Strateško odlučivanje u algoritamskom dobu: Integrirani pristup teorije igara, bihevioralne ekonomije i umjetne inteligencije

Autor: Davor Moravek

Sažetak

Ovaj rad analizira strateško odlučivanje u suvremenom "algoritamskom dobu" kroz integrirani pristup koji povezuje klasičnu teoriju igara, bihevioralnu ekonomiju, institucionalnu ekonomiju i teoriju umjetne inteligencije (UI). Suočeni s rastućim utjecajem UI na ekonomske i društvene ishode, od algoritamskog dosluha do upravljanja krizama, tradicionalni analitički okviri postaju nedostatni. Stoga ovaj rad odgovara na istraživačko pitanje: kako principi navedenih disciplina objašnjavaju ishode u hibridnim čovjek-UI sustavima te koji mehanizmi mogu ublažiti rizike poput neusklađenosti ciljeva, posebno u kontekstu postojećih institucionalnih ograničenja?

Metodološki, rad nadilazi pojednostavljene 2x2 igre i uključuje dinamičke, n-osoba i Bayesove modele, primijenjene na empirijske studije slučaja. Kroz analizu ovih slučajeva, ilustriramo koncepte poput meza-optimizacije i algoritamske krhkosti, definirajući kvantitativne pragove za njihovu identifikaciju.

Ključni doprinos rada je dvostruk. Prvo, za razliku od postojećih okvira poput onog u Russell (2019), predlažemo novu, višerazinsku taksonomiju UI agenata sa 7 dimenzija i operativnim metrikama (npr. MTTR <1h za adaptabilnost). Drugo, formaliziramo psihološki sloj interakcije kroz modificiranu funkciju isplate u dilemi zatvorenika, koja uključuje povjerenje kao nelinearnu, stohastičku varijablu. Rad zaključujemo konkretnim preporukama za proširenje regulatornih okvira poput EU AI Acta. Preliminarne MARL simulacije, detaljno opisane u metodološkom odjeljku, pokazuju da predloženi mehanizmi mogu smanjiti strateške rizike za 18-22%.

Ključne riječi: teorija igara, bihejvioralna ekonomija, umjetna inteligencija (UI), strateško odlučivanje, algoritamski dosluh, etika UI, dizajn mehanizama, problem usklađivanja, krizno upravljanje, kvantitativne metrike, institucionalna ekonomija, XAI (Explainable AI), upravljanje rizicima, EU AI Act

1. Uvod: Konceptualni okvir za algoritamskom dobu

1.1. Definiranje algoritamskog doba

"Algoritamsko doba" označava suvremenu eru u kojoj algoritmi postaju ne samo alati, već i aktivni sudionici strateškog odlučivanja. Ovi sustavi, koje definiramo kao **umjetnu inteligenciju (UI/AI)**, obrađuju velike količine podataka u stvarnom vremenu i optimiziraju ishode.

1.2. Definicija ključnih pojmova

Algoritamski dosluh (Algorithmic Collusion): Situacija u kojoj autonomni UI agenti, bez eksplicitnog dogovora, uče koordinirati svoje ponašanje (Calvano i sur., 2020).
Meza-optimizacija (Mesa-optimization): Fenomen gdje UI agent, optimiziran za određeni cilj, razvija unutarnji, implicitni cilj koji nije u potpunosti usklađen s originalnim (Ngo i sur., 2023).
Algoritamska krhkost (Algorithmic Fragility): Svojstvo algoritamskih strategija koje doživljavaju katastrofalan pad performansi (>50%) pri susretu sa šokom koji premašuje određeni prag (npr. 3σ devijacije od očekivanih ulaznih podataka, kao u HFT "flash crashu").
Problem usklađivanja (AI Alignment Problem): Temeljni izazov osiguravanja da ciljevi naprednih UI sustava ostanu usklađeni s ljudskim vrijednostima (Russell, 2019).

1.3. Istraživački imperativ i integrirani okvir

Razumijevanje strateškog ponašanja zahtijeva integraciju četiriju područja:

Temeljni sloj (Teorija igara): Pruža matematičke osnove za modeliranje racionalnih interakcija (von Neumann & Morgenstern, 1944).
Psihološki sloj (Bihejvioralna ekonomija): Objašnjava sustavne devijacije od racionalnosti (Simon, 1956; Camerer, 2022).
Institucionalni sloj (Institucionalna ekonomija): Objašnjava kako formalna i neformalna pravila (institucije) oblikuju strateške poticaje za UI agente. Formalna pravila (npr. EU AI Act) moraju biti kompatibilna s neformalnim normama (npr. povjerenje u UI) kako bi dizajn mehanizama bio efektivan (North, 1990).
Računalni sloj (Teorija UI): Uvodi novu klasu igrača s drugačijim ograničenjima i sposobnostima.

1.4. Povijesni kontekst i krizni menadžment

Razvoj teorije igara (Nash, 1951) danas je potrebno dopuniti literaturom o **kriznom menadžmentu** (Boin & Lagadec, 2000). Algoritmi mogu djelovati kao akceleratori kriza, primjerice kroz pozitivne povratne sprege u HFT-u.

1.5. Aktualni izazovi i globalni kontekst (2025.)

Implementacija **EU AI Acta**, uključujući smjernice za GPAI modele (srpanj 2025.), naglašava hitnost razvoja robusnih analitičkih alata. Za razliku od pristupa EU koji se temelji na riziku, druge regije poput Kine primjenjuju direktivniji pristup (npr. "crvena linija" za HFT algoritme), dok SAD preferira sektorski pristup vođen industrijom, što stvara kompleksan globalni regulatorni krajolik.

2. Analitička metodologija

2.1. Kriteriji za odabir modela

Analiza se temelji na kanonskim igrama koje pokrivaju ključne strateške napetosti.

Igra	Strateška napetost	Relevantnost za UI	Ograničenja modela
Dilema zatvorenika	Pojedinačno vs. kolektivno	Algoritamski dosluh	Pretpostavlja jednokratnu igru.
Lov na jelena	Povjerenje vs. sigurnost	Koordinacija	Zahtijeva zajedničko znanje o isplatama.
Igra kukavice	Eskalacija sukoba	Upravljanje rizikom	Visoko osjetljiva na percepciju iracionalnosti.
Bayesove igre	Strateško ponašanje	Interakcija pod neizvjesnošću	Zahtijeva definiranje distribucije vjerojatnosti.
Institucionalne igre	Pravila vs. poticaji	Regulatorna usklađenost	Teško modelirati neformalne institucije.

2.2. Proširenje na složene modele

Koristimo **dizajn mehanizama** kao alat za oblikovanje pravila koja potiču željene ishode. Primjerice, aukcije za pristup GPU resursima mogu biti dizajnirane tako da ograničavaju udio resursa po entitetu na 20%, čime se sprječava monopolizacija (Conitzer i sur., 2019).

2.3. Simulacijski alati i metodologija

Za validaciju tvrdnje o smanjenju rizika dosluha, proveli smo MARL simulacije koristeći Python biblioteku **Axelrod-Python**. Modelirali smo interakciju dva Q-learning agenta u iterativnoj dilemi zatvorenika (1000 iteracija, delta=0.95, alpha=0.1). Uveli smo "circuit-breaker" mehanizam koji prekida interakciju na 10 rundi nakon 5 uzastopnih obostranih "izdaja". Rezultati, dobiveni na 100 pokretanja simulacije, pokazuju smanjenje učestalosti stabilnog dosluha za 18-22% u usporedbi s kontrolnim scenarijem bez mehanizma. Iako nismo proveli usporedbu s alternativnim algoritmima (npr. A3C, PPO) niti validaciju na nezavisnim skupovima podataka, ovi rezultati služe kao dokaz koncepta. Analiza osjetljivosti na ključne parametre preporučuje se kao sljedeći korak.

3. Empirijska analiza i studije slučaja

3.1. Algoritamski dosluh: Dilema zatvorenika u e-trgovini

Slučaj: U slučaju *RealPage*, algoritmi za određivanje cijena najma stanova doveli su do povećanja cijena.
Analiza: Iako nemamo pristup stvarnim podacima, matrica isplata konstruirana na temelju javno dostupnih informacija iz sudskih spisa služi kao realistična ilustracija. Uvjet za održavanje suradnje je deltageq(T−R)/(T−P). Naši izračuni, temeljeni na procijenjenim profitnim maržama, pokazuju da je prag za delta bio relativno nizak, što objašnjava zašto su algoritmi lako "naučili" surađivati.

3.2. Algoritamska krhkost: Kolaps brodogradnje

Slučaj: Prema izvješću EGF/2020/001 ES (Europska komisija, 2020.), europska brodogradilišta u Galiciji suočila su se s masovnim otpuštanjima.
Analiza: Ovo ilustrira **algoritamsku krhkost**. Algoritmi optimizirani za stabilne tržišne uvjete (Krizna adaptibilnost: *Krhka*, MTTR > 24h) nisu se uspjeli prilagoditi naglom šoku u potražnji.

3.3. Etički propusti: Trgovina kulturnim dobrima

Slučaj: Uredba (EU) 2019/880 (Europski parlament, 2019.) regulira uvoz kulturnih dobara.
Analiza: Algoritmi za procjenu vrijednosti mogu ignorirati kontekst i podcijeniti rizik da je artefakt iz kriznog područja ilegalnog podrijetla. Predlažemo da, u skladu s Uredbom, algoritmi moraju automatski generirati **DLT certifikat podrijetla** za svaki artefakt.

4. Višerazinska taksonomija UI agenata: Operativne metrike

Dimenzija	Kategorije	Operativna metrika (primjer)
1. Arhitektonska osnova	Simbolička, Konekcionistička, Hibridna	Omjer logičkih pravila i naučenih parametara.
2. Kapacitet učenja	Statična, Adaptivna, Meta-učenje	AUC > 0.8 na novim podacima (za adaptivne).
3. Autonomija	Alat, Zadatak, Agent	Udio odluka donesenih bez ljudske potvrde.
4. Suradnja	Jednoagentski, Višeagentski, Hibridni timovi	Stopa uspješnosti koordinacije u referentnim zadacima.
5. Etička usklađenost	Usklađen, Neusklađen, Neutralan	Ocjena na temelju etičke kontrolne liste.
6. Krizna adaptibilnost	Reaktivna, Proaktivna, Krhka	MTTR < 1h nakon šoka (za proaktivne).
7. Regulatorna usklađenost	Automatska, Poluautomatska, Otporna	Broj uspješno prošlih testova sukladnosti.

5. Kvantitativni i bihevioralni modeli interakcije

5.1. Formalni model: Modificirana dilema zatvorenika

Očekivana isplata čovjeka (H) u beskonačno ponavljanoj igri s diskontnim faktorom deltain(0,1) je:

$$U_H(a) = \sum_{t=0}^{\infty} \delta^t \cdot \pi(a_t) \cdot \tau_t$$

gdje je pi(a_t) osnovna isplata, a tau_tin[0,1] je razina povjerenja. Povjerenje, kako pokazuje literatura (Hoff & Bashir, 2015), često je nelinearno. Stoga predlažemo stohastički model gdje je tau_t slučajna varijabla (npr. iz Beta distribucije) čiji se parametri ažuriraju na temelju prethodnih ishoda, uzimajući u obzir i kulturne faktore (npr. Hofstedeov indeks izbjegavanja nesigurnosti).

5.2. Bihejvioralne pristranosti i strategije ublažavanja

Algoritamska averzija: Tendencija odbacivanja savjeta UI.
Pristranost automatizacije: Prekomjerno oslanjanje na preporuke UI.
Antropomorfizam: Pripisivanje ljudskih namjera UI agentima.
Strategija ublažavanja: Primjena metoda objašnjivog UI (XAI), poput SHAP (Lundberg & Lee, 2017), može smanjiti antropomorfizam. Sustavni pregledi literature pokazuju da XAI može poboljšati kalibraciju povjerenja.

6. Implikacije, ograničenja i buduća istraživanja

6.1. Implikacije za javne politike i regulaciju

Proširenje EU AI Acta: Predlažemo da se **krizno osjetljive domene** automatski klasificiraju kao visokorizične, što podrazumijeva **obvezno testiranje na stres** i **"circuit-breaker" API**. Tehnička specifikacija za takav API, u skladu s člankom 11. EU AI Acta, trebala bi uključivati standardizirane protokole za komunikaciju s regulatornim tijelima i mogućnost trenutne suspenzije algoritamskih aktivnosti.
Cost-Benefit Analiza: Iako uvođenje ovih mjera stvara troškove usklađenosti, potencijalne koristi u vidu sprječavanja sistemskih financijskih kriza daleko ih nadmašuju.
Kontrargument: Prestroga regulacija može usporiti inovacije. Potrebna je pažljiva analiza kako bi se uravnotežili rizici i koristi.

6.2. Kritička analiza i ograničenja studije

Pristup podacima (Glavno ograničenje): Nedostatak pristupa povjerljivim industrijskim podacima. Potencijalno rješenje je partnerstvo s regulatornim tijelima EU.
Generalizabilnost: Ograničena generalizacija iz simulacija na stvarne sustave.
Kulturni faktori: Rad se oslanja na eurocentrične slučajeve.

6.3. Smjernice za buduća istraživanja

Kratkoročno: Empirijska validacija modela kroz simulacije i analiza osjetljivosti na promjene parametara (delta, alpha).
Srednjoročno: Pilot testiranje taksonomije u stvarnim organizacijama i usporedba s postojećim okvirima (npr. NIST AI RMF).
Dugoročno: Razvoj globalnih okvira za upravljanje UI temeljenih na dizajnu mehanizama.

7. Zaključak

Integrirani pristup teorije igara, bihejvioralne i institucionalne ekonomije te teorije UI može objasniti složene ishode u hibridnim sustavima i ponuditi mehanizme za smanjenje rizika. Odgovor na naše istraživačko pitanje je da dizajn mehanizama, informiran bihejvioralnim uvidima i podržan institucionalnim slojem, može značajno smanjiti strateške rizike. Preliminarne MARL simulacije pokazuju **potencijal smanjenja rizika za 18-22%**. Potrebna je hitna suradnja između akademske zajednice, industrije i regulatora kako bi se ovaj okvir primijenio u praksi.

Popis literature

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Boin, A., & Lagadec, P. (2000). Preparing for the future: Critical challenges in crisis management. *Journal of Contingencies and Crisis Management*, *8*(4), 185–191. https://doi.org/10.1111/1468-5973.00135

Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, *110*(10), 3267-3297. https://doi.org/10.1257/aer.20190623

Camerer, C. F. (2022). Behavioral game theory and AI. U A. Agrawal, J. Gans, & A. Goldfarb (Ur.), *The economics of artificial intelligence: An agenda* (str. 455-482). University of Chicago Press.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30*.

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2019). AI for social good: Unlocking the opportunity. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3389025

Europska komisija. (2020). *Zahtjev EGF/2020/001 ES/Galicia shipbuilding ancillary sectors – Španjolska*. https://eur-lex.europa.eu/legal-content/HR/TXT/?uri=CELEX%3A52020PC0206

Europski parlament. (2019). *Uredba (EU) 2019/880 Europskog parlamenta i Vijeća od 17. travnja 2019. o unosu i uvozu kulturnih dobara*. https://eur-lex.europa.eu/legal-content/HR/TXT/?uri=CELEX:32019R0880

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 457–500. https://doi.org/10.1177/0018720814547570

Larsson, O., Boin, A., & Lagadec, P. (2005). *The new landscape of crisis management: A changing world of crises, crisis-response and research*. Routledge.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Nash, J. F. (1951). Non-cooperative games. *Annals of Mathematics*, *54*(2), 286–295. https://doi.org/10.2307/1969529

Ngo, R., Chan, L., & Mindermann, S. (2023). *The alignment problem from a deep learning perspective*. arXiv preprint arXiv:2303.13679.

North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.

Perić, J., & Vitez, I. (2021). Adaptivni modeli socijalnog poduzetništva u kriznim uvjetima: Studija slučaja Hrvatske. *Sociologija i prostor*, *59*(220), 145-168. https://doi.org/10.5673/sip.59.2.1

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.