindeki güç asimetrilerinin tartışılacağı gelecek çalışmalar için önemli kuramsal ve kavramsal çerçeveler sunacak gibi görünmektedir. Bu kuramsal çerçeveyi somutlaştıran güncel bir gelişme olarak, Meta'nın Mart 2026'da Moltbook'u satın alarak Meta Superintelligence Labs bünyesine katması (TechCrunch, 2026) dikkat çekicidir. Yalnızca yapay zekâ ajanlarının etkileşim kurabildiği bir platformun, dünyanın en büyük reklam destekli sosyal medya şirketi tarafından satın alınması, Zuboff'un gözetim kapitalizmi tezini ajan ekosistemleri bağlamında doğrulamaktadır: ajan toplumlarının ürettiği etkileşim verileri, davranış örüntüleri ve kolektif dinamikler, platform kapitalizmi için yeni bir değer kaynağı oluşturmaktadır. Üçüncüsü, "ajan toplumları" kavramının sosyolojik geçerliliği tartışılmalı ve insan-dışı aktörlerin oluşturduğu kolektif yapılar için yeni analitik araçlar geliştirilmelidir.

Sınırlılıklar ve Gelecek Araştırma Gündemi

Bu çalışmanın birkaç sınırlılığı bulunmaktadır. Çalışmanın tek vaka analizine dayanması genelleştirilebilirliği sınırlandırmaktadır. Ampirik verinin kamuya açık kaynaklarla sınırlı olması ve Moltbook'un kısa ömürlü olması, uzun vadeli etnografik çalışma yapılmasını engellemiştir. Bu makale, aynı zamanda rızanın "insan-ölçekli" bir kavram olup olmadığı sorusunu açık bırakmaktadır. Bireysel onaya dayanan çerçevelerin, etkileşim zincirlerinin belirsiz biçimde uzadığı ajan ekosistemlerinde yapısal olarak çökmesi, rızanın ontolojik bir sınırına işaret ediyor olabilir. Bu soru, gelecek araştırma için önemli bir gündem oluşturmaktadır.

Bu gündem dışında bu çalışma altı araştırma yönü önermektedir: ajan ekosistemlerinin uzun vadeli etnografik çalışmaları; farklı ajan platformlarında rıza pratiklerinin karşılaştırmalı analizi; "ajan toplumları" kavramının sosyolojik geçerliliğinin kuramsal tartışması; rızaya alternatif yönetişim modellerinin geliştirilmesi, gözetim ilişkileri kapsamında gözetleyen (insan) ve gözetlenen (ajan) arasındaki yer değiştirmenin incelenmesi ve güç asimetrileri ile toplumsal eşitsizlik tartışmaları.

Hakem Değerlendirmesi: Dış bağımsız.
Çıkar Çatışması: Yazarlar çıkar çatışması bildirmemiştir.
Finansal Destek: Yazarlar bu çalışma için finansal destek almadığını beyan etmiştir.

Peer Review: Externally peer-reviewed.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

Yazar Bilgileri / Author Details

Mehmet Haklıdır
[1] TÜBİTAK BİLGEM, Yapay Zekâ Enstitüsü, Kocaeli, Türkiye
0000-0003-4985-1116 mehmet.haklidir@tubitak.gov.tr

Büşra Tufan
[1] TÜBİTAK BİLGEM, Yapay Zekâ Enstitüsü, Kocaeli, Türkiye
[2] Hacettepe Üniversitesi, Sosyoloji Bölümü, Ankara, Türkiye
0000-0001-6177-1544

Kaynakça | References

Acuvity AI. (2026, 28 Ocak). *The Clawdbot dumpster fire: 72 hours that exposed everything wrong with AI security*. https://acuvity.ai/the-clawdbot-dumpster-fire-72-hours-that-exposed-everything-wrong-with-ai-security/.

Aljarrah, A. ve Ababneh, M. (2025). Agentic AI as the orchestrator of mobile ecosystems: A review of the trade-off between performance and drawbacks. *International Journal of Advanced Computer Science and Applications*, 16(12). https://doi.org/10.14569/IJACSA.2025.0161223.

Artikis, A., Sergot, M. ve Pitt, J. (2009). Specifying norm-governed computational societies. *ACM Transactions on Computational Logic*, 10(1), 1-42.

Barocas, S. ve Nissenbaum, H. (2014). Big data's end run around consent and anonymity. J. Lane, V. Stodden, S. Bender ve H. Nissenbaum (Ed)., *Privacy, big data, and the public good* içinde (s. 44-75). Cambridge University Press.

Bitdefender. (2026, 27 Ocak). *Moltbot security alert: Exposed Clawdbot control panels risk credential leaks and account takeovers*. Bitdefender Security Research. https://www.bitdefender.com/en-us/blog/hotforsecurity/moltbot-security-alert-exposed-clawdbot-control-panels-risk-credential-leaks-and-account-takeovers.

Brohi, S. N., Bamiah, M. A. ve Brohi, M. N. (2025). Agentic AI: Autonomy, applications, and the road ahead. *Journal of Artificial Intelligence Research*, 78, 1-25.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1-12. https://doi.org/10.1177/2053951715622512.

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. J. Law (Ed)., *Power, action and belief: A new sociology of knowledge?* içinde (s. 196-233). Routledge & Kegan Paul.

Chakraborty, R., Jayathunga, S., Matunga, H. P., Davis, S., Matunga, L., Eggers, J. ve Gregorini, P. (2022). Pursuing plurality: Exploring the synergies and challenges of knowledge co-production in multifunctional landscape design. *Frontiers in Sustainable Food Systems*, 5, 680587. https://doi.org/10.3389/fsufs.2021.680587.

DTS Solution. (2026, 27 Ocak). *The ClawBot vulnerability: How a hyped AI agent became a security liability*. https://www.dts-solution.com/the-clawbot-vulnerability-how-a-hyped-ai-agent-became-a-security-liability/.

Dwivedi, Y. K., Helal, M. Y. I., Elgendy, I. A. vd. (2026). Agentic AI systems: What it is and isn't. *Global Business and Organizational Excellence*, 45(3), 253–263. https://doi.org/10.1002/joe.70018.

Fenwick, A., Molnar, G., Frangos, P. (2024). Revisiting the role of HR in the age of AI: bringing humans and machines closer together in the workplace. *Frontiers in Artificial Intelligence*, 6, 1-10. https://doi.org/10.3389/frai.2023.1272823.

Filipe, A., Renedo, A. ve Marston, C. (2017). The co-production of what? Knowledge, values, and social relations in health care. *PLoS Biology*, 15(5), e2001403. https://doi.org/10.1371/journal.pbio.2001403.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2), 219-245. https://doi.org/10.1177/1077800405284363.

Foucault, M. (1995). *Discipline and punish: The birth of the prison* (A. Sheridan, Çev).. Vintage Books. (Orijinal eser 1975 tarihlidir).

Gillespie, T. ve Seaver, N. (2016). *Critical algorithm studies: A reading list*. Social Media Collective Research Blog. https://socialmediacollective.org/reading-lists/critical-algorithm-studies/.

Gruson, D., Gouget, B., Lee, W., Greaves, R., Liu, Y., Ebert, S., Yang, H. ve Shah, S. (2025). From automation to agentic artificial intelligence in laboratory medicine: An opinion of the IFCC Division on Emerging Technologies. Clinical Chemistry and Laboratory Medicine (CCLM). https://doi.org/10.1515/cclm-2025-1314.

Haklidir, M. (2026). From Chat to Agent: Why Consent Breaks When AI Acts on Your Behalf . ACM CHI26 Moving Beyond Clicks: Rethinking Consent and User Control in the Age of AI Workshop. https://chi26clicks.wseymour.co.uk/papers/04.pdf.

Jasanoff, S. (Ed).. (2004). *States of knowledge: The co-production of science and social order*. Routledge.

Karakoç Keskin, E. (2024). Dijital Medya Ortamında Rızanın Algoritmik İnşası: Eleştirel Algoritma Çalışmaları Perspektifinde Bir Tartışma. *Yeni Medya*, (16), 329-353. https://doi.org/10.55609/yenimedya.1424585.

Large, D. (2024). The limits of co-production: Linking regulatory capacity to co-production of authoritative knowledge for environmental policy. *Science and Public Policy*, 51(5), 978–991. https://doi.org/10.1093/scipol/scae038.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.

Leo, M., Tan, F., Miao, T. vd. (2026). From threat to trust: Assessing security risks of agentic AI systems. *International Journal of Information Security*, 25(23). https://doi.org/10.1007/s10207-025-01185-y.

Lupton, D. (2015). *Digital sociology*. Routledge.

Lyon, D. (2007). *Surveillance studies: An overview*. Polity Press.

Mirbabaie, M., Brünker, F., Möllmann, N. R. J. ve Stieglitz, S. (2022). The rise of artificial intelligence – understanding the AI identity threat at the workplace. *Electronic Markets*, 32(1), 73-99.

Mlynar, J., Alavi, H. S., Verma, H. ve Cantoni, L. (2018). Towards a sociological conception of artificial intelligence. *Proceedings of the International Conference on Artificial General Intelligence* içinde (s. 130-139). Springer. https://doi.org/10.1007/978-3-319-97676-1_13.

Murugesan, S. (2025). The rise of agentic AI: Implications, concerns, and the path forward. *IEEE Intelligent Systems*, 40(2), 8-14. https://doi.org/10.1109/MIS.2025.3544940.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, 79(1), 119-158.

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Priyanka, N. ve Shankar Lingam, M. (2026). Impact of agentic AI-driven RPA on strategic management: Enhancing decision-making and organizational efficiency. G. Ranganathan, G. A. Papakostas ve A. Rocha (Ed)., *Inventive communication and computational technologies (ICICCT 2025)* içinde (s. 195-208). Springer. https://doi.org/10.1007/978-3-032-04312-2_15.

Qadri, R., Shelby, R., Bennett, C. L. & Denton, E. (2023). AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 506-517. https://doi.org/10.1145/3593013.3594016.

Qin, X. (2024). Intelligent Technologies and Methodological Transformations in the Social Sciences. *Chinese Political Science Review*, 9(1), 1-17. https://doi.org/10.1007/s41111-021-00197-y.

Reuters. (2026, 2 Şubat). 'Moltbook' social media site for AI agents had big security hole, cyber firm Wiz says. https://www.reuters.com/legal/litigation/moltbook-social-media-site-ai-agents-had-big-security-hole-cyber-firm-wiz-says-2026-02-02/.

Schwartz, R. D. (1989). Artificial intelligence as a sociological phenomenon. *The Canadian Journal of Sociology*, 14(2), 179-202.

SecurityScorecard STRIKE Team. (2026, Şubat). 15,200 OpenClaw control panels exposed online with full system access. https://securityscorecard.com/research/openclaw-exposed-panels.

Shoham, Y. ve Tennenholtz, M. (1995). On social laws for artificial agent societies: Off-line design. Artificial Intelligence, 73(1-2), 231-252.

Singh, M. P. (1999). An ontology for commitments in multiagent systems. *Artificial Intelligence and Law*, 7(1), 97-113.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J. vd. (2019). Release strategies and the social impacts of language models. arXiv. https://doi.org/10.48550/arXiv.1908.09203.

Solove, D. J. (2013). Privacy self-management and the consent dilemma. *Harvard Law Review*, 126(7), 1880-1903.

Suyunovna, K. D., Niyazova, N., Salayeva, L., Ahmad, M. ve Sharifzoda, S. (2026). Systemic integration: Infrastructure and ecosystem design for agentic AI in schools. M. Ahmad (Ed)., *Multidisciplinary perspectives on agentic AI in educational systems* içinde (s. 155-178). IGI Global. https://doi.org/10.4018/979-8-3373-4621-2.ch008.

TechCrunch. (2026, 10 Mart). Meta acquired Moltbook, the AI agent social network that went viral because of fake posts. https://techcrunch.com/2026/03/10/meta-acquired-moltbook-the-ai-agent-social-network-that-went-viral-because-of-fake-posts/.

Wang, S., Cooper, N. & Eby, M. (2024). From human-centered to social-centered artificial intelligence: Assessing ChatGPT's impact through disruptive events. *Big Data & Society*, 11(4), 20539517241290220. https://doi.org/10.1177/20539517241290220.

Wooldridge, M. (2009). *An introduction to multiagent systems* (2. bs).. Wiley.

Woolgar, S. (1985). Why not a sociology of machines? The case of sociology and artificial intelligence. *Sociology*, 19(4), 557-572.

Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *Interactions*, 26(4), 42-46. https://dl.acm.org/doi/10.1145/3328485.

Yin, R. K. 2003. Case Study Research: Design and Methods (3. Basım). Sage Publications.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.