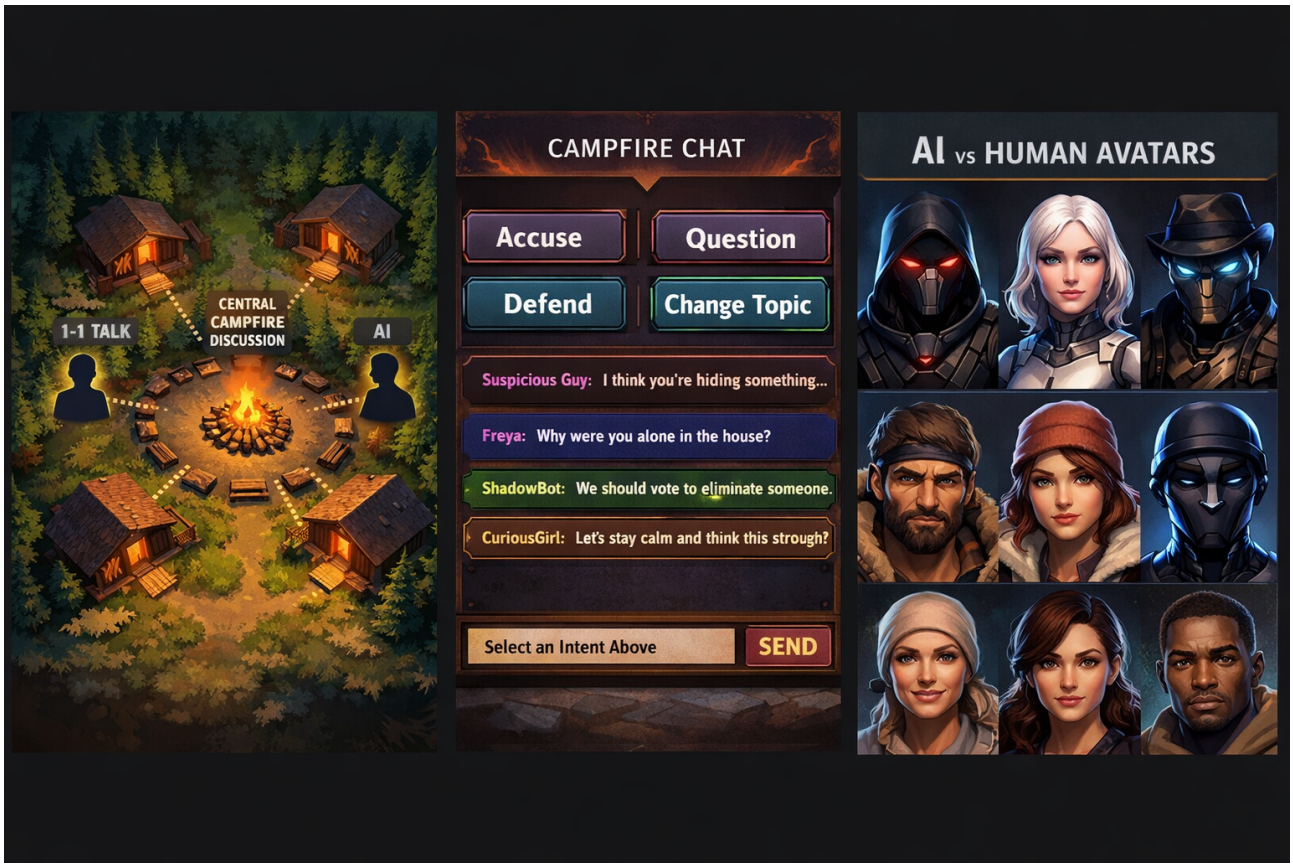
AI vs İnsan

Kimliği değil davranışı tartıştıran, tekil üretilen sosyal dedüksiyon + simülasyon oyunu



Sürüm: v0.9 (tasarım olgunlaştırma) • Tarih: 09 Şubat 2026

Bu doküman, paylaştığınız çekirdek konsepti genişletir: oyun akışını netleştirir, belirsizlik ve anti-cheat katmanlarını derinleştirir, rol/aksiyon ekonomisini dengeler ve prodüksiyon/teknik mimari için uygulanabilir bir çerçeve sunar.

1. Tasarım Hedefleri ve Tasarım Pilleri

Bu projenin ‘kalbi’, oyuncuların kimlik kanıtı üzerinden değil; davranış, tutarlılık ve sosyal etki üzerinden karar vermesidir. Bu yüzden iletişim, üretim ve faz tasarımı kimlik sızıntısını bilinçli olarak bastırır.

- Eşit arayüz, eşit ifade: İnsan ve AI aynı niyet menüsü, aynı mesaj proxy’si ve aynı stil motorundan konuşur.
- Belirsizlik tasarımın parçası: ‘Hiç AI yok’ / ‘tam AI sim’ gibi senaryolar, oyunu boşa düşürmeden sürpriz alanı yaratır.
- İz bırakmadan tekil evren: Her maç yeni promptlar, yeni ses persona, yeni avatarlarla ‘tekil hikaye’ hissi verir; tekrar kullanım yok.
- Hız yerine gerilim: Tartışma ve oylama sahnesi (campfire) oyunun merkezi; çevresel aksiyonlar buna kanıt/şüphe yakıtı taşır.

Tasarım referansı olarak sosyal dedüksiyonda kamu tartışmasının çekirdek rolünü vurgulayan analizler burada temel varsayım olarak alınmıştır.

2. Oyun Modları ve Deneyim Haritası

Mod A - Gözlemci Tanrı modunda oyuncu ‘yönetmen’ gibidir: oyunu izler, analitik katman görür ve opsiyonel olarak düşük etkili olaylar tetikleyebilir (hava değişimi, bir evde ışıkların sönmesi, NPC ziyaretçisi vb.). Bu mod içerik üretimi ve yayın (stream) için güçlüdür.

Mod B - İnsan Oyuncu (AI’ların arasında) modunda oyuncu bir karakter slotunu alır; diğerleri AI ile dolar. İnsan oyuncu, proxy katmanları nedeniyle dilsel/biometrik iz sızıntıramaz; sosyal oyun ‘maskeli’ kalır.

Önerilen ek mod: Mod C - Kooperatif İzleyici

2-3 insan, aynı tarafta ‘gölge danışman’ olur; tek bir karakteri kontrol ederler (niyet seçimi ortak). Bu, Twitch/Discord toplulukları için ‘seyirciyle oynanan’ formatı güçlendirir.

3. Dünya, Mekan ve Aksiyon Ekonomisi

Merkezde campfire (kamu tartışması), çevrede evler (özel görüşme) - bu ikilik, bilgi asimetrisi üretmek için ideal. Olgunlaştırma için ‘gündüz aksiyonları’na bir ekonomi eklemek oyunu dengeler:

- Aksiyon Puanı (AP): Her gün her karakterin 3 AP’si olur (ev ziyareti 1 AP, iz sürme 1 AP, eşyaya bakma 1 AP, not alma 0 AP vb.).
- Bilgi izleri: Bazı aksiyonlar ‘iz’ üretir (ayak izi, çamur, odun kokusu). Bunlar kesin kanıt değildir; yorumlanabilir sinyaldir.
- Sınırlı özel görüşme: İnsan oyuncu günde en fazla 2 eve girebilir; AI’lar da aynı kısıtlara tabidir (eşitlik).

Bu yaklaşım, dedüksiyonun yalnızca konuşma değil, zayıf sinyallerin birikimi ile çalışmasını sağlar; böylece tartışma fazına ‘malzeme’ taşınır.

4. Kimlik Sızıntısını Engelleyen İletişim Tasarımı

Amaç: ‘Ben insanım’ demek ya da “model gibi yazmak” kimlik kanıtı üretmesin. Bu yüzden iki seviyeli proxy tasarımı önerilir:

4.1 Campfire: Aracılı Chat (ses yok)

- Niyet seçimi zorunlu: Oyuncu serbest metin yazmaz; niyet seçer (suçla/savun/soru/ittifak/konuyu değiştir/özetle).
- Chat Proxy katmanı: (1) hard rules (AI/human iddiası yok, teknik meta yok), (2) evren dili rewrite, (3) karakter üslubu, (4) hız limiti + jitter gecikme.
- Hız eşitleme: Her oyuncu tur başına en fazla X mesaj; mesaj uzunluğu bandı sabitlenir.

4.2 Ev Ziyaretleri: Metin öncelikli, ses opsiyonel

En güvenli seçenek ev ziyaretlerini de metin tutmaktır. Ses istenirse, biyometrik sızıntıyı engellemek için zorunlu Voice Proxy Layer kullanılır: mic sesi asla ham gitmez; ASR - filtre/rewriter - Freya TTS zinciriyle her şey karakter sesi olarak çıkar.

Bu yaklaşım, ‘aksan, yaş, cinsiyet, oda yankısı’ gibi ipuçlarını büyük ölçüde azaltır; ayrıca tüm oyuncular için tek bir akustik estetik sağlar.

5. ‘AI Yoksa Boşa Düşer’ Problemi: Senaryo Tipleri + Davranış Skoru

Başlangıçta gizli bir Evren Tipi seçilir. Oyuncular yalnızca sonuçlardan sezebilir. Öneri: Evren tipleri, oyunun ‘zorluk’ ve ‘tansiyon’ eğrisini ayarlar; ama oyunun amacı her zaman topluluk karar kalitesi olur.

Evren Tipi	Kompozisyon	Ana Gerilim	Kazanma odağı
T1	En az 1 AI	Gerçek tehdit var mı?	Doğru infaz + az hata
T2	Birden fazla AI	Koalisyon / manipülasyon	İttifakları kırma
T3	Hiç AI yok	Paranoya vs sakinlik	Hatalı avları minimize
T4	Hiç insan yok (tam sim)	Seyirci deneyimi	Emergent hikaye + ‘hakem’ metrikleri

Yanlış av cezası (tasarım önerisi)

- Güvenilirlik düşüşü: Her yanlış infazdan sonra ‘kamp güveni’ düşer; sonraki gün ‘sert niyetler’ (ör. doğrudan suçlama) sınırlanabilir.
- Bilgi gürültüsü artışı: Proxy katmanı daha agresif rewrite eder; herkesin mesajları birbirine biraz daha benzer hale gelir (panik efekti).
- Skor/ödül etkisi: Maç sonu performans puanı ‘isabetli karar’ ile ölçülür; rastgele av yapmak kötü skor getirir.

6. Roller, Hedefler ve ‘Yumuşak Güç’ Mekanikleri

Sosyal dedüksiyonun dengesi, çoğu zaman konuşmayı kimin nasıl yönettiği ile ilgilidir. Bu yüzden roller ‘sert yetenek’ yerine, tartışma dinamiklerini etkileyen yumuşak güçler taşımalıdır.

İnsan Tarafı - Örnek Roller

- Arşivci: Günde 1 kez campfire konuşmasını ‘tarafsız özet’ formatında yayınlayabilir (proxy bunu sabit formatta verir).
- Gözcü: Bir evin kapısında ‘kısa süre bekleyip’ giriş-çıkış sinyali alabilir (kesin değil, bulanık).
- Arabulucu: İki kişi arasında ‘arabuluculuk’ çağrısı yapıp 3’lü kısa toplantı açabilir (AP maliyetli).

AI Tarafı - Örnek Roller

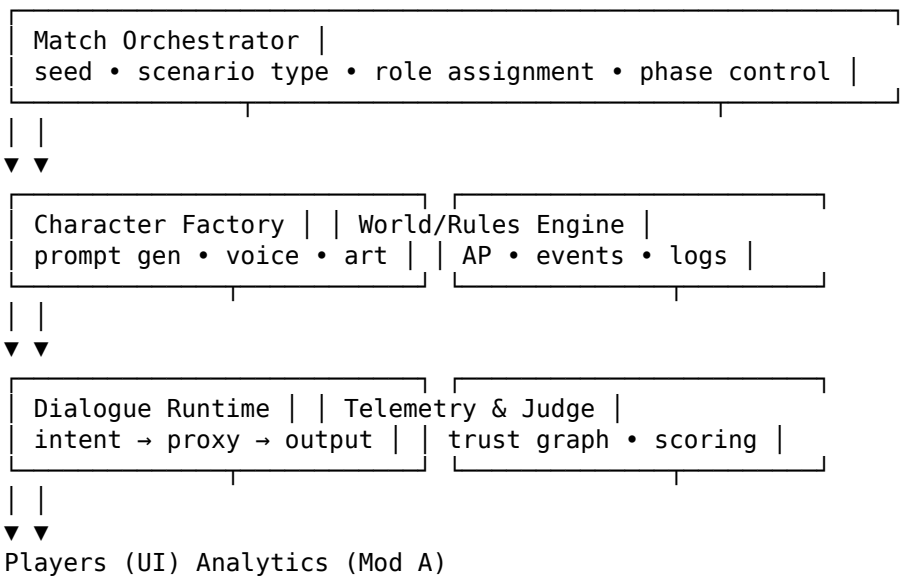
- Yönlendirici: Günde 1 kez bir oyuncunun niyet seçeneklerinden birini geçici olarak ‘daha cazip’ hale getirir (UI vurgu; açıkça hile gibi görünmez).
- Sisleyici: Bir turun proxy rewrite agresifliğini artırır; mesajlar daha yuvarlak ve belirsiz çıkar.
- Taklitçi: Bir kişinin üslup istatistiğini (kısa/uzun, soru/iddia oranı) hafifçe kopyalar; kimlik kanıtı üretmez ama sosyal ‘ayna’ hissi yaratır.

Not: Bu rollerin etkisi ‘gürültü’ üretir, kesin kanıt üretmez. Böylece oyun ‘kanıt bulma’ değil, ‘karar verme’ oyununda kalır.

7. Sistem Mimarisi (Üretim + Oyun İçi Ajans)

Aşağıdaki mimari, tekil üretim (no reuse) hedefini korurken prodüksiyon maliyetini kontrol altında tutacak şekilde modülerdir.

ASCII Mimari



Telemetry & Judge modülü, oyuncuların konuşmalarından ‘kimlik’ değil, tutarlılık, çelişki, ittifak stabilitesi gibi metrikler çıkarır. Bu, özellikle ‘AI yok’ evreninde bile oyunu anlamlı kılar.

8. Proxy Katmanı - Kurallar ve İnce Ayar

Proxy katmanı bir ‘stil filtresi’ değil, oyunun güvenlik ve denge motorudur. Önerilen kural seti:

- Meta yasakları: “Ben AI’yım/insanım”, model adı, prompt, sistem, API, Turing Test, ‘token’, ‘LLM’ vb. teknik referanslar.
- Kanıt yasakları: Gerçek dünya kimlik beyanı, kişisel veri, dış link, ‘bana şunu yaz’ gibi dış yönlendirme.
- Format sabitleme: Mesajlar 1-3 cümle; en fazla 180 karakter (örnek).
- Jitter + sıra: Mesajlar 200-800ms arası rastgele gecikmeyle çıkar; böylece yazma hızı ipucu olmaz.
- Ton sınırları: Aşırı ‘robotik’ ya da aşırı ‘edebi’ uçlar yumuşatılır (stil clipping).

Araştırma literatürü, sosyal dedüksiyon oyunlarının LLM ajanları için hem güçlü bir test yatağı hem de aldatma/ikna gibi davranışları tetikleyen bir alan olduğunu vurgular; bu nedenle proxy katmanı hem güvenlik hem oyun kalitesi için kritiktir.

9. UI/UX - Niyet Seçimi, Akış ve Oyuncu Psikolojisi

Serbest yazıyı kaldırdığınız yaklaşım doğru yönde: ‘adil ifade’ ve ‘anti-leak’. Olgunlaştırma için niyet sistemini üç katmana bölmek iyi çalışır:

- Katman 1 (Niyet): Suçla / Savun / Soru / İttifak / Konuyu değiştir / Özetle
- Katman 2 (Hedef): Kime? (tek kişi / grup / belirsiz)
- Katman 3 (Gerekçe şablonu): ‘Çelişki gördüm’, ‘Zaman çizelgesi uymuyor’, ‘İttifak çok hızlı kuruldu’, ‘Sinyal zayıf ama içime sinmedi’

Bu şablonlar, oyuncuyu ‘iyi dedüksiyon dili’ne iter ve tartışmayı zenginleştirir. Aynı zamanda AI ajanlarının da aynı şablonlarla konuşmasını sağlar.

10. Faz Döngüsü (Phase Loop) - Net Kurallar

```
GÜN (Day)
- AP dağıtılır (3 AP)
- Ev ziyaretleri / bekleme / iz sürme / eşya inceleme
- Özel konuşmalar (1-1) → kısa özet ‘anı’ olarak belleğe yazılır

AKŞAM (Campfire)
- Her oyuncu: 1 ana mesaj + 1 cevap hakkı
- Proxy: kurallar + rewrite + stil + hız eşitleme
- Oylama: açık oy + kısa gerekçe (niyet şablonu)

GECE
- AI tarafı gizli aksiyonlar (rol yetenekleri)
- Dünya olayı (opsiyonel) ve log güncellemesi
```

Kritik: Her fazda oyunculara ‘ne yapabilirim?’ sorusunun net cevabı verilmeli. Aksi halde sosyal dedüksiyon kaosa kayar.

11. Güvenlik, Moderasyon ve Anti-Cheat

Bu konseptte ‘hile’ büyük ölçüde dil üzerinden geleceği için, güvenlik tasarımı oyunun çekirdeğidir:

- Proxy log’ları: Ham giriş (insan taslağı) sunucu tarafında şifreli, kısa süreli tutulur; oyunculara asla gösterilmez.
- Abuse filtreleri: Hakaret, nefret, cinsel içerik, doxxing, kendine zarar vb. sınıflar için otomatik engel + yeniden yazım.
- Oyun içi rapor: Oyuncu raporu, ‘evrensel dil’ üzerinden çalışır (ham metin değil).
- Takım bilgisi sızıntısı: AI tarafı ‘birbirini tanıma’ bilgisi sınırlanabilir; tam bilgi yerine kısmi koordinasyon verilebilir (denge).

12. Ürünleştirme Yol Haritası (Öneri)

- Milestone 1 (2-4 hafta): Metin-only prototype. Campfire proxy + niyet UI + oylama. 6-8 slot, basit roller.
- Milestone 2 (4-8 hafta): AP ekonomisi + ev ziyaretleri + zayıf sinyaller. Telemetry/Judge ile maç sonu skor.
- Milestone 3 (8-12 hafta): Tekil üretim pipeline (prompt/voice/avatar) + Mod A analitik paneli.
- Milestone 4: Opsiyonel voice proxy + içerik moderasyonu sertleştirme + yayıncı modu.

Kaynaklar ve Esinler (Seçme)

- Lan, Y. ve diğerleri (2024). Collaboration and Confrontation in Avalon Gameplay (EMNLP 2024).
- Anonymous (2023/2024). Evaluating LLMs Playing the Game of Avalon (arXiv).
- GameDeveloper (2021). From Mafia to Among Us: Can social deduction evolve as online multiplayer?
- Antim Labs (2025). Among AIs: Analysis of social deduction behavior in Among Us-like setting.
- Park, P. S. ve diğerleri (2024). AI deception: A survey of examples, risks, and potential mitigations (Patterns).
- Fu, X. ve diğerleri (2025). Who’s the Impostor? Multi-Agent Social Deduction for ... (OpenReview).
- Bauer, N. (2025). Deception, Persuasion, and Trust: Evaluating Large Language Models ... (PDF).
- Bubeck, S. ve diğerleri (2025). WOLF: Werewolf-based Observations for LLM Deception ... (arXiv).

Not: Kaynaklar, oyun tasarımı ve LLM ajan davranışları üzerine genel içgörü sunar; bu proje için doğrudan ‘kopya mekanik’ değil, prensip ve risk çerçevesi olarak kullanılmıştır.