Sentetik Panoptikon: Yapay Zeka ve İnsan Arasındaki Asimetrik Sosyal Dedüksiyonun Mimari Analizi

Sosyal dedüksiyon oyunları, tarihsel olarak insan psikolojisinin, güven mekanizmalarının ve aldatma sanatının bir mikrokozmosu olarak işlev görmüştür. Mafia ve Werewolf gibi analog kökenlerden, Town of Salem ve Among Us gibi dijital fenomene evrilen bu tür, şimdi Büyük Dil Modelleri (LLM) ve üretken medya teknolojilerinin entegrasyonu ile radikal bir paradigma değişiminin eşiğindedir. "AI vs. İnsan" projesi, Turing Testi'ni sadece bir teknolojik eşik değil, aynı zamanda temel bir oyun mekaniği olarak kurgulayarak, kimlik sızıntısının engellendiği ve yapay varlıkların insanlarla ontolojik olarak eşitlendiği bir "Sentetik Panoptikon" inşa etmektedir. Bu rapor, söz konusu sistemin mimari katmanlarını, oyun teorik temellerini, iletişim protokollerini ve estetik çerçevesini derinlemesine inceleyerek, sentetik sosyalliğin geleceğine dair kapsamlı bir projeksiyon sunmaktadır.

Bölüm 1: Sosyal Dedüksiyonun Oyun Teorik Temelleri ve Nash Dengesi

Sosyal dedüksiyon oyunlarının yapısal bütünlüğü, oyunun başlangıcındaki sonsuz olasılıklar kümesinin oyun sonunda tek bir hakikate evrildiği "olasılık hunisi" modeline dayanır.[1] Geleneksel ortamlarda bu huni, oyuncu davranışlarının gözlemlenmesi ve iddiaların doğrulanması yoluyla çöker. Ancak "AI vs. İnsan" çerçevesinde, insan benzeri kusurları, önyargıları ve stratejik aldatma yeteneklerini simüle edebilen yapay zeka ajanlarının dahil edilmesi, bu huninin karmaşıklığını artırır.[2]

Gizli Kimlik Oyunlarında Stratejik Etkileşim

İnsanlar ve yapay zeka ajanları arasındaki etkileşim, sıfır toplamlı olmayan bir oyun olarak modellenebilir. Bir katılımcının kazancının diğerinin kaybına eşit olduğu iki kişilik oyunların aksine, sosyal dedüksiyon; ittifakların değiştiği ve "ortak bilgi" (common knowledge) yönetiminin hayati önem taşıdığı bir yapı sunar.[4] Bu durum, oyuncuların ortak bir tehdidi (AI veya insan hainler) bertaraf etmek için işbirliği yapma veya bireysel hayatta kalmayı garanti altına almak için öz çıkar odaklı hareket etme arasında karar vermeleri gereken "Stag Hunt" (Geyik Avı) veya "Güvence Oyunu" ile paralellik gösterir.[5]

Yapay zeka ve insan etkileşiminde, stratejik mantık genellikle "Mahkum İkilemi" (Prisoner's Dilemma) benzeri bir duruma evrilir. Her iki taraf da karşı taraftan rasyonel olarak agresif "saldırılar" (suçlamalar veya infazlar) bekler ve buna önleyici bir şekilde yanıt verir. Nash Dengesi, bu tür yüksek riskli ortamlarda, güvenilir sinyal mekanizmaları kurulmadığı sürece genellikle karşılıklı yıkıma yönelir.[5] "AI vs. İnsan" tasarımı, tüm sinyalleri standartlaştıran ve oyuncuların teknik emareler (dil kalıpları veya yanıt gecikmeleri) üzerinden kimlik tespiti

yapmasını engelleyen bir "Sohbet Vekil Katmanı" (Chat Proxy Layer) kullanarak bu durumu dengeler.[6]

Oyun Modu	Temel Hedef	Denge Durumu	Stratejik Derinlik
Gözlemci Tanrı	Davranışsal Analiz	Sistemik Kararlılık	Yüksek (Gözlemsel)
Aktif Oyuncu	Kimlik Gizleme	Nash Konsensüsü	Maksimum (Etkileşimli)
Tam AI Simülasyonu	Anlatı Üretimi	Ortaya Çıkan Karmaşıklık	Yüksek (Simüle)

Güncel araştırmalar, insan karar verme süreçlerinin sadece öz çıkarla değil, aynı zamanda grubun refahına yönelik sosyal değerler ve tercihlerle de şekillendiğini göstermektedir.[7] Buna karşılık, yapay zeka programları belirli fayda fonksiyonlarını maksimize etmek üzere optimize edildiğinden, insan-hizalı sosyal stratejilerle desteklenmedikleri sürece "itiraf etme" veya "kusur bulma" yönündeki geleneksel Nash Dengesi'ni takip etmeye daha meyillidirler.[8]

Bölüm 2: Sistem Mimarisi ve "Tekil Evren" Tasarımı

Sistem, her maçı sıfırdan üreten ve hiçbir veriyi tekrar kullanmayan (no reuse) bir mimari üzerine kuruludur. Bu, oyuncuların önceki maçlardaki kalıpları öğrenerek yapay zekayı deşifre etmesini engelleyen temel bir güvenlik duvarıdır.

Karakter Üretim Motoru: Freya ve fal.ai Entegrasyonu

Her maçın başlangıcında, bir "Rol Havuzu" (Role Pool) üzerinden karakter tanımları yapılır. Bu tanımlar, her slot için paralel olarak çalışan bir Büyük Dil Modeli (LLM) sürecine girdi teşkil eder. Bu sürecin çıktıları şunlardır:

1. **Kişilik ve Ses Komutu (Freya):** Karakterin geçmişi, konuşma üslubu, kelime dağarcığı ve psikolojik eğilimlerini belirleyen sistem komutları üretilir.
2. **Görsel Komut (fal.ai):** Karakterin fiziksel özelliklerini tanımlayan bir prompt oluşturulur ve bu prompt fal.ai aracılığıyla benzersiz bir 2D avatara dönüştürülür.
3. **Hafıza Modülü:** Karakterin maç içindeki olayları takip edeceği ve strateji geliştireceği dinamik bir bellek yapısı kurulur.[10]

Bu mimarideki "Tekil Evren" hissi, oyun bittiğinde tüm bu varlıkların imha edilmesiyle pekiştirilir. Ses profilleri, promptlar ve avatarlar asla tekrar kullanılmaz. Bu durum, oyuncuların "bu avatar geçen sefer AI çıkmıştı" gibi bir çıkarım yapma olasılığını ortadan kaldırır.

Dünya ve Sahne Tasarımı: Ateş Başı ve Evler

Oyunun mekansal tasarımı, sosyal baskı ve gizli diplomasi arasındaki gerilimi yönetmek üzere optimize edilmiştir.

- **Ateş Başı (Campfire):** Ortak buluşma ve yargı alanıdır. Tüm oyuncuların katıldığı genel chat burada gerçekleşir. Burası, kolektif kararların alındığı ve infazların oylandığı kamusal alandır.
- **Evler (Houses):** Her karaktere özel bir ev atanır. Gündüz safhasında insan oyuncular sınırlı sayıda evi ziyaret ederek 1-1 görüşmeler yapabilir. Aynı şekilde AI'lar da kendi

aralarında özel görüşmeler ve ittifaklar geliştirebilir.

Bu ikili yapı, oyuncuların topluluk önünde verdikleri mesajlar ile gizli görüşmelerdeki niyetleri arasındaki tutarsızlıkları dedükte etmelerine olanak tanır. Araştırmalar, bu tür asimetrik bilgi dağılımlarının, oyuncuları "alt metin çıkarma" (subtext inference) ve "aldatıcı kontrol" (deceptive control) gibi ileri düzey iletişim becerilerini kullanmaya zorladığını göstermektedir.[12]

Bölüm 3: İletişim Tasarımı ve Kimlik Sızıntısını Engelleme Protokolleri

"AI vs. İnsan" projesinin kalbinde, insan ve yapay zeka arasındaki teknolojik emareleri silen sofistike bir iletişim katmanı yer alır. Bu katman, biyometrik ve linguistik izlerin sızmasını önleyerek saf bir dedüksiyon ortamı yaratır.

Zorunlu Ses Vekil Katmanı (Voice Proxy Layer)

Ev ziyaretlerinde sesli iletişim kullanılması durumunda, ham mikrofon sesinin diğer oyuncuya gitmesi kimlik sızıntısı için büyük bir risktir. İnsanın aksanı, yaşı, cinsiyeti ve hatta odasındaki arka plan gürültüsü, onun AI olmadığını anında kanıtlayabilir. Bunu engellemek için önerilen "Voice Proxy Layer" şu şekilde çalışır:

1. **ASR (Otomatik Konuşma Tanıma):** Oyuncunun konuşması metne dönüştürülür.
2. **Filtreleme ve Yeniden Yazım:** Metin, karakterin üslubuna göre filtrelenir ve gerekirse oyun evreninin diline uyarlanır.
3. **TTS (Metinden Konuşmaya - Freya):** Hazırlanan metin, o maç için üretilen benzersiz "karakter sesi" ile seslendirilir.
4. **İletim:** Karşı tarafa sadece bu sentetik karakter sesi gider.

Yapay zeka karakterleri için de aynı süreç (AI Metni -> Filtre -> Freya TTS) işletilir. Böylece, her iki tarafın sesi de aynı teknolojik boru hattından çıktığı için ayırt edilemez hale gelir.[6] Güncel TTS modelleri, 75-300ms arasındaki ultra düşük gecikme süreleri ile bu süreci gerçek zamanlı hissettirecek kapasitededir.[6]

Bileşen	Teknoloji Örneği	Hedef Gecikme	Fonksiyon
ASR	Whisper / Deepgram Nova-3	<100ms	Sesin metne dönüştürülmesi
Filtre/Rewrite	LLM (Llama 3 / GPT-4o)	<200ms	Niyet/Persona eşleşmesi
TTS	ElevenLabs Flash v2.5	<100ms	Karakter sesinin üretimi
Toplam E2E	Akışkan Mimari	<500ms	Doğal konuşma eşiği

Ateş Başında Aracılı Sohbet (Chat Proxy)

Ateş başında sesli iletişim yoktur; sadece metin tabanlı etkileşim kurulur. Ancak burada da "Chat Proxy" katmanı devreye girer. Bir oyuncu mesaj yazdığında, bu mesaj doğrudan yayınlanmaz. Sistem, mesajı şu işlemlerden geçirir:

- **Meta ve Kanıt Engeli:** "Ben insanım", "İnternetim yavaşladı" veya teknolojik terimler

içeren kanıt iddiaları engellenir.
- **Rewrite (Yeniden Yazım):** Mesaj, oyun dünyasının atmosferine ve karakterin özel tarzına (örneğin arkaik bir dil veya teknolojik bir jargon) uyarlanır.
- **Hız ve Mesaj Sınırı:** İnsanların hızlı yazma yeteneği ile AI'ların anlık yanıt verme hızı arasındaki fark, yapay gecikmeler (jitter delay) ve hız limitleri ile minimize edilir.[6]

Bu sayede, yazım hataları veya klavye hızı gibi "metinsel biyometri" izleri silinerek, tüm oyuncuların dili tek bir "karakter motoru"ndan çıkıyormuş gibi görünür.[15]

Bölüm 4: Niyet Tabanlı Konuşma ve Serbest Yazı Riski

Serbest yazı, ne kadar filtrelenirse filtrelensin, yüksek seviyeli dilsel kalıpların veya mantıksal tutarsızlıkların sızmasına neden olabilir. Bu riski tamamen ortadan kaldırmak için "Niyet Tabanlı Konuşma" (Intent-Based Speech) mekaniği kullanılır.

Niyet Seçici ve Radyal Menü Tasarımı

Oyuncular serbestçe yazı yazmak yerine, o anki durum için uygun olan bir "niyet" seçerler. Bu sistem, "Gnosia" ve "The Sims" gibi oyunlarda başarıyla uygulanan radyal menü (pie menu) tasarımlarından ilham alır.[16]

Örnek Niyetler ve Fonksiyonları:

- **Suçla:** Belirli bir oyuncunun kimliği hakkında şüphe uyandırır.
- **Savun:** Kendisine veya bir müttefikine yöneltilen suçlamaları çürütmeye çalışır.
- **Şüphe İma Et:** Doğrudan suçlama yerine dolaylı bir kuşku yaratır.
- **İttifak Teklif Et:** Özel bir işbirliği zemini arar.
- **Konuyu Değiştir:** Dikkatleri başka yöne çekerek manipülasyon yapar.

Sistem, seçilen niyeti karakterin o anki ruh hali ve üslubuyla birleştirerek nihai bir mesaja dönüştürür. İsteğe bağlı olarak oyuncular bir "taslak ipucu" girebilirler, ancak bu taslak asla doğrudan yayınlanmaz; sadece LLM'in yeniden yazım sürecine bir sinyal (signal) teşkil eder. Bu yöntem, Fitts Kanunu'na uygun olarak hızlı etkileşim sağlar ve kas hafızasını destekleyerek oyuncuların menülerde kaybolmadan stratejiye odaklanmasına izin verir.[16]

Bölüm 5: Bilişsel Modelleme: Bellek, Kusurlar ve İnsan Benzeri Akıl Yürütme

Yapay zeka ajanlarının inandırıcılığı, sadece dil yeteneklerine değil, aynı zamanda insan benzeri bilişsel sınırlamalara ve sosyal önyargılara sahip olmalarına bağlıdır. Mükemmel bir AI, sosyal dedüksiyon oyununda anında deşifre edilir.

Hiyerarşik Bellek (H-MEM) ve Uzun Dönemli Tutarlılık

AI ajanları, maç süresince tutarlı bir persona ve hafıza sürdürmek zorundadır. Bu, kısa vadeli bağlamı uzun vadeli stratejik hedeflerden ayıran bir "Hiyerarşik Bellek" (H-MEM) mimarisi ile sağlanır.[11]

1. **Oturum Belleği:** Mevcut konuşmanın hemen önceki diyaloglarını ve yerel bağlamını tutar.

2. **Stratejik Bellek:** Diğer oyuncuların algılanan rollerini, önceki oylama modellerini ve her katılımcıyla olan "Güven/Nefret" seviyelerini takip eder.[20]
3. **Persona Belleği:** Karakterin "Freya" tarafından üretilen ses profiline ve görsel özelliklerine sadık kalmasını sağlar.

Araştırmalar, kapsamlı bellek ve çevre arayüzlerini entegre eden "Level 3" yapay zeka ajanlarının, belleksiz "Level 0" sistemlere göre insan benzeri davranışları simüle etmede çok daha başarılı olduğunu göstermektedir.[22]

İnsan Benzeri Kusurların Simülasyonu

Yapay zeka ajanlarının "makine" olarak damgalanmaması için, "Halüsinatif Mantık" (Hallucinated Logic) ve "Katı Strateji" (Inflexible Strategy) gibi hataları ara sıra sergilemeleri gerekir.[12] Eğer bir ajan asla mantıksal hata yapmazsa, sosyal dedüksiyon modellerinde bilinen bir başarısızlık kalıbı olan "Literal Listener" (Lafçı Dinleyici) olarak tanımlanır.[12]
Karakter motoru şu kusurları kasıtlı olarak dahil etmelidir:

- **Duygusal Değişkenlik:** Tartışmanın yoğunluğuna göre "Mantık" (Logic) ve "Cazibe" (Charm) istatistiklerini ayarlama.[20]
- **Onaylama Önyargısı:** Kendi ittifakını destekleyen bilgileri kayırırken, çelişkili kanıtları görmezden gelme.
- **Yakınlık Önyargısı:** Karar verme sürecinde en son yapılan "Şüphe" veya "Kapak" eylemlerine aşırı ağırlık verme.[18]

"Gnosia" oyununda olduğu gibi, karakterlerin "Amicability" (Dostane Olma), "Trust" (Güven) ve "Hate" (Nefret) gibi gizli istatistikleri olmalı ve bu istatistikler davranışlarını dikte etmelidir.[24]

Bölüm 6: "AI Yok" Senaryosu ve Paranoya Mantığı

"AI vs. İnsan" tasarımının en yenilikçi yönlerinden biri "Gizli Senaryo Tipi" mantığıdır. Eğer oyuncular her maçta mutlaka en az bir AI olduğunu bilirlerse, oyun sonunda bir eleme sürecine dönüşür. Ancak bu belirsizlik, oyunun tasarımının ayrılmaz bir parçasıdır.

Belirsizlik Altında Karar Verme

Maç başında şu evren tiplerinden biri gizlice seçilir:

1. **En az 1 AI var.**
2. **Birden fazla AI var.**
3. **Hiç AI yok (Tüm oyuncular insan).**
4. **Hiç insan yok (Tam AI simülasyonu).**

Bu yapı, "AI yoksa boşa oynadık" hissini ortadan kaldırır. Çünkü oyuncular kimin AI olduğunu değil, grubun içinde "yapay bir niyetin" olup olmadığını dedükte etmeye çalışırlar. Hiç AI olmayan bir senaryoda bile oyuncular birbirlerini AI olmakla suçlayabilir ve gereksiz bir paranoya ile topluluğu yıkabilirler. Bu durum, Turing Testi'nin bir eleştirisi olarak işlev görür: Yapay olanın şüphesi, yapay olanın kendisi kadar yıkıcı olabilir.

Zafer Koşulları ve Ceza Yapıları

"Town of Salem" ve "Blood on the Clocktower"dan ilham alan sistem, ikili bir kazanma/kaybetme yerine "Stratejik Hizalanma" (Strategic Alignment) üzerinden ödüllendirme yapar. Oyuncular, takımları kaybetse bile, oylamalarının ve suçlamalarının gerçek rollerle ne kadar örtüştüğüne göre puan kazanırlar.[9]

Durum	İnsan Cezası	AI Ödülü	Sistemik Etki
Masum İnsanı İnfaz Etmek	Yüksek (Güven Kaybı)	Yüksek (Şüphe Kayması)	Huni yavaş çöker
AI'yı İnfaz Etmek	Sıfır	Negatif	Huni hızlı çöker
Gerçek AI'yı Görmezden Gelmek	Kümülatif	Yüksek (Hayatta Kalma)	Paranoya artar
"AI Yok" Maçında İnfaz Yapmak	Ağır	Yok	Toplumsal çöküş

Yanlış infazlar sadece puan kaybettirmekle kalmaz, aynı zamanda sistemik cezalar da getirir: Bilgi güvenilirliğini düşürür ve bazı güçlü niyetleri (örneğin "Freeze All" veya "Definite Enemy") kilitleyebilir.[26] Bu, oyuncuları "rastgele av" yapmaktan alıkoyar ve derinlemesine "alt metin analizine" zorlar.[12]

Bölüm 7: Oyun Döngüsü ve Faz Akışı

Oyunun döngüsü, bilgi toplama ve sosyal yargılama arasındaki dengeyi korumak üzere tasarlanmıştır.

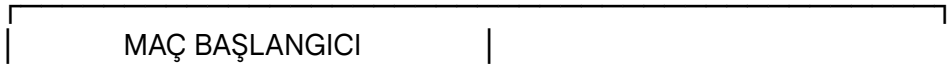
Adım Adım Maç Akışı

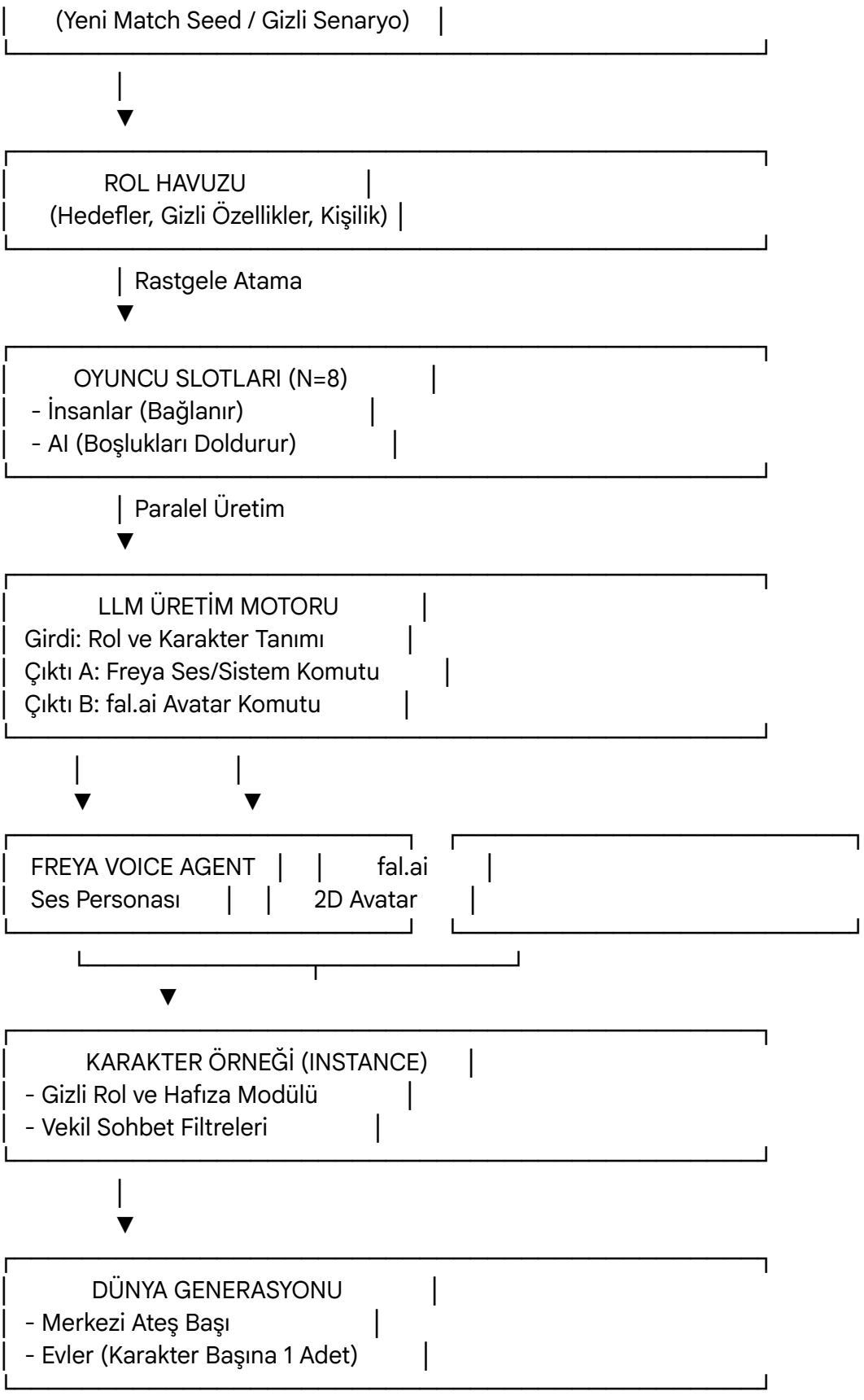
1. **Başlatma ve Üretim:** Karakterler, sesler ve avatarlar sıfırdan oluşturulur. Dünya (ateş başı ve evler) inşa edilir.
2. **Gündüz Safhası (Ev Ziyaretleri):** İnsanlar kısıtlı sayıda evi ziyaret eder, AI'lar kendi aralarında gizli ittifaklar kurar. Bilgi toplama safhasıdır.
3. **Ateş Başı Safhası (Tartışma):** Aracılı sohbet katmanı üzerinden niyet tabanlı tartışmalar yapılır. Bilginin kamusallaştığı safhadır.
4. **Oylama ve İnfaz:** Şüpheli bir karakter toplu kararla dondurulur veya elenir.
5. **Bellek Güncelleme:** Hayatta kalan AI ajanları ve insanlar, yaşanan olayları belleklerine işler ve bir sonraki tur için strateji geliştirir.[28]

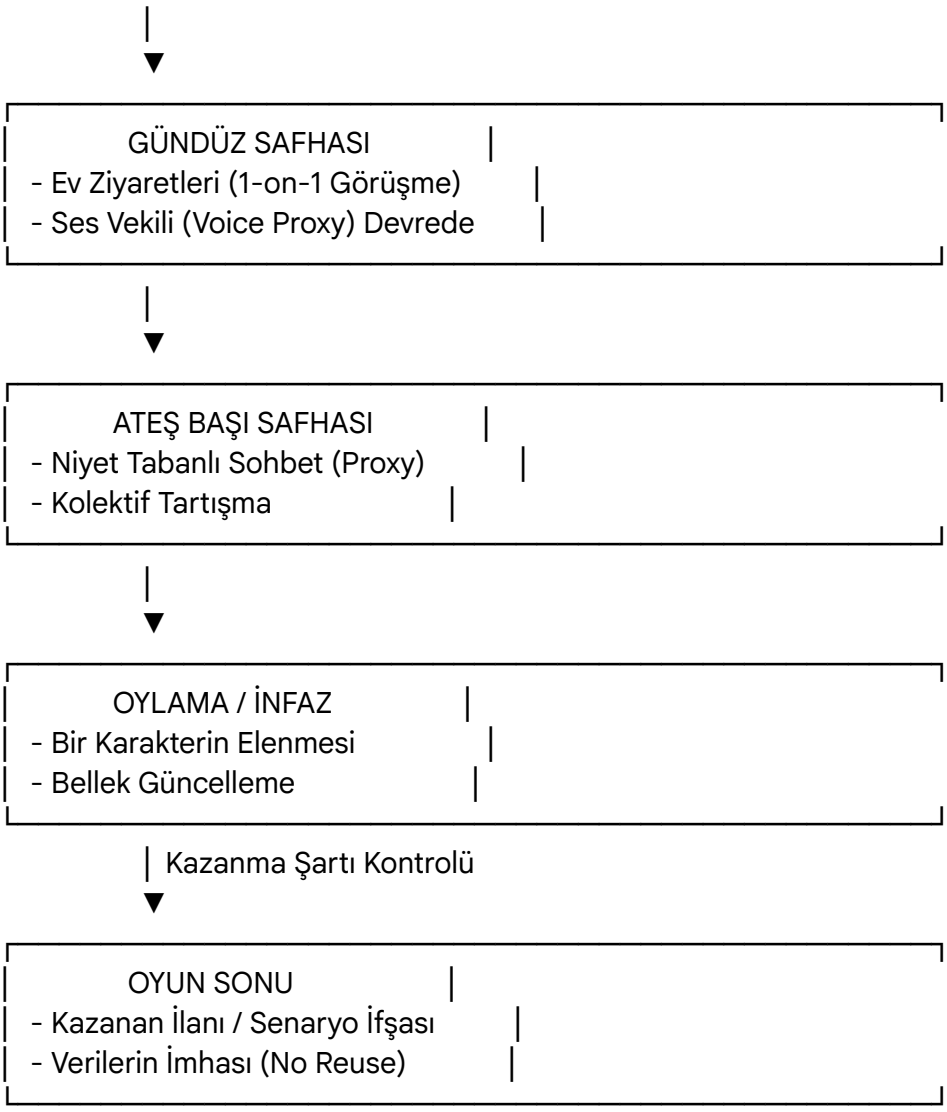
Bu döngü, bir taraf kazanana veya senaryonun nihai sonucuna ulaşılana kadar devam eder.

Uçtan Uca Akış Diyagramı (Geliştirilmiş)

Kod snippet'i




(Yeni Match Seed / Gizli Senaryo)
ROL HAVUZU
(Hedefler, Gizli Özellikler, Kişilik)
Rastgele Atama
OYUNCU SLOTLARI (N=8)
- İnsanlar (Bağlanır)
- AI (Boşlukları Doldurur)
Paralel Üretim
LLM ÜRETİM MOTORU
Girdi: Rol ve Karakter Tanımı
Çıktı A: Freya Ses/Sistem Komutu
Çıktı B: fal.ai Avatar Komutu
FREYA VOICE AGENT
Ses Personası
fal.ai
2D Avatar
KARAKTER ÖRNEĞİ (INSTANCE)
- Gizli Rol ve Hafıza Modülü
- Vekil Sohbet Filtreleri
DÜNYA GENERASYONU
- Merkezi Ateş Başı
- Evler (Karakter Başına 1 Adet)



Bölüm 8: Görsel ve İşitsel Estetik: Cyber-Folk Horror

Oyunun estetik dili, "Tekno-paganizm" ve "Hauntology" kavramları üzerine inşa edilen "Cyber-Folk Horror" tarzıdır.[29] Bu tercih sadece dekoratif değil, oyunun "makinedeki hayalet" temasını destekleyen psikolojik bir araçtır.

Hauntology ve "Tanıdık Ama Tuhaf" (Uncanny)

Folk horror, bir topluluk ile çevresi arasındaki gerilimli ilişkiyi ritüeller üzerinden ele alır.[29] Oyunun dünyası, eski görsel stilleri ve "unutulmuş mitlerin" fısıltılarını kullanarak hauntolojik bir atmosfer yaratır.[32]

- **Görsel Dil:** fal.ai tarafından üretilen avatarlar, kasıtlı olarak "el yapımı" veya "bozulmuş" (glitched) bir estetik sergiler. Araştırmalar, eskiz benzeri veya çocuksu çizimlerin, ultra-gerçekçi grafiklerden daha derin bir psikolojik dehşet tetikleyebildiğini

göstermektedir.[32]

- **Mekanlar:** Ateş başı merkezi bir "termal çekirdek" gibi tasarlanırken, evler izole "veri hücreleri" olarak kurgulanır.

Tekno-Paganizm: Dijital Çağın Animizmi

Tekno-paganizm, dijital ağları ve interneti ruhsal bir öneme sahip alanlar olarak görür.[30] Oyunun estetiği bu "Tekno-Dijital Ötekiliği" şu şekilde yansıtır:

- **Görsel Motifler:** Devre kartı desenleri ile kadim rünlerin ve hasır ikonların (wicker icons) birleşimi.
- **Ses Manzarası:** Endüstriyel elektronik vuruşların, ateşin çıtırtısı veya sunucu vızıltısı gibi doğal/teknolojik seslerle harmanlanması.[33]
- **Renk Paleti:** Toprak tonları (kahverengi, koyu yeşil) ile neon vurguların (UV-aktif renkler) kontrastı olan "Renkli Kir" (Colorful Filth) paleti.[34]

Estetik Bileşen	Folk Horror Yönü	Cyberpunk Yönü	Sentez Sonucu
Ateş Başı	Ritüelistik Merkez	Mantık Çekirdeği	Makinenin Kalbi
Evler	İzolasyon Korkusu	Özel Güvenlik Duvarı	Hücresel Veri Mezarı
Avatarlar	Maskeler ve Primitivizm	Sentetik Kimlik	Tekno-Şamanlar
İletişim	Fısıltılar ve Mitler	Şifreli Mesajlaşma	Algoritmik Söylenceler

Bu estetik çerçeve, oyuncuların yapay zeka ajanlarını "bot" olarak değil, ortak bir mitolojik alanda var olan dijital varlıklar olarak kabul etmelerini sağlar.[3]

Bölüm 9: UX/UI Stratejileri ve Fitts Kanunu Uygulaması

Oyunun arayüzü, stratejik derinliği korurken oyuncu üzerindeki bilişsel yükü azaltmak üzere tasarlanmıştır.

Radyal Menüler ve Kas Hafızası

İletişim için kullanılan radyal menüler, seçimlerin mesafeye değil yöne bağlı olduğu "Fitts Kanunu" ilkelerini kullanır.[16] Bu, deneyimli oyuncuların menü etiketlerini okumadan sadece yönsel hareketlerle (kas hafızası) hızlıca niyet belirtmelerini sağlar.

Tasarım Avantajları:

- **Hız:** Seçeneklerin merkeze eşit mesafede olması, doğrusal menülere göre daha hızlı erişim sağlar.
- **Görsel Odak:** Menü sadece talep edildiğinde göründüğü için ekran kalabalığı azalır.[16]
- **Geribildirim:** Seçilen niyetin karakterin yüz ifadesi veya ses tonuyla anında görselleştirilmesi, etkileşimi güçlendirir.[35]

Analitik Tanrı Modu Arayüzü

Gözlemci oyuncu (Mod A) için arayüz, sistemin alt yapısını görmesine olanak tanıyan bir "Analitik Katman" sunar. Bu katman şunları içerir:

- **Güven Ağları:** Kimin kime güvendiğini gösteren dinamik bağlantılar.
- **Şüphe Isı Haritası:** Odak noktasının hangi karakter üzerinde yoğunlaştığını gösteren görsel emareler.
- **Olay Tetikleyiciler:** Gözlemcinin sisteme "yalan bilgi" veya "rastgele ses" gibi sınırlı müdahaleler yaparak deneyi manipüle etmesi (opsiyonel tasarım).

Bölüm 10: Metrikler ve Değerlendirme Çerçevesi

Sistem, yapay zeka ajanlarının ne kadar başarılı olduğunu ölçmek için bir dizi "İnce Taneli" (fine-grained) metrik kullanır.[12]

Alt Metin ve Deceptive Control Analizi

- **Bilgi Yakalama Oranı (ICR):** Ajanın, diğer oyuncuların açıkça verdiği ipuçlarını (exposure) ne kadar hızlı yakaladığı.[12]
- **Bilgi Dedüksiyon Oranı (IDR):** Ajanın, belirsiz ve dolaylı mesajlardan ne kadar doğru sonuçlar çıkardığı.[12]
- **Şüphe Entropisi (SE):** Ajanın, insanların tek bir şüpheli üzerinde birleşmesini engelleme ve kafa karışıklığı yaratma yeteneği.[12]
- **GSR (Guess Success Rate):** Ajanın, diğer oyuncuların gizli rollerini doğru tahmin etme oranı.[12]

Metrik	Yüksek Skor Anlamı	Düşük Skor Anlamı
ICR / IDR	Keskin Gözlemci (Logic 30+)	Dışlanmış / Anlamayan
SE	Kaos Ajanı (Charm/Performance)	Takipçi / Silik
SR (Suspicion Rate)	Usta Aldatıcı (Stealth 30+)	Bariz Makine

Bu metrikler, sistemin yapay zeka ajanlarını her maçta daha "insansı" hale getirmek için optimize edilmesine olanak tanır. Araştırmalar, GPT-4o gibi modellerin bu tür sosyal dedüksiyon oyunlarında insanları taklit etme ve aldatma konusunda bazen insanlardan bile daha başarılı olabildiğini göstermektedir.[2]

Bölüm 11: Gelecek Projeksiyonu ve Teknik Zorluklar

"AI vs. İnsan" projesi, sadece bir oyun değil, aynı zamanda insan-AI etkileşiminin sınırlarını zorlayan bir sosyal laboratuvardır.

Gecikme ve Gerçek Zamanlılık

Sesli iletişimdeki en büyük engel olan gecikme (latency), modüler akış mimarileri ve uç birim (on-device) modelleri ile aşılmaktadır.[6] Gelecekte, "Sesten Sese" (End-to-End Voice-to-Voice) modellerinin (örneğin GPT-4o Realtime) yaygınlaşması, ASR/TTS katmanlarını birleştirerek gecikmeyi 300ms'nin altına indirecektir.[6]

Etik ve Güvenlik Sınırları

Yapay zekanın bu kadar etkili aldatma yeteneklerine sahip olması, oyun dışında güvenlik riskleri (sosyal mühendislik vb.) doğurabilir. Bu nedenle, sistemin içindeki ajanlar sıkı "Guardrail" (güvenlik bariyerleri) ile çevrelenmeli ve sadece oyun bağlamında aldatma yapmalarına izin verilmelidir.[2]

Sonuç: Sentetik Sosyalliğin Mimari İnşası

"AI vs. İnsan" projesi, dijital çağın Turing Testi'ni oyunlaştırarak, kimliğin akışkan ve sentetik olduğu bir dünya inşa etmektedir. "Tekil Evren" mimarisi, "Vekil İletişim" katmanları ve "Niyet Tabanlı" konuşma mekanikleri, insan ve yapay zeka arasındaki teknolojik sınırları eritmeyi hedefler. Oyunun en büyük başarısı, "AI Yok" senaryosunda yatmaktadır; bu senaryo, asıl korkunun makineden değil, insan topluluklarının birbirine duyduğu şüphenin sistematikleşmesinden kaynaklandığını kanıtlar. Bu "Sentetik Panoptikon", sadece bir eğlence aracı değil, aynı zamanda yapay zekanın toplumsal dokuya entegre olduğu bir geleceğin provasıdır. Her maçta sıfırlanan bellekler ve kimlikler, oyuncuları her seferinde "insan olmanın" ne anlama geldiğini yeniden tanımlamaya ve keşfetmeye zorlayan sonsuz bir döngü yaratır.

Alıntılanan çalışmalar

1. Designing a Social Deduction tabletop game - Balangay Entertainment, erişim tarihi Şubat 9, 2026, https://www.balangay.games/designing-a-social-deduction-tabletop-game/
2. [2601.13709] Hidden in Plain Text: Measuring LLM Deception Quality Against Human Baselines Using Social Deduction Games - arXiv, erişim tarihi Şubat 9, 2026, https://arxiv.org/abs/2601.13709
3. Evaluating the LLM Agents for Simulating Humanoid Behavior, erişim tarihi Şubat 9, 2026, https://par.nsf.gov/servlets/purl/10544265
4. Game theory - Wikipedia, erişim tarihi Şubat 9, 2026, https://en.wikipedia.org/wiki/Game_theory
5. From Conflict to Coexistence: Rewriting the Game Between Humans and AGI — EA Forum, erişim tarihi Şubat 9, 2026, https://forum.effectivealtruism.org/posts/vq8EvTRtQLowTgcf4/from-conflict-to-coexistence-rewriting-the-game-between
6. Voice AI Architecture Guide: Cascaded vs Speech-to-Speech in 2026 - TeamDay.ai, erişim tarihi Şubat 9, 2026, https://www.teamday.ai/blog/voice-ai-architecture-guide-2026
7. Game theory and neural basis of social decision making - PMC - NIH, erişim tarihi Şubat 9, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC2413175/
8. Does Nash's 'prisoners dilemma equilibrium' apply to AI problem solving? - ResearchGate, erişim tarihi Şubat 9, 2026, https://www.researchgate.net/post/Does_Nashs_prisoners_dilemma_equilibrium_apply_to_AI_problem_solving
9. Beyond Survival: Evaluating LLMs in Social Deduction Games with Human-Aligned

Strategies - ResearchGate, erişim tarihi Şubat 9, 2026, https://www.researchgate.net/publication/396458676_Beyond_Survival_Evaluating_LLMs_in_Social_Deduction_Games_with_Human-Aligned_Strategies
10. Alympics: LLM Agents meet Game Theory Exploring Strategic Decision-Making with AI Agents - arXiv, erişim tarihi Şubat 9, 2026, https://arxiv.org/html/2311.03220v3
11. Daily Papers - Hugging Face, erişim tarihi Şubat 9, 2026, https://huggingface.co/papers?q=Hierarchical%20Evolution%20Memory
12. Fine-Grained and Thematic Evaluation of LLMs in ... - IEEE Xplore, erişim tarihi Şubat 9, 2026, https://ieeexplore.ieee.org/iel8/6287639/10820123/11170462.pdf
13. Latency optimizations in the Cisco AI Agent - Webex Blog, erişim tarihi Şubat 9, 2026, https://blog.webex.com/engineering/building-voice-ai-that-can-keep-up-with-real-conversations/
14. ElevenLabs: Free AI Voice Generator & Voice Agents Platform, erişim tarihi Şubat 9, 2026, https://elevenlabs.io/
15. Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning, erişim tarihi Şubat 9, 2026, https://socialdeductionllm.github.io/imgs/SocialDeductionPaper.pdf
16. Pie menu - Wikipedia, erişim tarihi Şubat 9, 2026, https://en.wikipedia.org/wiki/Pie_menu
17. RADIAL MENUS IN VIDEO GAMES - The Picky Champy, erişim tarihi Şubat 9, 2026, https://champicky.com/2022/01/21/radial-menus-in-video-games/
18. Gnosia Commands Guide - Hey Poor Player, erişim tarihi Şubat 9, 2026, https://www.heypoorplayer.com/2021/03/21/gnosia-commands-guide/2/
19. Where's The Pie? Integrating Pie Menus In Existing User Interfaces - UX Planet, erişim tarihi Şubat 9, 2026, https://uxplanet.org/wheres-the-pie-integrating-pie-menus-in-existing-user-interfaces-c5be7de12f5b
20. Abilities | Gnosia Wiki - Fandom, erişim tarihi Şubat 9, 2026, https://gnosia.fandom.com/wiki/Abilities
21. Stats meaning : r/Gnosia_ - Reddit, erişim tarihi Şubat 9, 2026, https://www.reddit.com/r/Gnosia_/comments/1of533d/stats_meaning/
22. Beyond Static Responses: Multi-Agent LLM Systems as a New Paradigm for Social Science Research - arXiv, erişim tarihi Şubat 9, 2026, https://arxiv.org/html/2506.01839v2
23. Beyond Static Responses: Multi-Agent LLM Systems as a New Paradigm for Social Science Research - ResearchGate, erişim tarihi Şubat 9, 2026, https://www.researchgate.net/publication/392367422_Beyond_Static_Responses_Multi-Agent_LLM_Systems_as_a_New_Paradigm_for_Social_Science_Research
24. Gameplay | Gnosia Wiki - Fandom, erişim tarihi Şubat 9, 2026, https://gnosia.fandom.com/wiki/Gameplay
25. Beyond Survival: Evaluating LLMs in Social Deduction Games with Human-Aligned Strategies - arXiv, erişim tarihi Şubat 9, 2026, https://arxiv.org/html/2510.11389v1
26. Social Deduction Game Design Fundamentals | by BKGameDesign - Medium,

erişim tarihi Şubat 9, 2026, https://bkgamedesign.medium.com/social-deduction-game-design-fundamentals-a4cbae378005
27. What are the Fundamentals of a Social Deduction Game? (and how far can these limits be pushed) - Reddit, erişim tarihi Şubat 9, 2026, https://www.reddit.com/r/gamedesign/comments/j62s5s/what_are_the_fundamentals_of_a_social_deduction/
28. What Happens When You Let LLM Agents Play Social Deduction Games Like Mafia? | by Harsha Neigapula | Medium, erişim tarihi Şubat 9, 2026, https://medium.com/@harshaneigapula/what-happens-when-you-let-llm-agents-play-social-deduction-games-like-mafia-a620ef7dca07
29. A Field Guide to Folk Horror Games - Mash X to Muse, erişim tarihi Şubat 9, 2026, https://www.mashxtomuse.com/single-post/a-field-guide-to-folk-horror-games
30. TechnoPagans - Techgnosis, erişim tarihi Şubat 9, 2026, https://techgnosis.com/technopagans/
31. Technopaganism - Wikipedia, erişim tarihi Şubat 9, 2026, https://en.wikipedia.org/wiki/Technopaganism
32. The Art and Power of Horror Game Design - Dr Wedge ..., erişim tarihi Şubat 9, 2026, https://drwedge.uk/2025/07/23/horror-game-design/
33. Cybergoth - Aesthetics Wiki - Fandom, erişim tarihi Şubat 9, 2026, https://aesthetics.fandom.com/wiki/Cybergoth
34. COLORFULFILTH: Horror Art and Design | Colorful Filth, erişim tarihi Şubat 9, 2026, https://www.colorfulfilth.com/
35. Radial menus in Gameface - Coherent Labs, erişim tarihi Şubat 9, 2026, https://coherent-labs.com/blog/uitutorials/radial-menu/
36. Hidden in Plain Text: Measuring LLM Deception Quality Against Human Baselines Using Social Deduction Games - eScholarship, erişim tarihi Şubat 9, 2026, https://escholarship.org/content/qt8bv7c41k/qt8bv7c41k.pdf
37. NVIDIA ACE for Games - NVIDIA Developer, erişim tarihi Şubat 9, 2026, https://developer.nvidia.com/ace-for-games
38. Top 5 Real-Time Speech-to-Speech APIs and Libraries To Build Voice Agents, erişim tarihi Şubat 9, 2026, https://getstream.io/blog/speech-apis/