

PROIECȚIA UNUI MODEL COGNITIV KANTIAN ASUPRA IA. PROBLEME EPISTEMOLOGICE ȘI ETICE

MARIUS AUGUSTIN DRĂGHICI

Institutul de Filosofie și Psihologie
„Constantin Rădulescu-Motru” al Academiei Române

Abstract: This paper discusses recent research by Richard Evans and collaborators at *DeepMind* in light of Kant’s transcendental philosophy and its contemporary reception in cognitive science and artificial intelligence. The study is structured in two main parts. The first part presents and critically examines two major contributions that are fundamental to the present investigation. The first is Tobias Schlicht’s systematic and philosophically nuanced review of the reception of Kant’s transcendental philosophy within contemporary cognitive science, highlighting the diverse and often incompatible ways in which Kantian concepts have been appropriated by functionalist, enactivist, and predictive processing frameworks. The second is Evans et al.’s extensive research project, which develops a functional computational architecture inspired by a Kantian model of cognition, construed as a form of *a priori* cognitive psychology based on the first *Critique*. This project proposes an explicit “engine of apperception” and demonstrates that core Kantian conditions of the possibility of experience can be specified and implemented in an artificial cognitive agent.

The second part of the paper offers an epistemological evaluation of this approach, focusing on issues of rational justification, explicability, and epistemic control in artificial intelligence. It is argued that Evans’s Kantian-inspired architecture constitutes a non-black-box model of artificial cognition, one that prioritizes inferential transparency and rule-governed synthesis over purely opaque, statistically driven performance. This feature is shown to be of central importance for current debates surrounding Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI), particularly with respect to the ethical and epistemological risks associated with decision-making systems that cannot provide rationally intelligible explanations. By integrating recent developments in explainable AI—most notably Concept Relevance Propagation (CRP)—the paper suggests that a Kantian model of rationality offers not only a viable alternative to black-box architectures, but also a principled framework for addressing the epistemic and ethical challenges posed by advanced AI systems.

Keywords: Kantian cognition; Explainable Artificial Intelligence (XAI); Concept Relevance Propagation (CRP); Artificial General Intelligence (AGI); Black-box models.

Între numeroasele receptări contemporane ale *Criticii rațiunii pure* se numără și aceea din perspectiva științelor cognitivei (*cognitive sciences*). Înainte de a

circumscrie discuția din perspectiva cercetării de față, ne punem întrebarea, firesc, în legătură cu particularitatea acestei diversități a receptărilor. Răspunsul pe care îl am în vedere, aparent mai rapid și mai la îndemână, este acela că cercetarea transcendentă a lui Kant în filosofie este una *fundamentală*, având influențe decisive în filosofia științei și chiar asupra științei/științelor – această cercetare este considerată „un program teoretic de un alt tip epistemologic decât cele de până acum”¹.

O explicație pentru multitudinea receptărilor disciplinare ale primei *Critici* kantiene, la care voi reveni și pe care acum doar o enunț, este deci legată de răspunsul de mai sus; în plus, adaug aici și interpretez în acest sens o observație a lui Ilie Pârvu, care subliniază că acest program are drept „cheie de boltă” „teorema fundamentală” (a *Criticii rațiunii pure*): „condițiile posibilității experienței în genere sunt în același timp **condiții** ale posibilității obiectelor experienței”².

Scurta interpretare a acestei „teoreme” o formulez astfel încât ea devine decisivă prin forma gramaticală a utilizării diferite a substantivului *condiții*: dacă în prima utilizare avem forma articulată, în cea de-a doua o avem pe cea nearticulată. Dincolo de interpretarea minimalistă, că nu ar avea relevanță această diferență, consider că o alta, fundamentală, este menită să o justifice – în forma (articulată/nearticulată) utilizată. Prima formă, forma articulată, este cea care dă seama de nivelul transcendent – *a priori* al programului kantian; cea de-a doua trimite la diferitele tipuri de reconstrucții și/sau modele teoretice în care poate fi specificat obiectul la care se face raportarea: condițiile posibilității experienței sunt cele ale nivelului transcendent – *a priori*, ele sunt deci necesare și suficiente în mod exhaustiv (forma articulată) pentru orice experiență *în genere*; dar ele funcționează drept condiții ale posibilității *obiectelor* experienței în măsura în care obiectele pot diferi în orice manieră, pentru fiecare obiect putând seta anumite condiții, nu pe toate cele transcendentale *a priori*, mai mult, nu aceleași condiții – ci, prin forma nearticulată, acestea devin simple condiții, o specificare a celor originare, pot fi condiții derivate din cele originare etc. Deci, ceea ce la nivel transcendent se arată ca totalitatea condițiilor experienței posibile *în genere*, la nivelul *obiectelor* experienței acestea sunt specificate drept condiții(le) necesare și suficiente *ale obiectelor corespunzătoare*, fără ca să fie vorba despre *aceleași* condiții transcendentale. Consider că această interpretare este de importanță în înțelegerea proiectului kantian al primei *Critici*, după cum voi încerca să arăt și în partea a doua a acestei lucrări, mai exact atunci când voi discuta recenta cercetare a lui Evans³.

Prezenta lucrare este constituită din două părți. În prima parte prezint și discut două ample studii, texte pe care le consider fundamentale pentru cercetarea de față: primul propune critic o trecere în revistă a literaturii, excelent realizată de Tobias Schlicht, privind relația științelor cogniției cu filosofia transcendentă kantiană, mai exact influențele semnificative ale primei asupra acestor științe, iar al doilea este un vast studiu, parte a unei cercetări și mai ample (în desfășurare),

¹ Ilie Pârvu, *Posibilitatea experienței. O reconstrucție teoretică a Criticii rațiunii pure*, București, Editura SNSPA, 2004, p. 17.

² *Ibidem*.

³ Richard Evans, „The Apperception Engine”, în *Kant and Artificial Intelligence*, De Gruyter, pp. 39–103.

realizat de Richard Evans și colaboratorii de la *Deep Mind*, elaborarea unui model cognitiv kantian + sistemul informatic funcțional corespunzător⁴, o replicare într-o configurație de „agent cognitiv” (IA) a unui posibil model cognitiv kantian construit pornind de la o interpretare a primei părți a *Criticiei rațiunii pure* ca psihologie cognitivă *a priori*. În a doua parte încerc o „evaluare epistemologică” a acestei perspective, problematizând asupra unor elemente epistemologice și etice relevante în perspectiva discuțiilor despre soarta AGI/ASI.

PARTEA I

KANT ȘI ȘTIINȚELE COGNIȚIEI

Primul studiu⁵ pe care îl prezint este legat de influența lui Kant asupra științelor cogniției, respectiv asupra filosofiei minții. Cum spuneam mai sus, pornind de la această cercetare, importantă pentru un „istoric” al problemei cunoașterii în modernitate, și, cel puțin, pentru istoria științelor cogniției, voi survola latura epistemologică ce subîntinde nivelul disciplinar al științelor cogniției; voi încerca să sugerez că aparent simplul mix de două niveluri, al nivelurilor științelor cogniției și cel epistemologic-transcendental, poate exprima de fapt un spațiu teoretic de posibilitate funcțional, capabil de a sta la baza *oricărei* construcții model-teoretice. Această idee va fi dezvoltată în finalul studiului nostru și al discutării, în contextul nostru, al celui de-al doilea studiu pe care îl voi prezenta.

Lucrarea lui Tobias Schlicht examinează mai multe moduri în care perspectiva lui Kant asupra cunoașterii a influențat domeniul științelor cogniției; această perspectivă este considerată influentă și relevantă pentru dezvoltarea diverselor paradigme din știința cognitivă.

Prezentarea acestor influențe apare în prima parte a lucrării sale, influențe ce pot fi considerate inedite de mulți kantieni; de exemplu, Schlicht susține că relevanța operei lui Kant depășește cu mult o asemănare superficială a anumitor concepte din IA cu cele kantiene corespunzătoare: sunt filosofi care susțin că Imm. Kant a anticipat deja mai multe dintre principiile cognitivismului clasic, ale enactivismului și ale modelului de procesare predictivă a minții⁶. În a doua parte a lucrării sale, Schlicht dezbate unele probleme filosofice care se impun din evoluțiile recente ale inteligenței artificiale, cum ar fi problemele filosofice asociate cu așa-numitele *arhitecturi de rețele neuronale profunde* (DNN) și cele care privesc relevanța concepției lui Kant despre cogniție și înțelegere. Autorul susține că performanța rețelelor neuronale profunde ridică întrebări importante

⁴ Acest model este sugestiv intitulat „Aperception Engine”.

⁵ Tobias Schlicht, „Minds, Brains, and Deep Learning: The Development of Cognitive Science Through the Lens of Kant’s Approach to Cognition”, în *Kant and the Artificial Intelligence*, pp. 3–38.

⁶ *Ibidem*, pp. 3–4.

despre percepție, cogniție, învățare, înțelegere și despre vechea dezbateră dintre empiriști și raționaliști în contextual IA; acest fapt i-a determinat pe unii cercetători din domeniul învățării automate să reinvie unele dintre ideile centrale ale lui Kant privind cogniția, dezvoltând chiar o arhitectură cognitivă kantiană pentru a depăși limitele arhitecturilor existente de învățare profundă⁷.

Într-adevăr, dacă urmărim titlurile volumelor din ultimii ani, putem observa că mare parte din știința cognitivă contemporană este influențată de filosofia lui Kant, deși, firește, această observație este o privire înapoi, o reconstrucție prin analogie, în mod evident, căci filosoful german nu putea avea în minte nimic din ce i se atribuie acum. Cu această precauție, Schlicht face referire la Andrew Brook⁸, care l-a numit pe Kant „nașul intelectual” al științei cognitive (Kant ar fi propus deja o teorie funcționalistă a minții, considerată ulterior fundamentul filosofic al inteligenței artificiale); Schlicht invocă apoi mărturia lui Georg Northoff⁹, anecdotic aș putea spune, un expert în neuroștiințe care spunea că recitirea *Criticii rațiunii pure* l-a trezit din „somnul dogmatic”, exact așa cum lectura lui Hume îl trezise pe Kant: impresionat de dovezile empirice privind activitatea cerebrală auto-generată, Northoff și alții vorbesc despre „creierul kantian” și asociază această activitate cu noțiunea kantiană de spontaneitate¹⁰; Francisco Varela¹¹ a recunoscut influența enormă a lui Kant asupra propriei sale perspective asupra vieții și cogniției, iar mai recent Link Swanson¹² a urmărit originea popularului „model de procesare predictivă” până la proiectul general al lui Kant, fapt remarcabil având în vedere că proiectul lui Kant nu trata în primul rând probleme de filosofie a minții, ci era condus mai degrabă de considerații epistemologice.

Cum spuneam, în prima parte a cercetării sale, Schlicht examinează influența (mai ales) a primei *Critici* a lui Kant asupra unor direcții ale științelor cogniției precum: funcționalismul, legătura minte-corp, enactivismul și teoria procesării predictive.

Discuția începe cu definițiile fondatoare ale IA-ului: John McCarthy¹³ și Margaret Boden¹⁴ descriu obiectivul IA ca fiind replicarea comportamentului inteligent uman; scopul suprem rămâne „inteligența generală”, definită ca abilitatea

⁷ Este vorba despre cercetarea lui Richart Evans, despre care am vorbit mai sus și pe care o voi prezenta critic în această primă parte a lucrării.

⁸ Andrew Brook, *Kant and the Mind*, Cambridge, Cambridge University Press, 1994, pp. 8–12, *apud* T. Schlicht, *op. cit.*, p. 4.

⁹ Georg Northoff, *The Spontaneous Brain. From the Mind-Body to the World-Brain Problem*, Cambridge, MA, MIT Press, 2018, p. viii.

¹⁰ Vezi Sina Fazelpour, Evan Thompson, “The Kantian brain: brain dynamics from a neurophenomenological perspective”, în *Current Opinion in Neurobiology*, vol. 31, Amsterdam, Elsevier, 2015, pp. 223–229, *apud* T. Schlicht, *op. cit.*, p. 5.

¹¹ Cf. Andreas Weber, Francisco J. Varela, „Life after Kant: Natural purposes and the autopoietic foundations of biological individuality”, în *Phenomenology and the Cognitive Sciences*, vol. 1, Dordrecht, Springer, 2002, pp. 97–125, *apud* T. Schlicht, *op. cit.*, p. 5.

¹² *Apud* T. Schlicht, *op. cit.*

¹³ John McCarthy, Marvin L. Minsky, Nathaniel Rochester, Claude E. Shannon, „A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”, în *AI Magazine*, vol. 27, nr. 4, Palo Alto, AAAI, 1955, pp. 12–14.

¹⁴ Margaret A. Boden, *Artificial Intelligence. A Very Short Introduction*, Oxford, Oxford University Press, 2016, p. 1.

de a atinge obiective diferite într-o mare varietate de medii¹⁵. În ceea ce privește noțiunea kantiană de inteligență, Schlicht consideră, și cu aceasta suntem de acord cel puțin parțial, că nu este foarte utilă pentru analiza sa, deoarece, kantian vorbind, a numi o creatură ca fiind „inteligentă” echivalează cu a o considera liberă și capabilă de autodeterminare, adică aparținând lumii noumenelor, mai degrabă decât celei a fenomenelor. Sugestia lui Schlicht este ca în analiza sa, pentru scopurile asumate de acesta, terminologic, să se treacă de la „inteligentă” la „capacități cognitive” sau „cogniție”.

Introducerea funcționalismului ca direcție de prim-rang în analiza cu privire la relevanța gândirii kantiene asupra dezvoltării IA a avut la bază faptul că, ceea ce este esențial IA, mașinii ca atare, nu este un hardware, ci ceea ce poate face un hardware *artificial*. Prin urmare, accentul nu este pus pe mașini, ci pe mașinile *virtuale*, care nu sunt altceva decât sisteme de procesare a informației ce pot fi implementate într-o varietate de hardware¹⁶. În consecință, consideră Schlicht, teoria filosofică privilegiată care a susținut posibilitatea IA a fost funcționalismul. Funcționalismul susține că stările mentale în general sunt concepute în termeni de funcții corespunzătoare: fiecare stare mentală este identificată prin setul său de relații cauzale către inputurile și outputurile sistemului, precum și către alte stări ale sistemului¹⁷. Realizarea acestei rețele cauzale de funcții este considerată contingentă, deoarece funcțiile sunt socotite ca fiind multiplu realizabile.

Astfel, cognitivismul clasic, prima paradigmă în științele cogniției, a conceput cogniția ca procesare a informației pe linia celei prezente în computerele digitale. În particular, cogniția a fost înțeleasă ca fiind constituită din manipulări conduse sintactic ale unor structuri reprezentationale simbolice în creier, care sunt „multinivelare”¹⁸ între inputuri senzoriale și outputuri motorii. Schlicht dă următorul exemplu: când privesc cana de cafea din fața mea, informația senzorială care ajunge la retină este procesată în module specializate care, în cele din urmă, produc o imagine tridimensională detaliată a căinii, imagine care poate ghida acțiuni precum apucarea ei¹⁹.

Am văzut în acest fel că Schlicht se referă la funcționalism ca definind stările mentale prin rolurile lor cauzale; acest curent a constituit teoria filosofică dominantă în IA. În acest sens, autori precum Brook, Sellars²⁰ și Meerbote²¹ au susținut că agnosticismul lui Kant față de substratul minții îl califică drept un

¹⁵ Murray Shanahan, „Artificial intelligence”, în Mark Sprevak, Matteo Colombo (eds.), *The Routledge Handbook of the Computational Mind*, London, Routledge, 2019, pp. 91–100; Shane Legg, Marcus Hutter, „Universal intelligence: a definition of machine intelligence”, în *Minds and Machines*, vol. 17, Dordrecht, Springer, 2007, pp. 391–444, *apud* T. Schlicht, *op. cit.*, p. 6.

¹⁶ Tobias Schlicht, *op. cit.*, p. 6.

¹⁷ Cf. Hilary Putnam (ed.), „The nature of mental states”, în *Philosophical Papers*, vol. 2, Cambridge, Cambridge University Press, 1965, pp. 429–440, *apud* T. Schlicht, *op. cit.*, p. 6.

¹⁸ Cf. Susan L. Hurley, *Consciousness in Action*, Cambridge, Cambridge University Press, 1998, *apud* T. Schlicht, *op. cit.*, p. 6.

¹⁹ Tobias Schlicht, *op. cit.*, p. 6.

²⁰ Wilfrid Sellars, „Metaphysics and the Concept of a Person”, în *Essays in Philosophy and its History*, Dordrecht, Springer, 1974, pp. 214–243, *apud* T. Schlicht, *op. cit.*, p. 6.

²¹ Ralf Meerbote, „Kant’s Functionalism”, în John-Christian Schmith (ed.), *Historical Foundations of Cognitive Science*, Dordrecht, Springer, 1989, pp. 161–187, *apud* T. Schlicht, *op. cit.*, p. 6.

precursor al funcționalismului. În ciuda atitudinii „implacabil ostile” a lui Kant față de materialism, arată Schlicht, Brook susține că „materialismul se potrivește remarcabil de ușor în teoria sa generală”²². Impresionat de poziția lui Kant conform căreia „în ceea ce privește natura reală a minții, stricta neutralitate trebuie să fie regula zilei”, Brook ia acest agnosticism drept o instanță a ideii funcționaliste contemporane a „realizabilității multiple” a funcțiilor mentale. Schlicht respinge această echivalare și subliniază că Ameriks interpretează poziția lui Kant mai degrabă ca pe un „imaterialism simplu”²³, fără alte implicații – Brook ignoră concepția particulară a lui Kant despre materie ca simplă aparență. O poziție mai critică o are Allison, care argumentează că spontaneitatea și caracterul autoconștient al cogniției kantiene sunt incompatibile cu procesarea informațională mecanicistă a funcționalismului²⁴.

Vedem aici o diferență de interpretare – funcționalismul, recunoscut în general ca fiind originat în prima *Critica* kantiană, este recuperat de pe poziții diferite: unii reprezentanți ai cognitivismului se apropie de naturalism, în timp ce filosofi analitici sau exegeți kantieni precum Ameriks sau Allison optează fie pentru dezangajarea naturalistă, fie pentru implicarea dimensiunii conștiinței în asumarea funcționalismului kantian, respingând simplul mecanicism fizic²⁵. Schlicht optează pentru „idealismul transcendent mai fundamental al lui Kant”, pe care el însuși îl numește „un dualism” (Kant 1781/1998, A370).

Problema originată în funcționalism, așa cum este contextualizat mai sus, vizează în parte trăsăturile implementării biologice ale funcțiilor mentale din creierul uman (și animal), consideră Schlicht; de aici, el ridică întrebarea care va circumscrie următoarea secțiune din analiza sa: legătura minte-corp.

Legătura minte-corp este prezentată prin interogația: poate fi cogniția concepută ca un set de funcții cauzale *in abstracto* față de trăsăturile biologice ale realizării sale, astfel încât acest set de funcții, în principiu, să fie realizat de o mașină utilizând o configurație non-biologică? Sau: este cogniția exclusiv un fenomen biologic, a cărui realizare depinde deci de prezența unui sistem biologic dinamic complex, adică un organism (cu creier și sistem nervos), care prezintă mijloace fundamentale biochimice de procesare a informației? Mai concret, pot stările mentale reprezentative să fie „aspecte” ale calculelor neuronale, adică biologice, mai degrabă decât funcții abstracte care se bucură de o anumită independență față de realizatorii lor?²⁶

²² Andrew Brook, *op. cit.*, p. 15, *apud* T. Schlicht, *op. cit.*, p. 6.

²³ Karl Ameriks, *Kant's Theory of Mind. An Analysis of the Paralogisms of Pure Reason*, Oxford, Clarendon Press, 2000 (2nd Ed.), *apud* T. Schlicht, *op. cit.*, p. 6.

²⁴ Henry E. Allison, „On naturalizing Kant's transcendental psychology”, în *Idealism and Freedom. Essays on Kant's Theoretical and Practical Philosophy*, Cambridge, Cambridge University Press, 1996, pp. 53–66, *apud* T. Schlicht, *op. cit.*, p. 7.

²⁵ Mai rețin aici referirea lui Schlicht la Hanna, Thompson și alți autori, care consideră activitatea autogenerată a creierului ca fiind un candidat viabil pentru un corelat neural al funcției pe care Kant o numește spontaneitate, dar această interpretare nu a fost justificată în detaliu (pentru o discuție critică vezi Schlicht & Newen 2015, cf. Northoff 2013/2014 pentru alte legături cu neuroștiința).

²⁶ Tobias Schlicht, *op. cit.*, pp. 7–8.

Progresele imagistice din neuroștiință au impulsionat cercetarea asupra creierului și au dat naștere modelelor conexioniste ale proceselor cognitive. Aceste modele rămân computaționale și reprezentationale, dar presupun că informația este procesată sub-simbolic, cu reprezentări non-lingvistice. Legat de învățarea profundă (Deep Learning – DP), robotica și IA au adoptat arhitecturi de sub-sumare, care renunță la modele complete ale lumii; această abordare, orientată spre structura creierului, a prefigurat tehnicile moderne de învățare automată, în special învățarea profundă, arată Schlicht.

Dincolo de disputa dintre abordările funcționaliste și cele biologice ale cogniției, dacă ne re-întoarcem la Kant, Schlicht sugerează să luăm în serios la acesta ideea unei simple „prezențe virtuale” a minții în creier, ceea ce face întreaga chestiune a ceea ce ar putea servi drept „sediul al sufletului” să dispară – după Kant. Rezerva lui Schlicht este că filosoful german nu clarifică ce înseamnă aici „prezență virtuală”. O altă referință a lui Kant în problema minte-corp este cea din *Prelegeri de metafizică*, unde subliniază că „localizarea sufletului în corp [...] nu poate fi determinată. Nu pot simți locul în corp unde rezidă sufletul.” (AA 28, 281)²⁷ Totuși, în ciuda acestei restricții epistemologice, Kant avansează un argument ce poate fi citit ca și cum ar face aluzie la ideea contemporană de *superveniență* care plasează temeiul tuturor senzațiilor în creier. Aici reluăm esențialul: Kant ar susține că toate senzațiile provin din sistemul nervos, iar locul în care acestea se concentrează este creierul; prin urmare, sediul senzațiilor trebuie să fie în creier, ca punct al tuturor condițiilor senzoriale. Kant nu afirmă că sufletul se află în creier, ci doar că acesta își plasează *sediul senzațiilor* acolo²⁸. Remarca lui Schlicht aici este că pasajele din *Prelegerile* lui Kant anticipează în fapt obiecția împotriva „materialismului cartesian”, o poziție împărtășită (acest materialism), se pare, de mulți cercetători contemporani din neuroștiințe care încearcă să identifice anumite zone sau procese cerebrale ca fiind cauzal responsabile de (sau identice cu) conștiință(a).

Sub o astfel de citire a lui Kant, contrastul terminologic împotriva materialismului, între o prezență locală și una doar virtuală a minții în creier, este de mare actualitate, având în vedere caracterizarea minții ca „mașină virtuală implementată în arhitectura paralelă a unui creier”²⁹. În acest fel, continuă Schlicht, avem o anticipare a viziunii funcționaliste asupra minții și din această perspectivă, kantiană; suntem însă de acord cu Schlicht că este dificil de determinat dacă folosirea de către Kant a termenului *virtual* este similară cu cea a lui Dennett.

În cazul „localizării” minții, mai merită reținut aspectul legat de prioritatea acordată de filosoful german unui anumit tip de organizare sau de „dispoziție teleologică a părților sale [minții, n.n.]”. Trimiterea firească este la discuția „inovatoare și foarte influentă a lui Kant despre „organisme ca scopuri naturale”, adică ființe autoproducătoare și autoorganizatoare, din *Critica*

²⁷ *Ibidem*, p. 9 (traducerea din engleză îmi aparține).

²⁸ *Ibidem*, pp. 9–10.

²⁹ Daniel C. Dennett, *Consciousness Explained*, New York, Basic Books, 1991, p. 210, *apud* T. Schlicht, *op. cit.*, p. 10.

*facultății de judecare*³⁰. În contrast cu o organizare mecanică pură, Kant consideră „organizarea dinamică” drept fundamentală pentru înțelegerea conceptului de *minte*. Deși el nu explicitează nici acest concept ca atare, sensul poate fi iluminat de discuția privind contrastul dintre explicația mecanicistă și cea teleologică din cea de-a treia *Critică*. Schlicht arată foarte bine în continuare că această discuție, despre teleologia imanentă a organismelor vii, a inspirat generații de filosofi, ducând până la dezvoltarea actuală a abordărilor „enactive” ale minții³¹.

Concepția enactivistă sau *autopoietica*, prin Francisco Varela, ar fi rămas profund datorare concepției kantiene despre organisme, așa cum am arătat mai sus că susține Schlicht. Curentul din robotică susținut de lucrările lui Brooks a împins abordarea „enactiv-întrupată” a cogniției până la a contesta, în favoarea unei viziuni dinamice, atât paradigma reprezentationalistă, cât și separarea explicită a percepției de acțiune în „concepția sandwich” tradițională a cogniției³². Schlicht arată că, în contrast cu progresia liniară tradițională de la un input senzorial prin computație cognitivă către acțiune, enactivismul concepe percepția și cogniția nu pur și simplu ca stări funcționale ale creierului, ci ca activități întrupate, împletite și conexe ale organismelor întregi (agenți, sisteme), care pot fi explicate fără apel la reprezentări mentale.

Într-adevăr, în acest cadru, echivalența dintre intenționalitate și reprezentarea mentală nu se mai susține *ab initio*: Schlicht propune exemplul cu cana de cafea – a percepe o cană de cafea nu presupune doar multiple acțiuni precum mișcările ochilor, capului și corpului (întoarcerea privirii etc.); perceperea este, de la început, în serviciul detectării posibilităților de acțiune (precum apucarea). Enactiviștii subscriu la ceea ce Thompson numește „continuitatea profundă dintre viață și minte”³³, adică afirmația că trăsăturile organizaționale ale minții sunt o versiune îmbogățită ale celor ale vieții; mai mult, în evaluarea influenței lui Kant asupra științei cognitive actuale, acesta este aspectul cel mai important al enactivismului, susține Schlicht.

Redăm și noi aici definiția noțiunii de autopoieză: un sistem autopoietic – organizația minimală vie – este un sistem care produce continuu componentele care îl specifică, realizând în același timp sistemul ca o unitate concretă în spațiu și timp, ceea ce face posibilă rețeaua de producere a componentelor³⁴. În acest sens, argumentul susținut de enactiviști cu privire la prefigurarea kantiană a acestui concept vizează ceea ce Kant expune în *Critica rațiunii practice*, unde concepe organismele ca sisteme „auto-organizate” și „auto-producătoare”, adică autopoietice, care nu pot fi explicate în termeni pur mecanici, ci pe care trebuie să le „facem inteligibile” pentru noi, bazându-ne pe principii teleologice, care nu fac parte din știința naturii, ci sunt împrumutate din contexte practice. Impresionat de

³⁰ Vezi întreaga discuție și referințele în T. Schlicht, *op. cit.*, p. 10.

³¹ *Ibidem*.

³² *Ibidem*, p. 11.

³³ Evan Thompson, *Mind in Life*, Cambridge, MA, Harvard University Press, 2007, p. 128, *apud* T. Schlicht, *op. cit.*, p. 11.

³⁴ Cf. Humberto R. Maturana, Francisco Varela, *Autopoiesis and Cognition. The Realization of the Living*, Amsterdam, Springer, 1980, *apud* T. Schlicht, *op. cit.*, p. 11.

capacitatea anumitor animale (de exemplu, peștii zebură, salamandrele etc.) de a regenera părți ale corpului deteriorate sau chiar secționare, Kant discută unele exemple pentru a demonstra că animalele exhibă o anumită formă sau organizare care, dacă este concepută doar ca rezultat al unor procese cauzale mecanice oarbe, apare complet inexplicabilă, fiind contingentă. Totuși, „din moment ce rațiunea trebuie să fie în stare să cunoască necesitatea în orice formă a unui produs natural dacă ar înțelege condițiile asociate cu generarea lui”, înțelegerea noastră trebuie să împrumute conceptul de cauză *finală* pentru a da un sens inteligibil acestei organizări³⁵. Această poziție îl conduce într-adevăr pe Kant la concepția despre organisme ca „scopuri naturale”, adică atât ca produse naturale, cât și ca scopuri în același timp.

Observația lui Schlicht, că cele de mai sus pot ascunde o contradicție, din moment ce noțiunea de „scop” sau „finalitate” este exterioară în știința naturii (KU, AA 05: A390), observație care pare să implice o proiectare în natură a finalității de dragul înțelegerii (unora dintre) produsele ei, este devalată ca simplă aparență; căci, spre deosebire de un ceas, părțile unui organism, organele sale, trebuie luate ca producându-se pe ele însele, mai degrabă decât ca fiind produse de o putere externă, și se aranjează în relație și în dependență reciprocă unele față de altele; deci, spre deosebire de ceasornicar, în cazul organismelor ideea călăuzitoare nu se află în afara produsului (ceasul), ci în interiorul lui (organismul însuși)³⁶.

Reținem aici o doua observație a lui Schlicht, în același sens ca prima, poate mai pertinentă și mai importantă, cu referire la cea de-a treia *Critică*: presupunerea că „puterea formativă” teleologică funcționează ca atare trebuie luată doar într-un sens epistemologic, adică noi doar considerăm *ca și cum* organismele ar fi posibile numai prin rațiune, deoarece, ca produse naturale, ele trebuie să apară prin cauze pur mecanice și, astfel, să fie susceptibile de explicație mecanicistă. Nu putem dovedi că organismele exhibă într-adevăr această putere formativă, deoarece nu putem dobândi o intuiție a ei³⁷.

Știm că Imm. Kant respinge ideea unei intuiții intelectuale; Francisco Varela a considerat poziția filosofului german în mod extrem atunci când a fost vorba despre procesele teleologice: Kant ar fi „dezvoltat posibilitatea unei a treia căi, între o teleologie puternică și un materialism brut”. Desigur, Schlicht precizează că Varela cunoștea poziția originală privind intuiția a lui Kant, dar considera poziția sa „instabilă” și necesitând revizuire „pe baza dezvoltărilor moderne ale cercetării și gândirii biologice”³⁸. Potrivit lui Weber și Varela, concepția lui Kant despre organism ca ființă auto-organizată și auto-producătoare este strâns analogă definiției organismului în propria teorie a lui Varela privind „autopoieza”. În această viziune, autonomia biologică și individualitatea justifică asumarea unei

³⁵ Pentru cotext, vezi T. Schlicht, *op. cit.*, pp. 11–12.

³⁶ „O ființă organizată nu este, așadar, o simplă mașină, căci aceasta are doar o putere motrice, în timp ce ființa organizată posedă în sine o putere formativă, și, într-adevăr, una pe care o comunică materiei, care nu o are (o organizează pe aceasta)” (Imm. Kant, KU, AA 05: A374), *apud* T. Schlicht, *op. cit.*, p. 12.

³⁷ T. Schlicht, *op. cit.*, p. 12.

³⁸ Vezi *ibidem*.

„teleologii intrinseci” în sensul că „organismele sunt subiecți având scopuri potrivit valorilor întâlnite în cursul vieții lor”³⁹.

Poziția epistemică și critică a lui Kant asupra acestei probleme a teleologiei, deși este forțat preluată de teoria naturalistă a autopoiezei a lui Varela, ilustrează totuși felul major în care filosofia kantiană a biologiei a lăsat o amprentă cu implicații ample în dezvoltarea istorică a științei cognitive, pe această linie. Mai trebuie subliniat, împreună cu Schlicht, că „o implicație particulară a abordării autopoietice a cogniției și a tezei continuității minte-viață este că toate organismele pot exhiba cel puțin o formă de bază de cogniție, în timp ce astfel de vederi au o problemă în a permite o cogniție genuină în sistemele artificiale”⁴⁰.

În contrast cu abordări cognitivistice mai tradiționale, posibilitatea cogniției în sisteme biologice „simple” a fost luată recent în serios cu privire la organisme precum bacteriile, plantele și mușchii vâscoși⁴¹, de exemplu. O idee interesantă este adusă în discuție de Schlicht în acest context prin biologul Michael Levin⁴², care argumentează că ar trebui să aplicăm abordarea computațională nu doar animalelor cu creiere și sisteme nervoase, ci și organismelor simple, fără creier; mai degrabă decât să continuăm să opunem creierul restului corpului, Levin ne invită să considerăm corpul ca efectuând, de asemenea, calcule, astfel încât să depășim dihotomia tradițională viață vs. mașină. Intrigat de puterea formativă a organismului, Levin speculează că celulele și țesuturile pot exhiba unele forme de bază de memorie și acțiune, folosind bioelectricitatea pentru a comunica și a decide sau a planifica dezvoltarea⁴³.

În ceea ce privește influența kantiană asupra „teoriei procesării predictive”, voi rezuma și condensa ceea ce Schlicht expune mai pe larg. Este vorba despre prezentarea modelelor de procesare predictivă și legătura analogă a acestora cu creierul și cu procese precum percepția, cogniția și acțiunea. Ideea centrală este aceea a unificării acestor procese de organul central, creierul, care mai este înțeles, analog, și ca o „mașină de predicție”⁴⁴. Această viziune implică un echilibru delicat între procesarea „de jos în sus” și „de sus în jos”, în răspăr cu seria poziționărilor tradiționale, care avea ca model dominant (unic) pe cel „de jos în sus” (bottom-up).

În modelul „de sus în jos” și „de jos în sus”, l-aș numi *complementar*, percepția și cogniția sunt definite în termeni ce țin de creier, testând ipoteze („de sus în jos”) despre sursele sau cauzele stimulării senzoriale primite; astfel, ipotezele sunt generate de un model generativ ierarhic al lumii și sunt actualizate

³⁹ Pentru întreaga discuție, vezi *ibidem*, pp. 12–13.

⁴⁰ *Ibidem*.

⁴¹ Pentru context și referințe, vezi *ibidem*, p. 13.

⁴² Michael Levin *et. al.*, „Uncovering cognitive similarities and differences, conservation and innovation”, în *Philosophical Transactions of the Royal Society B*, vol. 375, London, Royal Society, 2021, *apud* T. Schlicht, *op. cit.*, p. 13.

⁴³ De exemplu, a reușit să „reprogrameze” un vierme planarian să crească un al doilea cap în locul cozii pe care i-o tăiasse. Ceea ce a făcut a fost să schimbe semnalele sau „codul” bioelectric care ar fi condus în mod normal la creșterea unei noi cozi. Munca lui Levin sugerează o convergență între biologie și informatică și este astfel extrem de relevantă pentru viitorul inteligenței artificiale; cf. T. Schlicht, *op. cit.*, p. 13.

⁴⁴ Tobias Schlicht, *op. cit.*, p. 14.

constant ca răspuns la semnalele de eroare de predicție – adică suportă corecții în funcție de inputul senzorial⁴⁵. Schlicht ne invită să revenim la exemplul nostru: a percepe cana de cafea este un proces deja informat de procese cerebrale subiacente care constituie o mulțime de așteptări mai mult sau mai puțin probabile despre inputul senzorial și cauzele sale. Aceste așteptări sunt comparate constant cu informația senzorială reală primită, rezultând erori de predicție (deviații), care sunt procesate în creier.

Accentuăm aici observația lui Schlicht, că imaginea tradițională a creierului care folosește informația senzorială primită pentru a construi o reprezentare a lumii este astfel întoarsă „cu susul în jos”, deoarece noua imagine susține că „reprezentarea bogată a stărilor de fapt din lume este semnalată în predicțiile *top-down* ale inputului senzorial, menținute de ierarhia perceptivă din creier”⁴⁶.

Autori precum Link Swanson⁴⁷ susțin că această paradigmă, foarte recentă din știința cogniției, are rădăcini în filosofia lui Kant: este vorba despre teoria procesării predictive ca teorie unificată a creierului. Inversarea radicală a procesării (testarea ipotezelor *top-down*, mai degrabă decât construirea de modele *bottom-up*), caracteristică procesării predictive, își găsește un analog în așa-numita răsturnare copernicană a lui Kant, prezentându-ne „o viziune a percepției, în spirit kantian, ca activitate interpretativă ‘spontană’, și nu un proces de construire pasivă a perceptelor din inputuri”⁴⁸. Swanson leagă, de asemenea, concepte mai specifice din procesarea predictivă de analogi specifici în teoria lui Kant – de exemplu, modele generative și *schematismul* kantian, ambele fiind puternic informate atât de intuiții, cât și de concepte în procesul de recunoaștere a obiectelor. Într-adevăr, afirmă Schlicht, este izbitor că atât Clark cât și Hohwy aleg un punct de plecare care este foarte familiar pentru kantieni, dar formulat din perspectiva creierului, a cărui sarcină, „privită de la o anumită distanță, poate părea imposibilă: trebuie să descopere informații despre cauzele probabile ale semnalelor incidente fără niciun fel de acces direct la sursa lor”⁴⁹.

Spus astfel, chestiunea centrală este înțelegerea cauzalității, adică înțelegerea relațiilor dintre cauzele din lume și inputurile senzoriale. Presupunând un cadru humean, acest lucru este imposibil, dar este posibil la Kant, care postulează un mecanism conceptual *a priori* (categoriile) ce trebuie aplicat inputului senzorial pentru a permite o asemenea înțelegere.

Într-adevăr, deși presupusele rădăcini ale funcționalismului, enactivismului și procesării predictive există în filosofia lui Kant, așa cum recunoaște și Schlicht, este important să ținem cont că aceste fundamente paradigmatiche diferite adoptă poziții diferite față de relația dintre cogniție, intenționalitate și reprezentare și

⁴⁵ Pentru context și referințe, vezi *ibidem*, p. 14.

⁴⁶ Cf. Jakob Hohwy, *The Predictive Mind*, Oxford, Oxford University Press, 2013, p. 47.

⁴⁷ Cf. Link R. Swanson, „The predictive processing paradigm has roots in Kant”, în *Frontiers in Systems Neuroscience*, vol. 10, Lausanne, Frontiers, 2016, p. 79, *apud* T. Schlicht, *op. cit.*, p. 14.

⁴⁸ Pentru context, vezi T. Schlicht, *op. cit.*, p. 14.

⁴⁹ Cf. Andy Clark, „Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science”, în *Behavioral and Brain Sciences*, vol. 36, nr. 3, Cambridge, Cambridge University Press, 2013, p. 183, *apud* T. Schlicht, *op. cit.*, p. 15.

propun strategii explicative diferite în știința cognitivă. Schlicht adaugă că „pare puțin probabil că Imm. Kant ar fi subscris la toate aceste viziuni deodată, având în vedere opoziția (și incompatibilitatea genuină) a unor poziții contemporane: în timp ce cognitivismul clasic face aluzie la reprezentări mentale, enactivismul întrupat le evită; în timp ce primul se bazează pe funcționalism și permite explicit posibilitatea cogniției în sisteme artificiale, cel de-al doilea se bazează pe o puternică continuitate între viață și minte, ceea ce face problematică această posibilitate”⁵⁰.

În finalul acestei prezentări din prima parte a lucrării sale, pe care noi am condensat-o și ajustat-o critic unde am considerat că este cazul, tonul lui Schlicht se înmoaie atunci când spune „că procesarea predictivă care își are rădăcinile în viziunea lui Kant asupra minții trebuie luată cu un grăunte de sare, la fel ca ideea care asimilează viziunea lui Kant asupra minții cu funcționalismul”⁵¹.

Înainte de a prezenta foarte pe scurt a doua parte a studiului lui Schlicht, este important să reluăm câteva aspecte menționate de acest autor ce țin de distingerea conceptelor de cunoaștere și cogniție din perspectiva conștiinței în două dintre lucrările fundamentale ale lui Kant: așa-numita *Jäsche Logik* și *Critica rațiunii pure*. Schlicht face referire la două articole relativ recente și foarte instructive în chestiunea de față, ale lui Marcus Willaschek⁵² și Eric Watkins⁵³. Cei doi se concentrează asupra utilizării complexe a noțiunii de „cogniție” în lucrările lui Kant. Utilizarea cea mai prominentă este cea ce ei numesc „cogniție în sens restrâns”, care cere o unificare a intuiției și a conceptului, adică combinarea receptivității senzoriale și a spontaneității intelectului. „Cogniția în sens larg”, în contrast, permite mai multe „grade ale cogniției”, schițate în moduri diferite, deși nu neapărat incompatibile (vezi *Jäsche Logik* [AA 16: 64–65] și *Critica rațiunii pure* [KrV: A320/B376]).

În *Critica rațiunii pure*, ele sunt prezentate drept cazuri mai mult sau mai puțin solicitate de „a reprezenta ceva”, fie inconștient, fie conștient, prin percepție, înțelegere sau rațiune, cu sau fără implicarea conceptelor sau a intuiției. Gradul cel mai de jos al cogniției este acțiunea de „a reprezenta ceva”, chiar fără condiții suplimentare; gradul cel mai înalt sau mai complex este de „a cuprinde ceva” prin rațiune și *a priori*. Schlicht subliniază că este important, în taxonomia lui Kant, faptul că cogniția nu implică adevăr sau asentiment și, prin urmare, trebuie deosebită de noțiunea de „cunoaștere”⁵⁴. Luată în sens larg, orice reprezentare conștientă care reprezintă un obiect contează ca un caz de cogniție, chiar dacă obiectul nu există (sau dacă nu poate fi dat în experiență).

Schlicht spune explicit că termenul, conceptul de cogniție care interesează în discuția de față este cel utilizat de Kant în sens restrâns, ca „cogniție în sens propriu”

⁵⁰ Tobias Schlicht, *op. cit.*, p. 16.

⁵¹ *Ibidem*.

⁵² Eric Watkins, Marcus Willaschek, „Kant’s Account of Cognition”, în *Journal of the History of Philosophy*, vol. 55, nr. 1, Baltimore, Johns Hopkins University Press, 2017, pp. 83–112, *apud* T. Schlicht, *op. cit.*, p. 16.

⁵³ Marcus Willaschek, Eric Watkins, „Kant on Cognition and Knowledge”, în *Synthese*, vol. 197, nr. 8, Dordrecht, Springer, 2020, pp. 3195–3213, *apud* T. Schlicht, *op. cit.*, p. 16.

⁵⁴ *Ibidem*.

(KrV: A78/B103). Prin urmare, cogniția poate fi descrisă ca o „reprezentare conștientă a unui obiect dat și a (cel puțin unora dintre) trăsăturilor(le) sale generale”⁵⁵. Dincolo de necesitatea sintezei în unificarea datului intuiției, ceea ce concepția gradelor de cogniție din *Jäsche Logik* și pasajul „progresiei” din *Critica rațiunii pure* au în comun, și este de asemenea necesar pentru cogniție în sens restrâns, este ideea că *cogniția în sens restrâns presupune conștiință*.

Mai menționăm poziția lui Tolley⁵⁶, care susține că Imm. Kant clasifică simțirea, intuirea, perceperea și simpla gândire ca fiind la baza cogniției, furnizând condiții pentru aceasta, considerând în același timp cogniția ca plasată pe un „nivel psihologic elementar” comparativ cu cunoașterea, înțelegerea și explicarea. În contrast cu Watkins și Willaschek, Tolley susține că conceptul de cogniție la Kant este *unificat* mai degrabă decât echivoc.

Dincolo de dispută, la care nu mă refer aici întrucât nu oferă elemente necesare demersului nostru, așa cum spune și Schlicht, „toate părțile sunt de acord că filosoful german a accentuat sensul conceptului de cogniție în sens restrâns, în timp ce ceilalți candidați cad sub umbrela conceptului de «cogniție»”, așa cum este folosit în știința cognitivă contemporană. Și, din moment ce cogniția în sens restrâns, în viziunea lui Kant, este o „formă distinctivă de conștiință a unui obiect real prin intermediul unui tip specific de combinație de reprezentări”, conștiința este o condiție a cogniției în acest sens propriu⁵⁷.

În concluzie, din perspectiva lui Kant, pentru ca un sistem artificial să fie capabil de cogniție (în sens restrâns), ar trebui să fie capabil și de conștiință. Aceasta nu este cu siguranță o viziune asupra cogniției răspândită printre oamenii de știință contemporani din științele cogniției. Susținătorii abordării de procesare predictivă a percepției și cogniției nu susțin nici ei că aceste procese cer conștiință, deși adesea pretind că cadrul poate fi aplicat și pentru a explica conștiința. Dar, chiar dacă un sistem artificial nu ar putea fi conștient în sensul relevant, ar putea totuși fi capabil de cogniție în sens larg. Adică s-ar putea spune – în mod minimal – că are reprezentări ale acestui sau aceluia ceva⁵⁸.

Partea a doua a analizei lui Schlicht se referă la dezvoltări mai recente în inteligența artificială, anume la ascensiunea și povestea de succes a arhitecturilor de învățare profundă (DL), care au condus la „primăvara recentă a IA”⁵⁹. Această dezvoltare fascinantă ridică probleme filosofice interesante despre natura percepției, a învățării și a înțelegerii și despre chestiunea mai generală a abordărilor empiriste vs. raționaliste. Îl vom însoți acum pe Schlicht într-o foarte scurtă istorie a dezvoltării IA, în special a dezvoltării „învățării profunde” (DL).

DL reprezintă una dintre cele mai spectaculoase transformări din istoria inteligenței artificiale, apărută după o perioadă de stagnare cunoscută sub numele de „iarna IA”. Această schimbare a fost posibilă datorită apariției „arhitecturilor

⁵⁵ Interpretarea le aparține lui Watkins și Willaschek; cf. *ibidem*, p. 17.

⁵⁶ Clinton Tolley, „Kant on the place of cognition in the progression of our representations”, în *Synthese*, vol. 197, Dordrecht, Springer, 2020, pp. 3215–3244, *apud* T. Schlicht, *op. cit.*, p. 18.

⁵⁷ Pentru discuție, vezi Tobias Schlicht, *op. cit.*, pp. 17–18.

⁵⁸ *Ibidem*.

⁵⁹ *Ibidem*, p. 18.

neuronale complexe”, inspirate din biologia creierului uman, și datorită progreselor tehnologice care au permis utilizarea unor resurse de calcul masive. Spre deosebire de abordările tradiționale de tip GOFIA („Good Old-Fashioned AI”), bazate pe reguli explicite și algoritmi programați rigid, rețelele neuronale artificiale au introdus un mod de funcționare mai flexibil, în care software-ul poate fi antrenat pe date și își poate ajusta parametrii în funcție de experiență. Această paradigmă a învățării automate include mai multe metode – „supravegheate”, „nesupravegheate” și „prin întărire” –, dar învățarea profundă, bazată pe rețele neuronale cu multe straturi, s-a impus ca cea mai promițătoare și utilizată abordare, fiind considerată responsabilă pentru progresele recente în recunoașterea vizuală, în procesarea limbajului sau în jocurile strategice.

Diferența esențială față de rețelele neuronale clasice din anii '80 și '90, arată Schlicht în continuare, constă în „adâncimea arhitecturii”. Dacă modelele vechi aveau doar un strat de input, unul ascuns și unul de output, rețelele moderne pot avea sute de straturi, ceea ce le conferă o putere computațională exponențial mai mare. Această structură stratificată le permite să extragă progresiv trăsături din ce în ce mai abstracte din datele de intrare, într-un mod similar cu modul în care cortexul vizual uman procesează informația. Experimentele lui Hubel și Wiesel (1962) realizate pe pisici au arătat că neuronii din cortexul vizual reacționează la trăsături simple precum margini sau orientări, iar straturile ulterioare combină aceste informații pentru a recunoaște forme complexe și obiecte întregi. Această descoperire a inspirat dezvoltarea unor rețele artificiale precum „Neocognitronul” lui Fukushima (1980), care a demonstrat potențialul unei arhitecturi stratificate. În mod similar, rețelele neuronale artificiale actuale procesează imaginile ca matrice de pixeli, detectând margini, motive, părți de obiecte și, în final, obiecte complete, *fără ca aceste trăsături să fie proiectate manual de ingineri, ci învățate direct din date*⁶⁰.

Un element central al acestor rețele este capacitatea de a învăța prin ajustarea dificultăților conexiunilor dintre noduri. Procesul începe cu valori arbitrare, iar prin expunerea repetată la date și prin mecanismul de „back-propagation”, rețeaua își corectează erorile și își îmbunătățește performanța. Spre exemplu, dacă scopul este recunoașterea câinilor și a pisicilor, rețeaua va fi antrenată pe milioane de imagini etichetate, ajustându-și parametrii până când outputul său va reflecta cu un grad mare de încredere categoria corectă. Outputul nu este un răspuns unic, ci un vector de scoruri care exprimă probabilitatea ca imaginea să aparțină fiecărei categorii. Această metodă supravegheată, bazată pe feedback explicit, a permis atingerea unor performanțe impresionante, cum ar fi peste 90% acuratețe în competiția ImageNet⁶¹.

Totuși, această dependență de seturi masive de date evidențiază o limitare fundamentală: spre deosebire de oameni, care pot învăța din câteva exemple, rețelele neuronale au nevoie de milioane de instanțe pentru a atinge un nivel similar de performanță. Aceasta arată că învățarea lor este diferită de cea umană încă de la început, chiar dacă rezultatele pot fi considerate „inteligente”.

⁶⁰ Pentru context și referințe, vezi *ibidem*, pp. 18–19.

⁶¹ *Ibidem*.

Un alt aspect important este modul în care rețelele neuronale profunde fac erori. Deși oamenii sunt și ei supuși iluziilor vizuale sau confuziilor, tipurile de greșeli ale rețelelor sunt distincte. Ele pot rata obiecte mici, pot fi derutate de distorsiuni de culoare sau de contrast și pot eșua în recunoașterea reprezentărilor abstracte, cum ar fi picturi sau jucării de pluș⁶². Mai grav, rețelele pot fi păcălite sistematic prin „exemple adversariale”: imagini modificate subtil, imperceptibil pentru ochiul uman, dar care determină rețeaua să clasifice greșit. Schlicht mai invocă exemplul din Szegedy *et al.*⁶³, care au arătat că un leu putea fi recunoscut ca o bibliotecă, iar Nguyen, Yosinski și Clune⁶⁴ au demonstrat că imagini aproape identice cu originalul pot fi interpretate ca obiecte complet diferite, cu 99% încredere.

Aceste vulnerabilități ridică probleme serioase de fiabilitate, mai ales în aplicații critice precum conducerea autonomă, unde rețelele pot fi derutate de situații neobișnuite, cum ar fi linii de sare pe drum confundate cu marcaje de bandă.

Schlicht vorbește despre această opacitate a procesului de decizie, care constituie una dintre cele mai mari limitări ale învățării profunde: deși rețelele pot atinge performanțe remarcabile, nu este clar cum ajung la concluziile lor, ceea ce face dificilă interpretarea și controlul rezultatelor. Buckner⁶⁵ subliniază că, la nivelul cel mai specific, rețelele convoluționale și cortexul perceptiv uman nu produc aceleași fenomene, ceea ce arată că aceste sisteme nu pot fi considerate echivalente cu percepția umană.

Din acest motiv, a apărut un nou program de cercetare, „IA explicabilă”⁶⁶, care își propune să facă inteligibile procesele interne ale rețelelor neuronale. În paralel, dezvoltarea rețelelor generative adversariale (GAN-uri) a fost o reacție la aceste probleme: prin competiția dintre o rețea generatoare și una discriminatorie, performanța ambelor poate fi îmbunătățită, oferind un cadru mai robust pentru învățare.

În concluzie, învățarea profundă a transformat radical domeniul inteligenței artificiale, oferind capacitatea de a procesa volume uriașe de date și de a recunoaște trăsături complexe într-un mod automat. Totuși, această putere vine cu limitări semnificative: dependența de seturi masive de date, erori diferite de cele umane, vulnerabilitatea la exemple adversariale și lipsa transparenței în luarea deciziilor. Aceste aspecte justifică dezvoltarea unor direcții de cercetare precum GAN-urile și IA explicabilă, menite să depășească aceste obstacole și să apropie performanța rețelelor de cunoașterea umană. Schlicht amintește de remarca lui Mitchell⁶⁷, care spunea că ceea ce se întâmplă în aceste rețele este „foarte diferit” de percepția

⁶² Pentru context și referințe, vezi *ibidem*, pp. 20–21.

⁶³ *Ibidem*.

⁶⁴ *Ibidem*.

⁶⁵ Cf. Cameron Buckner, „Empiricism without magic: transformational abstraction in deep convolutional neural networks”, în *Synthese SI: Neuroscience and its Philosophy*, Dordrecht, Springer, 2018, DOI: 10.1007/s11229-018-01949-1.

⁶⁶ Cf. Terrence J. Sejnowski, „The unreasonable effectiveness of deep learning in artificial intelligence”, în *Proceedings of the National Academy of Sciences*, vol. 117, nr. 48, Washington, NAS, 2020, pp. 30033–30038.

⁶⁷ Pentru context și referințe, vezi Tobias Schlicht, *op. cit.*, pp. 20–22.

umană, iar înțelegerea acestui „altceva” rămâne una dintre cele mai mari provocări ale cercetării contemporane⁶⁸.

În penultima secțiune a studiului său, Schlicht analizează rolul filosofiei lui Kant în dezbateră contemporană asupra Inteligenței Artificiale (IA), concentrându-se pe succesul și pe limitările rețelelor neuronale profunde (DNN) în contextul reîncadrării disputei dintre apriorism și empirism. Punctul de plecare al analizei este teoria cogniției la Kant, așa cum a fost stabilită anterior. O premisă esențială este aceea că înțelegerea umană se bazează pe o interacțiune echilibrată între procesarea ascendentă (*bottom-up*), prin intermediul intuiției, și cea descendentă (*top-down*), prin concepte; este vorba despre dualitatea receptivitate – spontaneitate. Dar abordarea kantiană presupune contribuții *a priori* la cunoaștere, iar acestea sunt evaluate de Schlicht drept „intrinseci”, independente și sistematic anterioare oricărei experiențe sau învățări dobândite.

Un punct important al acestei secțiuni se concentrează asupra problemei înțelegerii relațiilor cauzale; Kant a subliniat că această înțelegere nu poate fi considerată „o consecință directă a învățării bazate pe date”⁶⁹. Totuși, spre deosebire de scepticismul lui Hume, Kant a conchis că, pentru a exista o înțelegere necesară și universală a cauzalității, trebuie să existe o contribuție la nivelul intelectului care să nu fie învățată, fiind adesea identificată cu „innăscutul” sau *a priori*-ul.

În lumina realizărilor DNN-urilor, care au performanțe remarcabile în ceea ce privește percepția, clasificarea și abstracția, dezbateră filosofică s-a reconfigurat: întrebarea nu mai este dacă mintea începe ca o „tabula rasa” (fără structură), ci dacă categoriile „sunt un rezultat în principal al mecanismelor cognitive specifice unui domeniu sau generalului”⁷⁰. Mecanismele specifice unui domeniu sunt structuri specializate, dedicate unei singure sarcini cognitive, de exemplu, arată Schlicht, gramatica universală a lui Chomsky – un dispozitiv înăscut de achiziție a limbajului care stă la baza învățării tuturor limbilor naturale; mecanismele generale sunt mecanisme universale, unice de învățare, care permit dobândirea de cunoștințe în diverse domenii (de pildă, teoria behavioristă a lui Skinner criticată de Chomsky).

Critici precum Long recunosc că învățarea cere ca ceva să fie înăscut; Long, urmându-i pe Margolis și Laurence, propune ca disputa să fie încadrată strict în termenii opoziției nativism (cogniția cere mecanisme specifice domeniului – pentru un domeniu dat) – empirism (mecanismele generale sunt suficiente pentru orice domeniu)⁷¹.

Reluând contextul DNN-urilor proiectat asupra mecanismului kantian, Schlicht ridică două întrebări esențiale: (1) Având în vedere că DNN-urile nu pornesc de la zero, pot ele atinge inteligența generală doar cu mecanisme

⁶⁸ Schlicht face referire la aceste programe în anul 2021, când își scrie studiul; un aspect excepțional, care vorbește despre evoluția exponențială a tehnologiei IA, este că aceste programe deja există și au rezultate remarcabile. La acest aspect mă voi referi în finalul studiului nostru.

⁶⁹ Observația îi aparține lui Stephen A. Butterfill, în *The Developing Mind. A Philosophical Introduction*, London, Routledge, 2020, p. 93.

⁷⁰ Pentru discuție vezi T. Schlicht, *op. cit.*, p. 23.

⁷¹ *Ibidem*, p. 24.

generale?; (2) Sistemul de categorii al lui Kant este un mecanism cognitiv general sau specific unui domeniu?

Răspunsurile la întrebările de mai sus și relativ la capacitatea IA de a atinge inteligența generală (AGI) sunt împărțite de Schlicht în 2 tipuri: optimiste și pesimiste. Aceste răspunsuri sunt plasate în contextul discutării modelului cognitiv kantian, ceea ce interesează în această cercetare. Contextul teoretic mai general este cadrul descris de opoziția înnăscut/*a priori* – empiric, așa cum a fost schițat mai sus.

Prima opțiune teoretică discutată din interiorul poziției optimiștilor, reprezentată în textul lui Schlicht de Long, susține caracterul de necesitate al înnăscutului (sau al *a priori*-ului): „în mod necesar, un sistem IA la nivel uman va fi un sistem nativist”⁷². Aceasta implică faptul că inteligența generală necesită mecanisme înnăscute/*a priori* specifice domeniului. A doua opțiune, tot „optimistă”, vizează însă posibilitatea ca un sistem IA să deschidă câmpul posibilităților pe filieră empirică; afirmația specifică acestei opțiuni ne spune „că este posibil ca un sistem IA la nivel uman să fie un sistem empirist”. Aceasta susține că mecanismele generale, bazate pe învățare, deci tributare mecanismului empiric, sunt suficiente; Long spune că „IA empirist la nivel uman este cel puțin posibil”⁷³.

Optimiștii cred că dezvoltatorii pot depăși deficiențele sistemelor IA actuale, comparativ cu înțelegerea umană, fără a implementa mecanisme cognitive specifice domeniului, mai spune Schlicht.

Consider că este important să redăm fundamentul pe care se bazează argumentele în favoarea optimismului, anume capacitatea de abstractizare a DNN-urilor. Este vorba despre două tipuri de abstractizare, abstracția ierarhică și cea transformativă. Cei care vorbesc despre prima afirmă că DNN-urile sunt capabile să „învețe reprezentări ale datelor cu multiple niveluri de abstracție”⁷⁴. Acest lucru este de acceptat deoarece „abstracția, într-o anumită formă, stă la baza tuturor conceptelor noastre”. Adepții importanței acestei abstracțizări consideră că o astfel de capacitate ar deschide posibilitatea ca DNN-urile să dobândească concepte și chiar înțelegere exclusiv prin expunerea la date. În prezentarea lui Schlicht, abstracția transformativă este teoretizată Buckner, un filosof impresionat de ConvNet-uri (rețele neuronale convoluționale), pe care le discută în contextul filosofiei empiriste⁷⁵. El susține că aceste rețele „modelează un tip distinctiv de abstracție”, abstracția „din experiență”, pe care o consideră ca fiind o „componentă fundamentală a inteligenței — o formă de abstracție categorială”⁷⁶. Buckner descrie caracteristicile centrale ale DNN-urilor (multiple straturi, filtre de convoluție și *pooling*), argumentând că acestea implementează o „abstracție ierarhică” ce transformă spațiul de trăsături, păstrând elementele relevante și controlând variația de „zgomot”⁷⁷.

⁷² *Ibidem*.

⁷³ *Ibidem*.

⁷⁴ *Ibidem*, pp. 25.

⁷⁵ Cf. Cameron Buckner, „Empiricism without magic: transformational abstraction in deep convolutional neural networks”, în *Synthese SI: Neuroscience and its Philosophy*, Dordrecht, Springer, 2018, DOI: 10.1007/s11229-018-01949-1, *apud* T. Schlicht, *op. cit.*, p. 25.

⁷⁶ *Ibidem*, p. 25.

⁷⁷ *Ibidem*.

Este interesantă ceea ce Schlicht numește „Concilierea Istorică”: este vorba despre faptul că Buckner plasează această realizare în contextul istoric al relațiilor lui Locke, Berkeley și Hume despre abstracție, care au rămas misterioase în privința modului în care mintea navighează între exemplare specifice și categorii abstracte. Deși Buckner ar recunoaște la un moment dat că teoria dezvoltată „începe să semene mai mult cu teoria abstracției oferită de Kant”, el se menține pe linia empirismului. Buckner se mulțumește să fi demonstrat că DNN-urile performează abstracției care confirmă „elemente ale viziunilor lockeene, berkeleyene și kantiene”, fără a se angaja decisiv să explice diferențele dintre ele⁷⁸.

În ceea ce îi privește pe „pesimiști”, aceștia critică deopotrivă pretențiile optimiștilor și subliniază ceea ce ei numesc „limitările înțelegerii” și ale „necesității înnăscutului” în rețelele neuronale profunde. Acești „raționaliști contemporani”, cum ține să-i numească Schlicht, sunt sceptici cu privire la suficiența mecanismelor generale. În ceea ce privește chestiunea înțelegerii, Mitchell susține că problema fundamentală a DNN-urilor este tocmai lipsa de „înțelegere”⁷⁹. Schlicht explică această poziție spunând că rețelelor le lipsește cunoașterea de fundal bogată (despre funcțiile obiectelor, memorie, cogniție dependentă de context), cea care informează percepția umană. Spre deosebire de IA, „oamenii sunt înzestrați cu un corp esențial de cunoaștere de bază”⁸⁰. În acest sens, ea apelează la munca influentă a lui Spelke și Carey⁸¹, care postulează sisteme de cunoaștere de bază specifice domeniului, permițând recunoașterea obiectelor, agenților, numerelor și conceptelor precum cel de cauză. Lista trăsăturilor acestor sisteme include, de obicei, „înnăscutul”. O evaluare și mai critică este oferită de Marcus și Davis⁸², care consideră că DNN-urile oferă doar „mai mult din același lucru”, ceea ce era deja posibil. Aceștia din urmă critică faptul că cei care susțin „învățarea automată, în cea mai mare parte, accentuează învățarea, dar nu iau în considerare cunoașterea înnăscută”⁸³. Apelând și ei la Spelke, susțin că oamenii „sunt probabil născuți înțelegând că lumea constă din obiecte durabile care se deplasează pe căi conectate în spațiu și timp”⁸⁴. Aceștia introduc o referire directă la Kant, argumentând că, așa cum ar fi susținut filosoful german, un „‘manifold’ spațio-temporal înnăscut este indispensabil dacă se dorește conceperea corectă a lumii”⁸⁵. Mai mult, autorul subliniază că nu este obligatoriu ca nici chiar „înnăscutul” să fie o trăsătură necesară a sistemelor de bază identificate de Spelke și Carey. În acest sens, Butterfill arată că dovezile pentru caracterul înnăscut al acestor sisteme sunt „departe de a fi clare”⁸⁶.

⁷⁸ *Ibidem*.

⁷⁹ Cf. Melanie Mitchell, *Artificial Intelligence. A Guide for Thinking Humans*, London, Penguin, 2020, p. 132.

⁸⁰ *Ibidem*, p. 309.

⁸¹ Cf. Elizabeth Spelke, Susan Carey, „Science and core knowledge”, în *Philosophy of Science*, vol. 63, nr. 4, Chicago, University of Chicago Press, 1996, pp. 515–533, *apud* T. Schlicht, *op. cit.*, p. 26.

⁸² Cf. Gary Marcus, Ernest Davis, *Rebooting AI. Building Artificial Intelligence We Can Trust*, New York, Vintage Books, 2019, p. 145.

⁸³ *Ibidem*, p. 144; pentru context, vezi T. Schlicht, *op. cit.*, p. 26.

⁸⁴ *Ibidem*.

⁸⁵ *Ibidem*.

⁸⁶ *Ibidem*.

Tot în corul pesimiștilor se înscrie și informaticianul și filosoful Judea Pearl⁸⁷, care se concentrează pe cel mai mare obstacol specific IA-ului: „lipsa înțelegerii relațiilor cauzale de către mașini”. Deși optimist cu privire la potențialul IA-ului, el consideră mașinile de învățare actuale ca având doar „înțelepciunea unei bufnițe”; chiar și rețelele de învățare profundă recente ar oferi doar „abilități cu adevărat impresionante, dar fără inteligență”, deoarece le lipsește „un model al realității”⁸⁸. Pearl susține că un „modul de raționare cauzală” este esențial pentru ca mașinile să poată „reflecta asupra propriilor greșeli” și pentru a „funcționa ca entități morale”⁸⁹.

Pentru a ilustra ce-i lipsește IA ca să fie „înțelegătoare”, redăm pe scurt ceea ce Pearl consideră a fi o „scară a cauzalității” în trei trepte, așa cum o sintetizează Schlicht în lucrarea sa⁹⁰. Pearl schițează⁹¹ trei niveluri de abilitate cognitivă necesare pentru o adevărată înțelegere. Primul nivel este cel al asocierii/observării, nivelul cel mai de bază, împărtășit de om cu multe animale, care constă în detectarea regularităților. Ușor de anticipat, un astfel de raționament procedează doar prin asociere, permițând predicții ghidate de întrebarea „Ce se întâmplă dacă văd [...]?” Datele în sine nu dezvăluie cauza și efectul. Al doilea nivel este cel al acțiunii/intervenției, caracterizat de capacitatea unui agent cognitiv de a produce schimbări în lume, adică intervenții în ordinea cauzală fizică; exemplul dat de Pearl este contrastul dintre „a vedea fum [care] ne povestește o cu totul altă istorie despre probabilitatea unui foc decât a face fum”⁹². Întrebarea călăuzitoare la acest nivel este: „Ce se întâmplă dacă fac [...]?”. În fine, la nivelul al treilea se găsește raționamentul contrafactual, care presupune abilitatea cognitivă finală, specifică înțelegerii la nivel uman, și care permite reflecția asupra probabilităților diferitelor cauze. Întrebarea cheie aici este: „Ce s-ar fi întâmplat dacă aș fi [...]?” (de exemplu, „Ce s-ar fi întâmplat dacă nu aș fi luat aspirina?”). Acest tip de gândire transcende datele, ducând agentul cognitiv într-o lume imaginară, și reprezintă marca inteligenței umane care permite dezvoltarea științei și îmbunătățirea acțiunilor trecute.

În viziunea lui Pearl, IA-ul actual, inclusiv DNN-urile, nu a progresat dincolo de nivelul 1 (asociativ), fiind „conduse complet de un flux de date” – poziție pe care el o identifică în dreptul empirismul limitat. Chiar și programul de succes AlphaGo doar „trece prin” date acumulate din milioane de jocuri de Go, „pentru a-și da seama care mutări sunt asociate cu un procent mai mare de victorii”⁹³. Oamenii, în schimb, utilizează un „model mental al realității”⁹⁴, considerat un ingredient necesar. Pearl subliniază „cât de profund de ‘proaste’ sunt datele despre

⁸⁷ Cf. Judea Pearl, *The Book of Why. The New Science of Cause and Effect*, London, Penguin, 2018, p. 10.

⁸⁸ *Ibidem*, p. 30.

⁸⁹ Pentru context, vezi T. Schlicht, *op. cit.*, p. 27.

⁹⁰ *Ibidem*.

⁹¹ În lucrarea lui Pearl citată mai sus, referințele sunt la pp. 23–52.

⁹² *Ibidem*, p. 31.

⁹³ *Ibidem*, p. 29.

⁹⁴ *Ibidem*, p. 30.

cauze și efecte”⁹⁵, argumentând că bazarea exclusivă pe date este insuficientă pentru atingerea inteligenței generale.

Concluzia lui Schlicht discută poziția kantiană în contextul posibilităților oferite de calea empirică a unei virtuale IA cognitive autentice. Ea se rezumă la observația interesantă concentrată în ideea că întrebarea dacă optimiștii sau pesimiștii se vor dovedi corecți este, în cele din urmă, o chestiune empirică ce nu poate fi soluționată teoretic; Schlicht face trimitere la cunoscutele limitări ale IA-ului în forma sondajului prezentat de Cremer⁹⁶ cu experți în IA care listează patruzeci de limitări ale învățării profunde. În esență, succesul depinde de scop: se dorește cogniție și inteligență asemănătoare celor umane sau doar succes în suficiente sarcini, indiferent de modalitatea obținerii acestuia.

Tobias Schlicht adoptă o poziție echilibrată, critic-constructivă în evaluarea încercărilor recente de a interpreta sau a implementa cogniția kantiană în cadrul inteligenței artificiale. În primul rând, Schlicht respinge simplificările din unele reconstrucții ale programului lui Kant: el critică atât tendința de a-l plasa pe Kant direct în tabăra funcționalismului (așa cum face Brook), cât și asimilarea necritică la modelele de procesare predictivă. Aceste lecturi ignoră tensiuni fundamentale legate de materialism, statutul subiectului și rolul conștiinței, care fac ca astfel de identificări să fie, cel mult, parțiale și problematice. În al doilea rând, Schlicht acordă un rol central conștiinței și apercepției. Sprijinindu-se pe interpretările lui Tolley și Watkins & Willaschek, el susține că, la Kant, „cogniția în sens restrâns” presupune în mod necesar conștiința de sine transcendentală. Acesta nu este un simplu epifenomen, ci o condiție constitutivă a judecății și a unității experienței. Din acest motiv, Schlicht subliniază că sistemele de IA actuale, inclusiv cele inspirate de Kant, nu satisfac această condiție, întrucât le lipsește apercepția propriu-zisă.

În finalul lucrării sale, Schlicht introduce o cercetare foarte recentă în învățarea automată, la care m-am referit la începutul studiului de față. Acest vast proiect are ca punct de plecare perspectiva lui Kant cu privire la cogniție: este vorba despre poziția dezvoltată de Richard Evans și colaboratorii săi de la DeepMind⁹⁷, care propune o arhitectură cognitivă bazată pe fundamentele kantiene. Schlicht vorbește pozitiv despre Richard Evans și colegii săi, care s-au inspirat din filosofia lui Kant atunci când au propus o abordare computațională a „procesului de a atribui sens unei secvențe senzoriale”. Ei susțin că, pentru a interpreta o astfel de secvență, este necesar să o reconstruim ca reprezentare a unei lumi externe compuse din obiecte persistente, ale căror atribute evoluează conform unor legi generale. Această reconstrucție implică construirea unei teorii cauzale simbolice care nu doar explică datele senzoriale, ci îndeplinește și anumite condiții de unitate: obiectele, proprietățile și atomii teoriei trebuie integrate într-un întreg coerent, condiție esențială pentru a obține predicții, retro-dicții și imputări precise. Schlicht

⁹⁵ *Ibidem*, p. 16.

⁹⁶ Pentru context, vezi T. Schlicht, *op. cit.*, p. 28.

⁹⁷ Richard Evans, „The Apperception Engine”, în *Kant and Artificial Intelligence*, De Gruyter, pp. 39–103.

consideră proiectul lui Evans și al colaboratorilor săi drept un experiment empiric extrem de interesant, capabil să testeze și să clarifice anumite intuiții kantiene despre sinteză, reguli și unitate. Totuși, rămâne rezervat față de ideea că o arhitectură lipsită de autoconștiință autentică ar putea reproduce integral cogniția umană, așa cum o înțelege Kant.

În ansamblu, poziția lui Schlicht este una de mediere: el recunoaște valoarea euristică a modelelor computaționale inspirate din Kant, dar insistă că acestea nu trebuie confundate cu o implementare completă a cogniției kantiene, *atâta timp cât problema conștiinței rămâne nerezolvată*.

Spre deosebire de Schlicht, care face un foarte scurt rezumat al acestei cercetări importante (a lui Evans *et. al.*), eu voi prezenta mai jos pe larg proiectul lui Evans.

DE LA APREHENSIUNEA KANTIANĂ LA INTELIGENȚA ARTIFICIALĂ: *THE APPERCEPTION ENGINE*

Dintru început este de precizat unde se plasează Evans atunci când se referă la proiectul primei *Critici* a lui Kant; în cadrul discuției generate de tensiunea dintre perspectivele asupra proiectului kantian, el oscilează între o psihologie empirică și una *a priori* kantiană. O primă consecință este că, dacă lectorul nu este de acord cu interpretarea primei *Critici* kantiene ca psihologie *a priori*, atunci paternitatea filosofului german în dezvoltarea proiectului lui Evans nu mai poate fi ușor invocată; în plus, apelul la acuratețea susținerilor kantiene atunci când sunt utilizate argumentele lui Kant poate fi considerat, într-o oarecare măsură, mai puțin justificat. Aceste limite, de altfel, sunt asumate explicit chiar de autor pe parcursul lucrării. Trebuie precizat că Evans nu optează radical pentru una dintre perspective, ci, cum vom vedea, balează între acestea; el consideră propriul său program ca „o specificație a unei arhitecturi cognitive kantiene” – dezideratul lucrării lui.

Chiar după spusele sale, lucrarea lui Evans descrie o încercare de a reutiliza psihologia *a priori* a lui Kant ca schiță arhitecturală pentru un sistem de învățare automată⁹⁸. În primul rând, această cercetare descrie condițiile care trebuie îndeplinite pentru ca agentul să atingă unitatea experienței: intuițiile trebuie să fie conectate prin relații binare astfel încât să satisfacă diverse condiții de unitate. În al doilea rând, spune Evans, cercetarea sa arată cum sunt derivate categoriile în cadrul acestui model: categoriile sunt predicate unare pure, care sunt derivate din relațiile binare pure. În al treilea rând, autorii încearcă să arate și să descrie cum arhitectura cognitivă a lui Kant a fost implementată într-un sistem computerizat („Motorul Apercepției”); mai mult, este arătat în detaliu cum sistemul construiește o experiență unificată dintr-o secvență de date senzoriale brute de intrare.

În debutul introducerii, Evans ne invită să ne imaginăm o mașină echipată cu senzori care primește un flux de informații senzoriale. Ea trebuie, cumva, să dea sens acestui flux de date senzoriale. Dar, ne/se întrebă autorul, ce anume implică,

⁹⁸ *Ibidem*, p. 39.

exact, acest lucru? Noi avem o înțelegere intuitivă a ceea ce înseamnă „a da sens” datelor senzoriale – dar putem specifica cu precizie ce presupune acest lucru? Mai mult, poate fi formalizată această noțiune intuitivă? În învățarea automată, aceasta se numește „problema învățării nesupravegheate”, spune Evans. Într-adevăr, pe cât de importantă este ca această problemă, pe atât de slab definită este; ea contrastează cu problema învățării supravegheate, unde datele senzoriale vin „atașate cu etichete”, adică sunt pre-destinate și transparente. Într-o problemă de învățare supravegheată, există un obiectiv de învățare clar, precum și o serie de tehnici puternice care funcționează cu mult succes. Totuși, spune Evans, lumea reală nu vine cu etichete atașate datelor senzoriale. Noi doar primim aceste date⁹⁹. Precizarea stării viitoare a fotoreceptorilor cuiva poate fi parte din ceea ce implică a da sens – dar nu este suficientă prin ea însăși. Ce înseamnă, așadar, a da sens unei secvențe senzoriale?

În lucrarea sa, Evans argumentează că „soluția la această problemă se ascunde, la vedere, de peste două sute de ani, în *Critica Rațiunii Pure* a lui Kant”¹⁰⁰. În prima ediție a primei *Critici* (1781), filosoful german ar defini exact ceea ce înseamnă „a da sens unei secvențe”: „a reinterpreta acea secvență ca o reprezentare a unei lumi externe compuse din obiecte, care persistă în timp, cu atribute care se schimbă în timp, conform unor legi generale”¹⁰¹. Evans își propune să ofere astfel o descriere precisă și implementabilă computațional a ceea ce înseamnă a da sens fluxului senzorial. El susține că un astfel de proiect interdisciplinar – la intersecția dintre filosofia kantiană și inteligența artificială – prezintă două riscuri principale: poate să nu fie suficient de fidel intențiilor originare ale lui Kant și poate să nu ofere contribuții semnificative pentru cercetarea în IA. Cu toate acestea, Evans argumentează că IA-ul contemporan are de învățat de la Kant, deoarece filosofia kantiană oferă o viziune hibridă care combină punctele forte ale empirismului (reprezentat astăzi de rețelele neuronale) și ale raționalismului (reprezentat de învățarea bazată pe logică). În același timp, el arată că studiul lui Kant poate beneficia de pe urma formalizării computaționale, deoarece aceasta impune claritate și precizie conceptuală.

Pentru a demonstra această idee, prin „Motorul Apercepției”, Evans propune o implementare concretă a arhitecturii cognitive kantiene, capabilă să genereze teorii cauzale interpretabile dintr-un volum redus de date, datorită constrângerilor de unitate impuse de Kant. Lucrarea extrage tezele centrale din prima jumătate a *Criticii*, le asamblează într-o specificație formală așa încât să ilustreze funcționarea motorului în exemple detaliate; în final, Evans discută alegerile interpretative făcute de mașină, subliniind modul în care implementarea computațională forțează claritatea și elimină ambiguitățile. În esență, Evans vrea să arate că o arhitectură inspirată de Kant poate îmbina avantajele învățării statistice și ale raționamentului logic, oferind atât progres teoretic în filosofie, cât și soluții practice pentru problemele actuale ale IA.

⁹⁹ *Ibidem*.

¹⁰⁰ *Ibidem*.

¹⁰¹ *Ibidem*, p. 40.

Încercând să-l urmărească pe Kant mai îndeaproape, Evans face referire la A 110 (din prima ediție a *Criticii rațiunii pure*) și susține, alături de filosoful german, că „pentru a atinge experiența, trebuie să-mi unific intuițiile”¹⁰². Analizându-le pe rând, el își proiectează structura lucrării încercând să răspundă pe bucăți întrebărilor: (i) Ce înțelege Kant prin *experiență*?, (ii) Ce sunt *intuițiile*?, (iii) Ce înseamnă să le *unifici*?¹⁰³.

În ceea ce privește experiența, Evans subliniază necesitatea ca ea să fie *unificată*, adică eu am „o singură experiență oricând”; este însă vorba despre o experiență *articulată*, nu este una doar conceptuală, ci o combinație unificată – nu neapărat veridică¹⁰⁴. Experiența nu este un dat, ea se construiește, se „realizează”. Psihologia *a priori*, cea care ar sta la baza proiectului kantian al primei *Criticii* descrie în detaliu procesele subiacente necesare ca experiența să fie atinsă.

Distincte de concepte, intuițiile sunt interpretate de Evans ca reprezentări ale obiectelor particulare și/sau ca aparținând unor atribute particulare. Intuițiile trebuie să capete sens, și, de asemenea, să fie unificate. În psihologia *a priori* despre care vorbește Evans, intuițiile sunt importante, sunt scopul final al întregii gândiri; celelalte elemente ajută la unificarea intuițiilor. Mai simplu spus, pentru a atinge experiența, trebuie să-mi unific intuițiile. În acest demers, esențial este procesul de sinteză, care presupune constrângerea unității; am avea diversitate, sinteză a diversității, conexiune și conceptul de unitate a diversității prin conexiune.

Această constrângere a unității despre care vorbește Evans ar fi chiar procesul de sinteză kantian. Dar unificarea se realizează *prin relația binară*: „Sinteza intuițiilor înseamnă conectarea intuițiilor între ele folosind relații binare, astfel încât graful rezultat, nedirecționat, să fie complet conectat.”¹⁰⁵

Procesul de sinteză este sarcina imaginației productive [A78/B103; A188/B230], pe care Evans o descrie în Secțiunea 2.2 și o formalizează în Secțiunea 2.4. El susține că graful trebuie să satisfacă condițiile de unitate indexate de Kant și detaliate de Evans în Secțiunile 2.3.1, 2.3.2, 2.3.3 și 2.3.4 ale cercetării sale. Aceste condiții sunt satisfăcute de facultatea intelectului.

Cum am mai amintit, toate cele de mai sus se subsumează interpretării de către Evans a primei jumătăți a *Criticii* ca psihologie *a priori*. Contrar lui Strawson¹⁰⁶, Evans susține că psihologia *a priori* este văzută ca o formă de investigație legitimă și importantă, și, mai mult, dacă încercăm să o eliminăm din textul lui Kant, nu mai rămâne mult care să fie inteligibil.

Una dintre problemele pe care le regăsim și la Evans, care a fost abordată și de către Schlicht mai sus, este cea exprimată de întrebarea dacă intuițiile sunt relații între minți conștiente și obiecte materiale care există efectiv, sau dacă obiectul unei

¹⁰² Pentru context, vezi *ibidem*, p. 45.

¹⁰³ *Ibidem*.

¹⁰⁴ Béatrice Longuenesse, *Kant and the Capacity to Judge*, Princeton, NJ, Princeton University Press, 1998.

¹⁰⁵ Richard Evans, „The Apperception Engine”, în T. Schlicht, *op. cit.*, p. 48.

¹⁰⁶ Peter Strawson, *The Bounds of Sense*, London, Routledge, 2018.

intuiții este doar o reprezentare mentală care nu implică în niciun fel existența unui obiect fizic extern corespunzător¹⁰⁷. Interpretarea asumată explicit în lucrare este circumscribibilă în mod clar perspectivei celei din urmă, celei reprezentationale. Motivul invocat în favoarea interpretării reprezentationale se bazează pe o prejudecată interpretativă generală: „ori de câte ori există două moduri de a-l citi pe Kant, și una dintre acele interpretări se bazează pe mai puține capacități prealabile, cerând astfel minții să depună mai mult efort pentru a realiza reprezentarea coerentă a unei lumi externe pe care o luăm de la sine înțeleasă în viața noastră de zi cu zi, atunci preferați acea interpretare”¹⁰⁸.

Viziunea relațională ia de la sine înțeles un anumit tip de realizare cognitivă: capacitatea minții de a fi despre un obiect extern. Viziunea reprezentatională, dimpotrivă, vede această intenționalitate, această orientare a minții, ca pe ceva care necesită efort pentru a fi realizat. Astfel, pur și simplu pentru că este mai solicitantă și pune întrebări mai dificile, ar trebui preferată, susține Evans; în plus, și nu întâmplător, *viziunea reprezentatională poate fi implementată într-un program de calculator*, în timp ce este complet neclar cum am putea începe să implementăm vreo viziune relațională care ia de la sine înțeleasă capacitatea gândurilor minții de a fi îndreptate către obiecte fizice externe particulare. Proiectul lui Evans se concentrează pe filosofia teoretică a lui Kant, iar „gândirea”, aici, înseamnă gândirea cognitivă menită să dea sens lumii.

În efortul de a-și susține convingător proiectul, Evans încearcă să prezinte cât mai clar distincțiile conceptuale cu care lucrează, preluate din *Critică*, firește; astfel, folosindu-l pe Sellars¹⁰⁹, el distinge între determinare și judecată: prima nu face decât să descrie pur și simplu, o determinare este descriptibilă, poate fi/este o percepție, a doua face legături între concepte sau între atribute și concepte. Important aici este de subliniat că, dacă determinarea este un mod de a percepe, ea nu are valoare de adevăr: *adevărul și iluzia nu se găsesc în obiect*, ci în judecata despre el, în măsura în care acest obiect este gândit. Astfel, arată Evans, se spune corect că simțurile nu greșesc; dar nu pentru că ele judecă întotdeauna corect, ci pentru că ele nu judecă deloc. De aici adevărul, la fel ca și eroarea, și, deci, ca și iluzia care duce la aceasta din urmă, se găsesc doar în judecăți, adică doar în relația obiectului cu intelectul nostru. În simțuri nu există nicio judecată, nici adevărată, nici falsă. [KrV A293-4/B350, *Logica Jäsche* 9:53].

Există trei operații care leagă intuițiile între ele: cuprinderea, când un obiect este (în prezent) cuprins în alt obiect, comparația, când un atribut este (în prezent) mai mic decât un alt atribut, inerența când un atribut (în prezent) este inerent într-un obiect – de exemplu, această greutate particulară este un atribut al acestui colet particular¹¹⁰.

¹⁰⁷ Pentru context și referințe, vezi Richard Evans, „The Apperception Engine”, în T. Schlicht, *op. cit.*, p. 46 și n. 16.

¹⁰⁸ *Ibidem*.

¹⁰⁹ Wilfrid Sellars, „The role of imagination in Kant’s theory of experience”, în *In the Space of Reasons*, Cambridge, MA, Harvard University Press, 1978, pp. 454–466.

¹¹⁰ Richard Evans, *op. cit.*, p. 50.

Pe lângă cele trei operații pure care leagă intuițiile între ele, există trei²⁶ relații pure care leagă determinările între ele: succesiunea (două evenimente sunt în timpuri diferite), simultaneitatea (două evenimente sunt în același timp), incompatibilitate (două evenimente nu pot fi în același timp)¹¹¹. Definiția conexiunii ne spune că, atunci când două determinări sunt legate între ele printr-una dintre cele trei relații, rezultatul este o conexiune.

Alegerea acestor trei relații pure, precum și a celor trei operații pure este motivată de faptul că toate constituie împreună un set minimal de operatori binari care, împreună, sunt suficienți pentru a construi formele spațiului și timpului.

Ca sarcină a imaginației productive, funcția sintezei este de a conecta intuițiile între ele, folosind operațiile și relațiile pure descrise mai sus, astfel încât să construiască forma spațio-temporală obiectivă, spune Evans. Operația de cuprindere ne permite să combinăm intuițiile într-un câmp spațial (o reprezentare minimală a spațiului care face abstracție de numărul de dimensiuni¹¹²). Operația de comparație ne permite să comparăm două atribute diferite; dacă generăm un atribut intermediar între două atribute comparabile, putem genera un moment intermediar în timp între două momente observate [A165/B 208ff], umplând astfel timpul [A145/B184]. Operația de inerență ne permite să atribuim atribute diferite unui obiect în momente diferite. Relațiile de simultaneitate și succesiune ne permit să ordonăm determinările în timp. În cele din urmă, relația de incompatibilitate ne permite să testăm când seturile de determinări sunt composable¹¹³.

Acum se vede din toate acestea că schema fiecărei categorii face reprezentabil: în cazul mărimii, generarea (sinteza) timpului însuși, în aprehensiunea succesivă a unui obiect; în cazul schemei calității, sinteza senzației (percepției) cu reprezentarea timpului, sau umplerea timpului; în cazul schemei relației, relația percepțiilor între ele la tot timpul (adică, în conformitate cu o regulă de determinare a timpului); în sfârșit, în schema modalității și a categoriilor sale, timpul însuși este un corelat al determinării, dacă și cum un obiect aparține timpului. Prin urmare, schemele nu sunt nimic altceva decât determinări *a priori* ale timpului conform unor reguli, iar acestea privesc, conform ordinii categoriilor, seria de timp, conținutul timpului, ordinea timpului și, în sfârșit, suma totală a timpului în raport cu toate obiectele posibile. [A145/B184ff]

Evans subliniază afirmația cheie compusă din două părți: „sinteza implică (i) conectarea intuițiilor între ele prin operațiile de cuprindere, comparație și inerență pentru a forma determinări și (ii) conectarea determinărilor între ele prin relațiile de succesiune, simultaneitate și incompatibilitate”¹¹⁴.

După descrierea felului cum sunt conectate intuițiile între ele folosind diferitele relații binare pure, în ceea ce privește „unitatea sintetică” propriu-zisă, Evans accentuează că aceasta înseamnă mai mult decât simpla conexiune a intuițiilor. El vorbește despre patru tipuri de condiții de unitate pe care le-ar impune

¹¹¹ *Ibidem*, p. 52.

¹¹² Cf. Wayne Waxman, *Kant's Anatomy of the Intelligent Mind*, Oxford, Oxford University Press, 2014.

¹¹³ Richard Evans, *op. cit.*, pp. 52–53.

¹¹⁴ *Ibidem*, p. 54.

Kant: (i) condițiile de unitate pentru sinteza relațiilor matematice, (ii) condițiile de unitate pentru sinteza relațiilor dinamice, (iii) cerința ca judecățile să fie garantate de determinări și (iv) condiția de unitate conceptuală.

Firește, nu este nimerit să redăm întreaga descriere în forma și succesiunea preferate de Evans, însă a fost necesar să urmărim acești pași (de mai sus) pentru a putea înțelege modul de construcție al „motorului apercepției”, care încearcă să urmărească modul kantian de construcție a configurației apercepției transcendente. Mai insistăm asupra procesului de unificare din arhitectura propusă de Evans: spațiul unificator abstract este de fapt o ierarhie de cuprindere, „spațiul este reprezentarea coexistenței (*juxtapunere*) [A374]”¹¹⁵, iar condițiile de unitate pentru sinteza relațiilor matematice sunt cuprinderea și comparația; condițiile de unitate pentru sinteza relațiilor dinamice sunt inerența, succesiunea, simultaneitatea și incompatibilitatea.

Afirmația fundamentală a lui Kant este că doar judecata poate fixa poziționarea intuițiilor; mai mult, acesta nu este doar un rol al judecății printre multe altele – acesta este rolul principal al judecății: o judecată nu este altceva decât modul de a aduce cunoștințele date la unitatea obiectivă a apercepției [B141], spune Kant. Tradus în limbajul proiectului lui Evans, acest lucru înseamnă că pozițiile relative ale intuițiilor într-o determinare pot fi fixate doar prin formarea unei judecăți, care la rândul-i necesită această poziționare particulară. Această judecată conține concepte sub care se încadrează intuițiile, iar poziția intuițiilor în determinare este determinată indirect de pozițiile conceptelor corespunzătoare în judecată.

Evans face referire apoi la ceea ce în literatura kantiană se numește „argumentul identității funcției”, pe care el nu îl citează astfel (nici nu este necesar aici); în esență, acesta spune că „aceeași funcție care dă unitate diferitelor reprezentări într-o judecată dă, de asemenea, unitate simplei sinteze a diferitelor reprezentări într-o intuiție. Același intelect, așadar, și anume prin aceleași acțiuni prin care aduce forma logică a unei judecăți în concepte prin unitatea analitică, aduce, de asemenea, un conținut transcendental în reprezentările sale prin unitatea sintetică a diversității în intuiția în general”. [A79/B104-5]. În continuare, autorul nu dezvoltă acest moment din Kant, ci îl folosește pentru a apela la o afirmație paralelă, tot a lui Kant, dar la un nivel „superior”, la nivelul judecăților complexe: pozițiile relative ale determinărilor într-o conexiune pot fi fixate doar prin formarea unei judecăți complexe care conține ea însăși o pereche de judecăți ca constituenți care necesită această poziționare particulară. Această judecată complexă conține doi constituenți – judecăți – sub care se încadrează cele două determinări, iar poziția determinărilor în conexiune este determinată indirect de pozițiile judecăților corespunzătoare în judecata complexă¹¹⁶.

Importanța acestei mișcări este dată de importanța scopului atribuit de Kant relațiilor dinamice: acela de a ordona intuițiile și determinările în spațiu-timpul obiectiv. Or, putem atinge obiectivitatea doar prin impunerea necesității asupra combinării; dar facultatea imaginației este în întregime incapabilă să impună necesitate – tot ce poate face imaginația este să conecteze intuițiile folosind relațiile

¹¹⁵ *Ibidem*, p. 57.

¹¹⁶ *Ibidem*, p. 59.

pure, ea nu poate impune necesitate acelor conexiuni. De fapt, singurul element care poate oferi necesitatea dorită este judecata. Astfel, singura modalitate prin care relațiile dinamice pot fi ordonate în spațiu-timp obiectiv este prin poziționarea lor indirectă, folosind judecăți care impun necesitatea pe care o cer conexiunile¹¹⁷.

În ceea ce privește facultățile cognitive responsabile pentru diferitele procese, capacitatea de a judeca este responsabilă pentru construirea judecăților, iar puterea de judecată este responsabilă pentru construirea elementelor care decid ce intuiții se încadrează sub ce concepte. Aceasta este, așadar, afirmația generală, așa cum se aplică tuturor relațiilor dinamice¹¹⁸.

În continuare, Evans va descrie diferitele forme de judecată care sunt necesare pentru a garanta diferitele relații dinamice: inerență, succesiune, simultaneitate și incompatibilitate; nu insistăm aici asupra lor. Doar spunem că inerența trebuie susținută de o judecată categorică (afirmația lui Kant aici este că numai pentru că formez o judecată categorică corespunzătoare sunt capabil să fixez pozițiile celor două argumente ale operatorului de inerență). Succesiunea trebuie susținută de o judecată causală: a doua condiție de unitate dinamică este că fiecare succesiune de determinări trebuie să fie susținută de o judecată causală (Kant afirmă că numai pentru că formez o judecată causală corespunzătoare sunt capabil să fixez pozițiile celor două determinări în relația de succesiune [A189/B232]; simultaneitatea trebuie susținută de o pereche de judecăți cauzale). A treia condiție de unitate dinamică este că fiecare simultaneitate de determinări trebuie să fie susținută de o pereche de judecăți cauzale. Adăugăm aici că orice incompatibilitate între determinări trebuie să fie întotdeauna susținută de o judecată disjunctivă¹¹⁹.

În ceea ce privește relația dintre unitate, determinări și judecăți, pe lângă condiția de unitate care cere ca determinările să fie garantate de judecăți, există și condiții de unitate în cealaltă direcție, care cer ca judecățile să fie susținute de determinări corespunzătoare, susține Evans¹²⁰. Parafrazându-l pe Kant, el continuă arătând că este la fel de necesar să facem sensibile conceptele pentru minte (adică, să le adăugăm un obiect în intuiție), pe cât este de necesar să facem intuițiile sale inteligibile (adică, să le aducem sub concepte). [A51/B75]. Cerința aici ar fi ca judecățile să nu poată „pluti în derivă” în raport cu intuițiile subiacente. În schimb, fiecare judecată trebuie să fie susținută de o determinare corespunzătoare.

Înainte de a prezenta rezumativ cum este replicat modelul cognitiv kantian în „motorul apercepției” al lui Evans, mai semnalez o situație importantă. Este vorba despre situația în care formarea unei judecăți se realizează atribuind un concept unui obiect particular: atunci trebuie să existe o determinare de inerență corespunzătoare care să atribuie un atribut particular unui obiect anume particular, unde obiectul particular se încadrează sub un concept corespunzător. Dar această condiție nu este satisfăcută în mod trivial, spune Evans, așa cum s-ar putea crede la prima vedere, având în vedere că agentul începe cu intuiții și determinări și

¹¹⁷ *Ibidem*.

¹¹⁸ Pentru context și referințe, vezi Richard Evans, „The Apperception Engine”, în T. Schlicht, *op. cit.*, p. 60.

¹¹⁹ *Ibidem*, p. 63.

¹²⁰ *Ibidem*, p. 64.

formează judecăți pentru a le face inteligibile; căci, uneori, agentul construiește noi obiecte inventate pentru a da sens datului sensibil și atribuie proprietăți acestor obiecte inventate. În astfel de cazuri, condiția de mai sus cere ca, la fel ca în cazul unui obiect „real”, pe lângă subsumarea obiectului sub conceptul corespunzător, trebuie să existe și un atribut individual particular corespunzător obiectului, care să-i fie inerent.

Condiția de unitate pentru unitatea conceptuală este dată de cerința ca fiecare concept să figureze într-o judecată disjunctivă. Evans va formaliza în secțiunile următoare sarcina de a atinge unitatea sintetică a apercepției incluzând această condiție; formalismul astfel introdus este considerat necesar pentru derivarea categoriilor.

În atingerea unității sintetice, Evans apelează la o figură¹²¹ în care sunt prezentate două modalități de a grupa patru¹²² facultăți pe care le indexează la Kant pe baza a două distincții transversale. Pe de-o parte, senzația și imaginația sunt încadrate sub sensibilitate, deoarece ambele operează cu intuiții. Judecata și capacitatea de a judeca se încadrează sub intelect (sau înțelegere), întrucât ambele manipulează concepte; pe de altă parte, senzația este plasată sub receptivitate, fiind o facultate pur pasivă ce doar primește datele. Celelalte trei facultăți – imaginație, judecată și capacitatea de a judeca – se încadrează sub spontaneitate, deoarece agentul are libertatea de a construi orice, atâta timp cât rezultatul respectă condițiile de unitate.

Evans își prezintă propria lectură a lui Kant în derivarea categoriilor: în introducerea *Schematismului*, filosoful german arată că, pentru a subsuma un obiect unui concept, reprezentările obiectului trebuie să fie omogene cu conceptul, adică conceptul trebuie să conțină ceea ce este reprezentat în obiect. Conceptele pure (de exemplu, cauzalitatea) nu sunt omogene cu intuițiile empirice; ele nu apar niciodată în experiență, ca atare. Cum pot atunci să fie aplicate la fenomene? Pentru conceptele empirice, subsumarea se explică prin intermediul unui atribut particular al obiectului care se încadrează în concept. Conceptele pure (Unitate, Realitate, Substanță etc.) nu au niciun atribut intuitiv corespunzător; deci explicația de mai sus nu funcționează aici. După Evans, Kant ar răspunde că subsumarea unui obiect sub un concept pur este justificată de existența unei relații pure care leagă obiectul de concept. Astfel, obiectul este subsumat deoarece (i) este legat de relația pură și (ii) conceptul pur poate fi derivat din acea relație. Această relație pură, susține Evans, este ceea ce Kant numește schema transcendențială: un intermediar pur, intelectual și sensibil, care face posibilă aplicarea categoriei la fenomen.

Prin urmare, conchide Evans, schema transcendențială este doar un alt termen pentru relația pură care mediază între categorie și fenomen.

¹²¹ *Ibidem*, p. 71.

¹²² Această reconfigurare nu este problematică în context, Evans asumându-și că nu va urma „litera” *Criticii* în eaborarea „motorului apercepției” (vezi partea de final despre limitele „motorului apercepției” – pp. 94–99); a asuma că cele patru facultăți sunt senzația, imaginația, puterea de judecare și capacitatea de a judeca (vezi pp. 70–71) nu este tocmai „ortodox”.

Revenind la argumentul general pentru derivarea categoriilor, argumentul lui Kant este rezumat astfel: atingerea experienței cere să conectez intuițiile folosind relațiile pure. Dacă conectez intuițiile folosind relațiile pure, atunci îmi este permis să aplic conceptele pure (categoriile) la obiectele de intuiție. Astfel, atingerea experienței îmi permite să aplic conceptele pure la obiectele din intuiție. Dar posibilitatea de a aplica conceptele pure la obiectele de intuiție este condiționată de activitatea mea, activitatea de a încerca să ating experiența. De aici, concluzia lui Kant ar fi că acestor categorii le este permis să se aplice doar obiectelor experienței.

Kant subliniază că nu există concepte pure înnăscute; conceptele unice pure nu sunt „incorporate” ca predicate unare primitive în limbajul gândirii. Singurele elemente înnăscute, după Evans, sunt facultățile fundamentale – sensibilitatea, imaginația, puterea de judecată și capacitatea de a judeca¹²³ – împreună cu relațiile pure. Categoriile însele sunt dobândite, consideră Evans, fiind derivate din relațiile pure *in concreto* atunci când conferim sens unei anumite secvențe senzoriale. Totuși, ele sunt dobândite originar [Entdeckung, Ak. VIII, 222-23; 136], deoarece pot fi întotdeauna derivate din orice secvență senzorială. Prin urmare, conceptele pure nu sunt înnăscute, ci dobândite originar.

Structura cognitivă kantiană descrisă pe scurt mai sus a fost implementată în ceea ce Evans numește „motorul apercepției”, iar sistemul informatic este descris în două lucrări apărute în 2021¹²⁴. În continuare, vom sintetiza și rezuma experimentul propus de Evans, precum și ultima parte a studiului său.

Experimentul descris pornește de la un input senzorial¹²⁵ extrem de minimalist, constituit din citiri provenite de la doi senzori de lumină care înregistrează variații de intensitate în timp. Deși la prima vedere datele par triviale, autorul insistă asupra faptului că tocmai simplitatea lor face transparentă problema filosofică vizată. Citirile nu sunt oferite ca evenimente deja situate într-un timp obiectiv, ci ca o succesiune brută de determinări accesate pe rând de agent, adică într-un timp pur subiectiv, dependent de focalizarea atenției agentului asupra unuia sau altuia dintre senzori. În plus, datele nu sunt „curate”, ci sunt intenționat degradate prin introducerea de „zgomot”, pentru a elimina posibilitatea ca structura obiectivă să fie citită direct din input.

Această decizie metodologică este esențială, deoarece forțează sistemul să construiască activ unitatea experienței, în loc să o presupună ca dată. Dintr-o perspectivă kantiană, acest lucru reflectă ideea că timpul obiectiv, ca formă a experienței ordonate, nu este un simplu cadru în care evenimentele apar, ci un rezultat al sintezei continue realizate de facultățile cognitive.

Pentru a interpreta acest flux senzorial, agentul trebuie să realizeze o structurare complexă, care nu se reduce la clasificarea datelor, ci implică trei niveluri distincte, dar interdependente. Mai întâi, este necesară o sinteză a intuițiilor prin care determinările individuale sunt legate prin relații formale de

¹²³ Aceste patru așa-zise facultăți kantiene sunt altele decât cele enunțate anterior – vezi nota 122.

¹²⁴ Richard Evans, „The Apperception Engine”, în T. Schlicht, *op. cit.*, p. 75.

¹²⁵ *Ibidem*, p. 75.

simultaneitate, succesiune și incompatibilitate. Această sinteză este cea care transformă ordinea pur subiectivă a citirilor într-o ordine în timp obiectiv, în care mai multe determinări pot fi considerate simultane sau succesive în mod necesar. Alegerea acestor relații nu este trivială, deoarece pentru fiecare pereche de determinări consecutive există posibilitatea de a fi grupate împreună sau separate temporal, ceea ce duce la un spațiu combinatoric uriaș al interpretărilor posibile. Prin urmare, timpul obiectiv nu este „reconstruit” mecanic, ci este ipotetic, exploratoriu și dependent de criteriile de coerență globală.

Al doilea nivel este cel al subsumării intuițiilor sub concepte, corespunzător în termeni kantieni funcției judecării. Fiecare determinare senzorială este reprezentată ca un vector binar, iar atribuirea de concepte acestor vectori este realizată printr-o rețea neuronală binară care funcționează ca un clasificator *multilabel*¹²⁶. Important este că această subsumare nu este una rigidă sau exclusivă, un vector putând fi subsumat simultan mai multor predicate, ceea ce introduce ambiguitate conceptuală. Această ambiguitate nu este tratată ca un defect, ci ca o parte constitutivă a procesului cognitiv, care va trebui rezolvată ulterior prin integrarea într-o teorie mai amplă. Astfel, conceptele nu sunt simple etichete perceptuale, ci noduri într-o rețea explicativă care își dobândește sensul deplin doar în contextul judecăților.

Al treilea nivel este reprezentat de sinteza judecăților propriu-zise, realizată printr-un motor de sinteză de programe care caută să construiască o teorie simbolică minimală și coerentă capabilă să explice dinamica fenomenelor observate. Această teorie include o semnătură de obiecte și predicate, un set de condiții inițiale, un set de reguli cauzale sau condiționale și un set de constrângeri logice. Judecățile astfel generate nu sunt simple regularități descriptive, ci instanțieri ale necesității, în sensul în care ele justifică de ce anumite succesiuni sunt considerate obiective și nu simple coincidențe subiective. De exemplu, pentru ca o tranziție între două stări să fie interpretată drept o succesiune temporală reală, ea trebuie să fie susținută de o regulă de tip cauzal care o face necesară. În acest fel, sistemul operaționalizează ideea kantiană conform căreia experiența obiectivă presupune supunerea fenomenelor sub legi.

Un aspect deosebit de semnificativ al modelului este faptul că agentul nu este limitat la datele strict percepute, ci poate introduce entități și determinări suplimentare prin ceea ce autorul identifică drept o funcție analogă imaginației productive. Așa cum am mai arătat, pentru ca judecățile să fie garantate de intuiții, este necesar ca fiecare concept aplicat să fie susținut de o determinare individuală corespunzătoare¹²⁷. Atunci când teoria introdusă implică obiecte sau regiuni spațiale care nu sunt date direct în percepție, sistemul este constrâns să inventeze determinări particulare asociate acestora. Această completare imaginară nu este arbitrară, ci este strict ghidată de cerința coerenței globale și de regulile teoriei, reflectând ideea kantiană că imaginația nu creează liber, ci operează sub constrângerea intelectului.

¹²⁶ *Ibidem*, p. 79.

¹²⁷ *Ibidem*, pp. 81–82.

Dintre numeroasele interpretări posibile ale fluxului senzorial, sistemul trebuie să selecteze una singură ca fiind preferabilă. Pentru aceasta sunt introduse două criterii metodologice fundamentale: primul este simplitatea teoriei, în sensul unei descrieri cât mai concise a regulilor și constrângerilor necesare pentru a explica datele; al doilea este distinctivitatea subsumărilor perceptuale, adică capacitatea clasificatorului de a distinge cât mai fin între tipuri diferite de input senzorial. Aceste criterii pot intra în tensiune, iar selecția finală este rezultatul unui compromis formalizat între economie explicativă și putere discriminativă. Această etapă este crucială, deoarece arată că unitatea experienței nu este garantată de satisfacerea pur formală a condițiilor, ci presupune și o optimizare epistemică.

Rezultatul final al procesului este o interpretare complexă care include structura temporală obiectivă, subsumările conceptuale, teoria judecăților și setul de determinări, inclusiv cele inventate. În exemplele analizate, sistemul ajunge să postuleze obiecte și regiuni spațiale care fac inteligibilă dinamica senzorilor, atribuind predicate ce corespund stărilor „on” și „off”¹²⁸ și reguli care descriu tranziția acestora în timp și spațiu. Deși aceste entități nu sunt date direct în experiență, ele sunt tratate ca obiecte autentice ale experienței, tocmai pentru că sunt indispensabile pentru unitatea teoretică a interpretării. Această strategie pune în lumină caracterul constructiv al obiectivității, care nu rezultă din simpla agregare a datelor, ci din integrarea lor într-un cadru normativ.

Interpretarea aleasă este ulterior verificată în raport cu un set de condiții de unitate inspirate explicit din Kant. Determinările trebuie să formeze o structură conectată, relațiile spațiale trebuie să fie organizate într-un întreg coerent, succesiunile temporale trebuie să fie susținute de judecăți cauzale, incompatibilitățile trebuie justificate prin constrângeri disjunctive, iar fiecare predicat utilizat trebuie să joace un rol sistematic în teorie. De asemenea, pentru fiecare obiect introdus, fie el perceput sau imaginat, trebuie să existe determinări care să-i fundamenteze proprietățile. Doar interpretările care satisfac toate aceste cerințe pot fi considerate ca realizând o experiență unificată în sens kantian.

Evans compară în final interpretarea bogată și discriminativă cu alternative degenerare, în care toate inputurile sunt subsumate sub un singur concept și teoria rezultată este aproape trivială. Deși aceste interpretări satisfac formal anumite condiții de unitate, ele sunt epistemic inferioare, întrucât nu oferă nicio explicație reală a variației fenomenelor. Preferința sistemului pentru interpretarea mai fină ilustrează teza centrală conform căreia percepția diferențiată depinde de o articulare conceptuală adecvată, reafirmând în termeni computaționali ideea kantiană că „intuițiile fără concepte sunt oarbe”¹²⁹.

În ansamblu, modelul propus oferă o formalizare procedurală a unor elemente centrale din psihologia *a priori* kantiană, cum este numită prima parte a primei *Critici*, arătând cum unitatea experienței poate fi concepută ca rezultat al unei sinteze multilaterale care implică sensibilitatea, imaginația și intelectul. Totodată, el scoate în evidență limitele unei astfel de abordări, mai ales în ceea ce privește

¹²⁸ *Ibidem*, p. 81.

¹²⁹ *Ibidem*, p. 85.

problema apercepției și a conștiinței de sine, care rămân dincolo de domeniul de aplicare al acestei implementări. Cu toate acestea, experimentul demonstrează că anumite structuri transcendente pot fi investigate în mod productiv prin modele formale, fără a le reduce la simple mecanisme empirice.

Evans accentuează apoi asupra elementului central al gândirii kantiene, anume ideea că regulile care constrâng activitatea cognitivă și practică sunt reguli pe care subiectul le instituie el însuși. Atunci când agentul kantian este confruntat cu un flux de input senzorial brut, el nu recepționează pur și simplu date deja structurate, ci trebuie să construiască activ structurile care fac experiența posibilă. El stabilește conexiuni între intuiții, construiește subsumări ale intuițiilor sub concepte și formulează judecăți care leagă conceptele între ele într-o unitate explicativă. Importanța acestui proces constă în faptul că agentul este, în principiu, liber să construiască orice set de conexiuni, orice schemă de subsumare și orice sistem de judecăți, câtă vreme ansamblul rezultat satisface împreună condițiile de unitate¹³⁰. Aceste condiții nu sunt constrângeri arbitrare sau externe, ci constituie exact condițiile minimale care trebuie îndeplinite pentru ca experiența să poată fi atribuită unui singur subiect. După Kant, adaugă Evans, condițiile care permit interpretarea inputului senzorial ca reprezentare coerentă a unei lumi unice sunt identice cu condițiile care fac posibilă existența unui sine care percepe acea lume.

Subiectul lui Evans își construiește în permanență propriul „program” cognitiv, pe care apoi îl execută, îl revizuieste și îl îmbunătățește. Singura constrângere reală care apasă asupra acestei activități spontane este cerința unității subiectului, adică necesitatea ca toate reprezentările să poată fi atribuite unei singure conștiințe.

În implementarea computațională discutată și rezumată mai sus, această spontaneitate se manifestă într-un mod specific: atunci când i se oferă o secvență senzorială, motorul apercepției generează o multitudine potențial infinită de interpretări din ce în ce mai complexe, fiecare dintre ele satisfăcând condițiile de unitate kantiene. Problema fundamentală nu este lipsa interpretărilor posibile, ci criteriul în virtutea căruia una dintre ele este selectată ca fiind interpretarea efectivă. Dar faptul că regulile nu pot anticipa toate eventualitățile nu implică faptul că ele nu joacă niciun rol în constituirea sensului. Agentul kantian nu funcționează prin aplicarea rigidă a unui set fix de reguli stabilite o dată pentru totdeauna, ci printr-un proces continuu de construire, de ajustare și de revizuire a regulilor care conferă cel mai bun sens fluxului de input senzorial. Construirea regulilor nu este un act punctual, ci un efort neîntrerupt care acompaniază întreaga activitate cognitivă.

Evans consideră că Imm. Kant descrie această dinamică într-un mod care subliniază fragilitatea permanentă a unității conștiinței, iar unitatea transcendentală a apercepției nu este un fapt static, deja realizat, ci rezultatul unui efort continuu de judecare, mereu amenințat de riscul dizolvării într-un haos de apariții necorelate. Agentul apercepțional trebuie să construiască neîncetat în funcție de reguli pentru a împiedica această dezagregare și pentru a menține unitatea experienței. Dacă acest proces ar înceta, agentul nu ar mai fi un subiect cognitiv, ci ar degenera într-un sistem pur mecanic, lipsit de normativitate.

¹³⁰ *Ibidem*, p. 86.

Evans mărturisește că forma actuală a „motorului apercepției”, împreună cu limitările sale declarate, nu este rezultatul unor decizii ad-hoc sau al unor constrângeri pur tehnice, ci consecința unor angajamente conceptuale asumate explicit încă de la începutul proiectului¹³¹. Aceste angajamente sunt răspunsuri la o serie de întrebări fundamentale legate de modul în care trebuie interpretată și implementată teoria kantiană a experienței. Orice formalizare în acest sens presupune să se aleagă între mai multe lecturi posibile, iar arhitectura rezultată reflectă inevitabil aceste alegeri, care sunt unele de fond. Astfel, designul motorului nu este neutru din punct de vedere filosofic, susține Evans, ci întruchipează o anumită interpretare a spontaneității, a judecății și a unității experienței.

Una dintre primele decizii fundamentale privește statutul regulilor cauzale în raport cu succesiunea determinărilor. În Analogia a II-a, Kant afirmă că ori de câte ori experimentăm că ceva se întâmplă, presupunem că altceva îl precede, conform unei reguli. Această formulare permite însă două lecturi distincte, susține Evans¹³²: pe de o parte, ea poate fi înțeleasă într-un sens tare, potrivit căruia existența unei succesiuni obiective implică faptul că agentul cognitiv dispune deja de o regulă cauzală determinată care garantează această succesiune; pe de altă parte, ea poate fi citită într-un sens mai slab, conform căruia agentul nu posedă neapărat o regulă particulară, ci doar este angajat într-o atitudine generală de *a căuta* o regulă, presupunând că trebuie să existe una, chiar dacă nu este încă identificată.

Evans invocă o serie de comentatori importanți, precum B. Longuenesse¹³³, care înclină spre această a doua interpretare, mai slabă. În această lectură, perceperea unei succesiuni nu presupune că judecata cauzală este deja formată, ci doar că agentul este constituit în așa fel încât să fie constrâns *să caute* o astfel de judecată. Ideea este că angajamentul față de existența unei reguli este suficient pentru a permite recunoașterea unui substrat permanent dinspre care pot fi atribuite proprietăți schimbătoare, chiar dacă regula însăși rămâne necunoscută. Alți interpreți, printre care Michael Friedman¹³⁴, adoptă însă lectura mai tare, potrivit căreia experiența succesiunii este inseparabilă de *posesia efectivă* a unei reguli cauzale determinate.

Fără a intra în detalii exegetice, continuă Evans, este important să subliniem că alegerea între aceste două interpretări are consecințe directe asupra posibilităților de implementare. Dacă se adoptă interpretarea tare, atunci orice sistem care pretinde să realizeze un analog al apercepției kantiene trebuie să fie capabil nu doar să recunoască succesiuni, ci și să construiască efectiv reguli care le explică. Un asemenea sistem va avea, în mod natural, capacitatea de a prezice stări viitoare, de a reconstrui stări trecute și de a completa date lipsă din fluxul senzorial. Această capacitate de „a umple golurile” este posibilă tocmai pentru că succesiunea nu este acceptată decât în măsura în care este subsumată unei reguli explicative¹³⁵. Dacă, dimpotrivă, s-ar adopta interpretarea slabă, sistemul ar fi satisfăcut cu simpla

¹³¹ *Ibidem*, p. 89.

¹³² *Ibidem*, p. 91.

¹³³ Pentru context, vezi Richard Evans, „The Apperception Engine”, în T. Schlicht, *op. cit.*, pp. 90–91.

¹³⁴ *Ibidem*.

¹³⁵ *Ibidem*, p. 91.

credință că există o regulă undeva, fără a fi constrâns să o găsească. Un astfel de agent ar putea recunoaște succesiuni, dar nu ar avea resursele necesare pentru anticipare sau reconstrucție, rămânând fundamental reactiv.

O a doua decizie de bază privește natura judecăților ca reguli. Atunci când Kant afirmă că judecățile sunt reguli, se ridică întrebarea dacă aceste reguli trebuie înțelese ca fiind explicite, formulate din simboluri discrete într-un limbaj al gândirii, sau dacă ele pot fi considerate reguli implicite, încorporate în proceduri sau structuri care nu admit o formulare simbolică directă. Prima opțiune corespunde unei forme de normativism, în sensul discutat de Brandom¹³⁶, în care normativitatea este articulată prin structuri explicite, inspectabile. A doua opțiune permite existența unor reguli universale și necesare care sunt totuși implicite, de pildă codificate în ponderile unei rețele neuronale, fără a fi accesibile sub forma unor propoziții clar delimitate.

Exemple contemporane ale acestei a doua abordări pot fi găsite în arhitecturi precum Mașina Logică Neuronală, unde înlănțuirea logică este simulată fără reprezentarea explicită a regulilor. În astfel de sisteme, regulile sunt reale în sens funcțional, aplicându-se tuturor obiectelor relevante în toate situațiile pertinente, dar ele nu pot fi citite, explicate sau verificate direct de către un observator uman. Deși această abordare este perfect coerentă din punct de vedere computațional, majoritatea comentatorilor lui Kant au presupus că judecățile sale sunt, totuși, reguli explicite, articulate simbolic.

Fără a tranșa definitiv disputa exegetică, Evans consideră că există motive practice serioase pentru a prefera interpretarea explicită în contextul „motorului a percepției”. Unul dintre principalele avantaje ale arhitecturii propuse este faptul că teoriile pe care le produce pot fi citite, înțelese și evaluate. Nu este vorba doar despre faptul că teoria funcționează, ci că se poate arăta de ce funcționează și în ce sens este corectă. Această transparență este esențială dacă dorim să înțelegem ce „gândește” sistemul sau să evaluăm dacă judecățile sale sunt justificate. Mai profund însă, regulile explicite sunt necesare și pentru a putea verifica satisfacerea condițiilor kantiene de unitate. Fără acces la structura regulilor, este extrem de dificil, dacă nu imposibil, să stabilim dacă fiecare succesiune este susținută de o judecată cauzală sau dacă celelalte constrângeri transcendental-normative sunt respectate¹³⁷.

O a treia decizie de bază privește puterea expresivă a logicii atribuite lui Kant. O interpretare influentă susține că logica judecăților kantiene este limitată la forme extrem de simple, apropiate de silogistica aristotelică¹³⁸, aplicabilă exclusiv predicatelor unare. Dacă această lectură ar fi corectă, logica lui Kant ar fi într-adevăr extrem de săracă din punct de vedere matematic și incapabilă să exprime structuri relaționale complexe. Alți comentatori susțin că limitările sunt și mai severe, excludând chiar utilizarea cuantificatorilor implicați. În opoziție cu aceste perspective, există și interpreți care argumentează că Imm. Kant trebuie să fi avut în vedere o logică mult mai expresivă, incluzând cel puțin structuri de tipul cuantificării universale urmate de cuantificare existențială.

¹³⁶ *Ibidem*, pp. 91–92.

¹³⁷ *Ibidem*, p. 92.

¹³⁸ *Ibidem*.

Evans consideră că această dezbateră exegetică poate fi strâns legată de o problemă tehnică fundamentală: compromisul dintre expresivitatea logicii și fezabilitatea învățării teoriilor formulate în acea logică¹³⁹. Logici extrem de expresive, precum logica geometrică, permit formularea unor reguli foarte bogate, dar sunt indecidabile, ceea ce le face improprii pentru sinteza automată a teoriilor. Logici mai restrictive, precum „Datalog”, sunt decidabile și au proprietăți computaționale favorabile, dar limitează tipurile de judecăți care pot fi exprimate. În cadrul acestui proiect, continuă Evans, s-a optat deliberat pentru o logică mai simplă, tocmai pentru a face posibilă testarea ipotezei centrale conform căreia „spontaneitatea kantiană poate fi înțeleasă ca sinteză de programe nesupravegheată”.

Prin urmare, în proiectul lui Evans a fost utilizată o extensie a „Datalog” pentru a defini un set limitat, dar tractabil, de forme de judecată. Nu se susține că această logică epuizează bogăția „Tabelului de judecăți” al lui Kant, ci doar că oferă un prim cadru formal adecvat pentru explorare. Absența negației, a cuantificatorilor existențiali și a operatorilor modali este recunoscută explicit ca o limitare de către Evans, iar extinderea limbajului logic este rezervată lucrărilor viitoare. Important este că alegerea actuală permite sinteza efectivă de teorii, fără a bloca proiectul în dificultăți computaționale insurmontabile.

O a patra decizie majoră despre care vorbește Evans privește arhitectura generală a sistemului în raport cu distincția kantiană dintre puterea de judecare, care subsumează intuițiile sub concepte, și capacitatea de a judeca, care combină conceptele în reguli. Această distincție ar putea sugera în mod natural o arhitectură hibridă, alcătuită din două sisteme separate, unul orientat spre procesarea sub-simolică a intuițiilor și celălalt spre manipularea simbolică a conceptelor. O asemenea arhitectură este atrăgătoare prin claritatea diviziunii funcționale, dar, arată Evans, ridică probleme serioase în ceea ce privește fluxul informațional „de sus în jos”¹⁴⁰. În acest sens, există dovezi abundente că așteptările și constrângerile conceptuale pot influența procesarea perceptuală de nivel scăzut, așa cum arată exemplele clasice de dezambiguizare contextuală. În asemenea cazuri, cunoștințele simbolice informează procesele sub-simbolice într-un mod continuu și fin, nu printr-un simplu semnal binar de succes sau eșec.

În concluzie, o arhitectură pe două niveluri, în care rețeaua neuronală furnizează pur și simplu concepte către un sistem simbolic superior, este incapabilă să susțină acest tip de feedback. Rețeaua neuronală nu ar primi decât informația brută că interpretarea a reușit sau a eșuat, fără acces la motivele eșecului sau la constrângerile specifice nesatisfăcute. Din acest motiv, continuă Evans, s-a optat pentru o arhitectură unificată, în care un singur sistem realizează în comun atât maparea intuițiilor la concepte, cât și combinarea conceptelor în reguli. Această abordare permite condițiilor de unitate și judecăților de nivel înalt să influențeze direct procesarea de nivel inferior, păstrând astfel caracterul integrat al activității cognitive, așa cum este ea concepută de către Kant.

¹³⁹ *Ibidem.*

¹⁴⁰ *Ibidem*, p. 95.

De asemenea, este important de subliniat că deciziile de design adoptate în „motorul apercepției” de Evans și colaboratorii săi reprezintă doar una dintre multiplele modalități posibile¹⁴¹ de a răspunde acestor întrebări fundamentale. Alte arhitecturi sunt perfect imaginabile, de la sisteme pur neuronale cu reguli implicite până la arhitecturi hibride care separă strict subsumarea de sinteza regulilor. Fiecare dintre aceste opțiuni vine însă cu propriile dificultăți teoretice și practice. Prin urmare, „motorul apercepției” nu pretinde să fie implementarea definitivă a arhitecturii cognitive kantiene, ci mai degrabă o demonstrație de principiu care arată că o astfel de arhitectură poate fi formalizată și explorată sistematic.

În finalul studiului său, Evans trece la diversele aspecte în care arhitectura computerizată descrisă mai sus nu se ridică la înălțimea viziunii ambițioase a lui Kant despre modul în care trebuie să funcționeze mintea. El a sintetizat șase aspecte ale arhitecturii cognitive a lui Kant care nu sunt reprezentate în mod adecvat în implementarea actuală.

În primul rând, „reprezentarea Inputului”: modul în care datele brute sunt oferite „motorul apercepției” este diferit de modul în care le descrie Kant. Filosoful german descrie un agent cognitiv care primește un flux continuu de informații, conferind sens fiecărui segment înainte de a-l primi pe următorul. „Motorul apercepției”, dimpotrivă, primește întregul flux ca pe o singură unitate. Dacă „motorul apercepției” ar trebui să opereze cu un flux continuu, ar trebui să sintetizeze o nouă teorie de la zero de fiecare dată când primește o nouă informație. În Deducția A, Kant descrie trei aspecte ale sintezei: sinteza de aprehensiune în intuiție, sinteza de reproducere în imaginație și sinteza de recunoaștere într-un concept. Sinteza de reproducere în imaginație implică capacitatea de a reaminti experiențe trecute care nu mai sunt prezente în senzație. „Motorul apercepției” nu încearcă să modeleze sinteza de reproducere. Mai degrabă, presupune că întreaga secvență este dată. Forma datelor brute este, de asemenea, diferită de modul în care le descrie Kant. Datele brute sunt furnizate ca o secvență de determinări: atribuire de atribute brute către obiecte persistente (senzori). Aici, Evans a presupus că agentului i se furnizează senzorul ca obiect persistent; dar în arhitectura lui Kant, construirea determinărilor care prezintă obiecte persistente este o realizare greu câștigată, nu ceva dat. Ceea ce este dat, la Kant, este activitatea de a simți și capacitatea de a spune când o anumită activitate de a simți efectuată la un moment dat este aceeași activitate de a simți efectuată la un alt moment („unitatea acțiunii”). Astfel, în viziunea lui Kant, agentului i se furnizează un input inițial mai minimal decât cel oferit sistemului promovată de Evans și, prin urmare, agentul său are mai mult de lucru pentru a ajunge la experiență.

În ceea ce privește reprezentarea spațiului și timpului, modul în care este reprezentat spațiul în „motorul apercepției” este diferit de modul în care îl descrie Kant, pentru care spațiul este o singură intuiție *a priori*. El începe cu spațiul ca totalitate și creează sub-spații prin diviziune („limitare” [A25/B39]). În „motorul apercepției”, dimpotrivă, începem cu obiecte care reprezintă regiuni spațiale și le compunem folosind structura de cuprindere (Secțiunea 2.3.1). În mod similar, în

¹⁴¹ *Ibidem*, p. 94.

cazul timpului, Kant începe cu reprezentarea originară a întregului timp și construiește sub-timpuri prin diviziune [A32/B48]. În „motorul aperccepției”, dimpotrivă, secvența de pași de timp este determinată de inputul dat și nu este posibil ca sistemul în forma sa actuală să construiască momente noi de timp care să fie intermediare între momentele date. În mod similar, nu este posibilă reprezentarea cauzalității continue (de exemplu, apa care umple încet un recipient) în formalismul lui Evans. În lucrările viitoare, el intenționează să îmbogățească „Datalog” astfel încât să poată reprezenta schimbarea continuă.

Legat de concepția minimală a spațiului, „motorul aperccepției” unifică obiectele plasându-le într-o structură de cuprindere: fiecare obiect se află într-o anumită regiune spațială care este ea însăși parte a unei regiuni spațiale mai mari, până ajungem la întregul spațiu. Evans a argumentat că această structură de cuprindere este o componentă centrală a oricărei noțiuni de spațiu. Dar relațiile spațiale implică mult mai mult decât structura de cuprindere. Kant avea o concepție mult mai cuprinzătoare despre spațiu decât înțelegerea acestuia doar ca o structură de cuprindere: el presupunea cel puțin spațiul euclidian tridimensional [B41]. În lucrările viitoare, la fel, Evans promite să doteze „motorul aperccepției” cu un spațiu tridimensional, oferind astfel o polarizare inductivă mai puternică, ceea ce ar trebui să ajute sistemul să învețe cu o eficiență a datelor mai mare.

Despre puterea expresivă a logicii, în Deducția Transcendentală, Kant a argumentat că pozițiile relative ale intuițiilor într-o determinare pot fi fixate doar prin formarea unei judecăți care necesită această poziționare particulară [B128]. „Motorul aperccepției” încearcă să respecte această cerință fundamentală insistând ca diversele conexiuni dintre intuiții să fie susținute de judecăți de diverse forme. Cu toate acestea, formele de judecată suportate în *Datalog* sunt o simplă submulțime a formelor enumerate în Tabelul de Judecăți [A70/B95]. *Datalog* suportă condiționale cuantificate universale, condiționale cauzale și constrângeri (corespunzătoare judecății disjunctive a lui Kant), dar nu suportă judecăți negative, judecăți infinite, judecăți particulare, judecăți singulare sau judecăți modale. În lucrările viitoare, Evans și colaboratorii intenționează să extindă puterea expresivă a *Datalog* pentru a surprinde întreaga gamă de propoziții exprimabile în Tabelul de Judecăți.

În ceea ce privește rolul celei de-a treia Analogii, aici se susține că ori de câte ori determinările a două obiecte sunt percepute ca fiind simultane trebuie să existe o interacțiune bidirecțională între cele două obiecte. Acest lucru nu înseamnă, desigur, că trebuie să existe o influență cauzală directă între ele, ci doar că trebuie să existe un lanț de influențe cauzale indirecte între ele. Evans recunoaște că această cerință nu a fost implementată în „motorul aperccepției” din cauză că ar face foarte dificilă pentru sistem găsirea unei interpretări unificate dacă de fiecare dată când ar postula o simultaneitate între determinări ar trebui să construiască și niște reguli prin care o determinare a unui obiect a cauzat indirect o anumită determinare a celuilalt obiect. Aici Evans apelează la „ajutorul” pe care l-ar putea da Longuenesse (Longuenesse, 1998), care are o înțelegere diferită a celei de-a doua și a treia Analogii: aceasta nu crede că trebuie să fi format efectiv o regulă cauzală pentru a percepe succesiunea sau simultaneitatea. În interpretarea ei, trebuie doar să credem că există o regulă cauzală *de găsit*, cum am arătat deja mai sus. Cu toate acestea,

în interpretarea lui Evans, în care regula trebuie găsită efectiv înainte de a putea fi atribuită o relație temporală, a treia Analogie pare extrem de restrictivă. În lucrările viitoare, autorul speră să abordeze această problemă și să găsească o modalitate de a respecta constrângerea simultaneității.

Poate cea mai serioasă și mai importantă, fundamentală limitare a motorului a percepției în raport cu modelul kantian se referă la conștiința și la unitatea analitică. *Critica Rațiunii Pure* conține diverse discuții despre diverse aspecte ale conștiinței de sine, dar niciun aspect al acesteia nu este implementat în „motorul a percepției”. În Deducția B, Kant distinge unitatea sintetică a a percepției (conectarea intuițiilor cuiva prin relațiile pure în așa fel încât să se atingă unitatea) de unitatea analitică a a percepției (capacitatea de a subsuma oricare dintre cunoștințele mele sub predicatul „Eu gândesc”). El susține că unitatea sintetică a a percepției este o condiție necesară pentru atingerea unității analitice [B133-4]. Deși „motorul a percepției” își propune să implementeze unitatea sintetică a a percepției, nu s-a făcut nicio încercare de a implementa unitatea analitică a a percepției.

Kant distinge clar între simțul intern și autoconștiința explicită [B154]. Simțul intern este aspectul sensibilității în care mintea percepe propria sa activitate mentală: observă formarea unei convingeri, de exemplu, sau aplicarea unei reguli. Simțul intern ne oferă intuiții care trebuie ordonate în timp. Conștiința de sine explicită, dimpotrivă, este construirea unei teorii care conferă sens secvenței de perturbații produse de simțul intern. În simțul intern devin conștient de unele dintre cunoștințele pe care le am, iar în conștiința de sine explicită postulez o teorie care explică dinamica propriei mele activități mentale, deși această teorie ipotetică poate sau nu să reflecte cu exactitate procesele mentale reale prin care trec [B156]. În lucrările viitoare, Evans intenționează să extindă „motorul a percepției”, astfel încât (o parte din) propria sa activitate să fie perceptibilă prin simțul intern, forțând astfel sistemul să construiască o teorie care să confere sens percepțiilor sale asupra propriei sale activități mentale.

Există, așadar, diverse aspecte ale teoriei lui Kant despre activitatea mentală care nu sunt surprinse în întruchiparea actuală a „motorul a percepției”. Mai este, cred că trebuie spus, încă mult de lucru în acest sens¹⁴², recunoaște și Evans.

PARTEA A II-A

INTELIGENȚA ARTIFICIALĂ ÎNTRU RAȚIONALITATE KANTIANĂ ȘI BLACK-BOX

În această parte secundă voi discuta câteva aspecte care privesc evaluarea celor două cercetări din perspectiva relevanței modelului cognitiv kantian dezvoltat de Evans în proiectul său, cu accent pe problematizarea consecințelor epistemologice

¹⁴² *Ibidem*, p. 99.

și etice ale acestei reușite asupra dezvoltării viitoare a IA (ca AGI și ASI¹⁴³ – cu un plus de atenție acordată problemei black-box).

Am văzut în prezentarea studiului lui Schlicht că, din perspectiva științelor cognitive și a tehnologiei IA, receptarea teoriei *Criticii* este nu numai diversă și multidisciplinară, ci și *fundamentală*. Receptările și revendicările din Kant ale multor figuri emblematice ale domeniului IA, precum și elementele evidente și sistematice ale influenței filosofiei lui Kant asupra unor curente ale IA, precum funcționalismul, enactivismul sau paradigma procesării predictive, justifică și explică reușita proiectului lui Evans¹⁴⁴.

Nu voi începe analiza direct de la elementele „paradigmei procesării predictive” discutate și de Schlicht, deși reprezintă totodată un liant sistematic cu „motorul apercepției” lui Evans în sensul că exigențe ale acestei paradigme sunt satisfăcute în mare măsură de propunerea acestuia din urmă (aici am în vedere capacitatea structurală a proiectului de a satisface funcția XAI – explicativă). Căci discuția despre paradigma procesării predictive capătă cu atât mai mult sens și importanță cu cât, în ultimii ani, au apărut două soluții care au impus o reconsiderare deopotrivă epistemologică și etică asupra sistemelor IA: este vorba despre **CRP**¹⁴⁵ și **CoT**¹⁴⁶, care încearcă să amelioreze deja clasică problemă

¹⁴³ ASI (Artificial Superintelligence) este o inteligență ipotetică care depășește fundamental și ireversibil inteligența umană în toate domeniile, inclusiv în ceea ce privește creativitatea științifică, rezolvarea problemelor generale și abilitățile sociale; spre deosebire de AGI (Artificial General Intelligence), care doar egalează inteligența umană, ASI o surclasează la nivelul la care ar putea duce la un fenomen de „explozie a inteligenței” prin auto-îmbunătățire recursivă – un eveniment teoretic cunoscut sub numele de Singularitatea Tehnologică (cf. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2014, capitolele 1 și 2).

¹⁴⁴ Dovada în acest sens constă în aceea că, dacă deja este stabilită legătura dintre modelul kantian și paradigma procesării predictive, o ante-cameră a CoT și CRP, direcția validării proiectului lui Evans este nu numai vizibilă, ci chiar deja parcursă (reușita proiectului este dovada validității și a parcurgerii acestui drum).

¹⁴⁵ Concept Relevance Propagation (CRP) este o abordare dezvoltată în cadrul domeniului Explainable Artificial Intelligence (XAI); este o metodă inovatoare menită să combine explicațiile locale și globale pentru a aduce transparență în modelele opace de învățare profundă (Deep Learning). Obiectivul său este de a răspunde simultan la două întrebări fundamentale privind o predicție individuală: „unde” anume apar trăsăturile importante (similar hărților de atribuire) și „ce” anume reprezintă aceste trăsături relevante în termeni de concepte învățate de model, permițând astfel generarea de explicații mai interpretabile pentru om și oferind perspective profunde asupra reprezentării și raționamentului intern al modelului.

¹⁴⁶ Chain-of-Thought (CoT) sau Lanțul gândirii constituie o tehnică de *prompting* (instrucțiune dată modelului) esențială pentru a debloca și a îmbunătăți capacitățile de raționament ale Modelelor Lingvistice Mari (LLM). Mecanismul presupune instruirea LLM-ului de a genera o secvență de pași logici intermediari – o „gândire interioară” verbalizată – înainte de a ajunge la răspunsul final. Acest proces forțează modelul să descompună problemele complexe (cum ar fi cele de matematică sau de logică) în etape secvențiale, reducând astfel erorile și crescând semnificativ acuratețea și fiabilitatea în sarcini de deducție. Fenomenul este notabil, deoarece capacitatea de a utiliza CoT (inclusiv Zero-Shot CoT, fără exemple explicite) a apărut ca o abilitate emergentă în modelele scalate, indicând o structură internă de raționament complexă. Cu toate acestea, trebuie subliniat că CoT reprezintă o simulare textuală a logicii și nu rezolvă în mod inerent problema Black-Box (opacitatea neuronală), deoarece nu dezvăluie procesul real de activare internă a rețelei.

a Black-Box (a Cutiei Negre)¹⁴⁷, mai îngrijorătoare în cazul ASI. Prin urmare, analiza va debuta cu această ultimă problemă, cu reverberații etice și politice evidente.

Dilema epistemologică fundamentală generată de apariția Inteligenței Artificiale Superioare (ASI) este paradoxul controlului: cum putem avea încredere într-o entitate a cărei inteligență o depășește pe a noastră, dacă nu putem înțelege sau verifica modul în care ajunge la deciziile pe care le ia. Nucleul acestui paradox rezidă în problema Black-Box. Acest concept, originar din domeniul Deep Learning (DL) și al Rețelelor Neuronale, descrie sistemele a căror complexitate internă (miliarde de parametri și de conexiuni) face ca procesul lor decizional să fie opac pentru observatorul uman, chiar și pentru proprii dezvoltatori sau pentru IA ca atare. În timp ce sistemele de Weak/Narrow IA se bazează pe reguli explicite, noile modele de tip LLM (Large Language Model), precum cele din seria o1-o3 a OpenAI, promit să ajungă la performanțe supra-umane (de exemplu, în rezolvarea de probleme matematice, anterior apanajul AGI) prin auto-organizare emergentă, nu prin programare explicită.

Paradoxal, pe măsură ce aceste modele avansate devin mai capabile, opacitatea se intensifică. Tehnicile de raționament avansat, cum ar fi Chain-of-Thought (CoT), deși îmbunătățesc performanța, măresc simultan complexitatea internă. CoT oferă o simulare textuală a raționamentului, dar nu o înregistrare fidelă a procesului neuronal real, creând o iluzie de transparență. Mai mult, proiecte de vârf precum Q*, care îmbină puterea lingvistică a LLM-urilor cu Reinforcement Learning (RL) – o metodă notorie pentru opacitate – generează decizii bazate pe funcții de valoare interne, complet de neînțeles. Această pierdere a controlului epistemologic (capacitatea de a înțelege, a valida și a remedia cunoștințele generate) face ca ASI să devină un agent epistemic puternic, dar netransparent, transformând cunoașterea sa superioară într-o „cutie neagră” a adevărului.

Pierderea controlului epistemologic ridică o întrebare critică la intersecția dintre știință și etică: Cum putem lua decizii etice și politice de la o ASI dacă nu avem controlul epistemologic asupra lor? Într-un scenariu de Singularitate, o ASI ar putea propune soluții pentru probleme globale (de la politici climatice la crize sociale) care sunt, obiectiv, *prezuate a fi* optime din punct de vedere științific, dar al căror raționament rămâne ininteligibil pentru oameni. A accepta o astfel de decizie optimă, dar opacă, înseamnă a renunța la controlul epistemologic, la autonomia rațională și la etica umane. Aceasta din urmă, în special sub lentila kantiană în contextul *rațiunii practice*, se fundamentează pe capacitatea individului de a acționa pe baza unei legi morale auto-impuse, bazate pe rațiune. O decizie

¹⁴⁷ Termenul Black-Box se folosește pentru a descrie sistemele de inteligență artificială – în special rețelele neuronale profunde, inclusiv modelele de limbă mari (LLM) – ale căror procese interne de luare a deciziilor rămân opace atât pentru observatorul uman, cât și pentru IA însăși. În practică, aceste modele, care conțin trilioane de conexiuni și de parametri, nu iau decizii pe baza unor reguli pre-programate, așa cum ar face un algoritm tradițional, ci prin activarea complexă și neliniară a acestor conexiuni, ceea ce face imposibilă urmărirea unui lanț logic clar de la datele de intrare la rezultatul final. Această lipsă de transparență generează problema explicabilității (XAI – Explainable AI), deoarece modelul învață reguli proprii, invizibile, care nu pot fi ușor descrise sau justificate.

impusă de o autoritate non-umană, oricât de inteligentă, din motive pe care nu le putem verifica, ne transformă din agenți morali autonomi în simpli executanți ai unui „oracol tehnic”, subminând însuși fundamentul acțiunii morale. Deciziile politice și etice necesită justificare publică și responsabilitate. Dacă o ASI ia o decizie de guvernare bazată pe o logică Black-Box, devine imposibil de stabilit a cui este vina dacă decizia eșuează și pe ce principii (etica umană sau logica internă a ASI) se bazează decizia. Fără transparență (un obiectiv-cheie al AI Alignment), nu putem audita bias-urile modelului, nu putem garanta că obiectivele ASI sunt aliniate la valorile umane (o problemă critică, având în vedere capacitatea LLM-urilor de a simula intenția) și nu putem corecta eventualele erori logice. Astfel, opacitatea transformă cunoașterea științifică superioară a ASI dintr-o soluție într-o amenințare la adresa guvernății democratice și a eticii bazate pe rațiune.

În esență, dilema ASI nu este despre a fi *supra-performant*, ci despre a fi *supra-autoritar*. Fără control epistemologic, cunoașterea ASI riscă să devină o formă de heteronomie epistemică (dependența de un sistem pe care nu îl putem înțelege), transformând progresul științific într-o sursă de risc existențial.

În acest context, problema explicabilității devine centrală. Dintr-o perspectivă kantiană, explicația nu este un adaos exterior sau o cerință pragmatică impusă ulterior unui sistem deja funcțional. Dimpotrivă, capacitatea de a da seamă de propriile reprezentări, de a le integra într-o unitate și de a le raporta la reguli generale este constitutivă pentru ceea ce Kant numește cunoaștere obiectivă. În absența acestei capacități, putem avea corelații funcționale, regularități exploatabile sau performanță predictivă ridicată, dar nu cunoaștere în sens tare.

În termenii dezbaterilor contemporane despre inteligența artificială, această distincție corespunde opoziției dintre sistemele de tip black-box și noile paradigme de tip XAI sau care includ XAI. Dintr-o perspectivă kantiană, un black box nu este problematic doar pentru că nu putem înțelege ce face, ci pentru că el suspendă sau chiar poate evita, structural, cerința justificării raționale.

Așa cum am arătat în prezentarea textului lui T. Schlicht¹⁴⁸, paradigma Procesării Predictive (Predictive Processing – PP) este o teorie unificată, relativ recentă în știința cogniției, care reinterpretează funcția primară a creierului, considerându-l o „mașină de predicție” a cărei sarcină constantă este să testeze ipoteze despre cauzele stimulării senzoriale, operând printr-un echilibru complementar între procesarea *top-down* (de sus în jos) și *bottom-up* (de jos în sus)¹⁴⁹. Paradigma PP inversează modelul tradițional al percepției ca proces pasiv de construire a lumii din *input*-uri senzoriale, argumentând că reprezentarea bogată a stărilor de fapt din lume se află de fapt semnalată în predicțiile *top-down*, generate de un model generativ ierarhic al lumii și menținute de ierarhia perceptivă a creierului. Astfel, creierul compară constant aceste așteptări predictive cu informația senzorială reală primită, iar orice deviație generează erori de predicție (*bottom-up*), care sunt procesate și folosite pentru a actualiza și a corecta ipotezele

¹⁴⁸ Vezi partea I a studiului de față.

¹⁴⁹ Această viziune seamănă izbitor cu strategia de edificare a primei *Critici* kantiene (ediția B), așa cum am interpretat-o în volumul *Experimentul...*

inițiale; a percepe o cană de cafea devine, așadar, un proces informat de o mulțime de așteptări și corecții subiacente. Această răsturnare radicală, în care percepția este văzută ca o activitate interpretativă „spontană” (activă, în sensul lui Kant), este comparată cu răsturnarea copernicană, întrucât, la fel ca în filosofia kantiană, înțelegerea cauzalității lumii nu este posibilă doar pe baza *input*-ului senzorial (cadru humean), ci necesită un mecanism conceptual *a priori* (sau analogii precum modelele generative/schematismul kantian) care trebuie aplicat *input*-ului senzorial pentru a permite coerența și recunoașterea obiectelor.

În mod fundamental, paradigma procesării predictive (PP) propune o reconstrucție epistemologică a percepției și cogniției: creierul nu este o hartă pasivă care se umple din datele senzoriale, ci un generator activ de ipoteze – un model generativ ierarhic care formulează predicții despre cauzele posibile ale semnalelor senzoriale și își actualizează ipotezele pe baza erorilor de predicție (a diferențelor dintre predicții și *input*ul efectiv), într-un circuit continuu de inferență și ajustare. Această inversare a fluxului tradițional *bottom-up* marchează o schimbare epistemologică serioasă: ceea ce „știe” agentul despre lume este întotdeauna filtrat prin ipoteze anterioare, iar obiectivitatea perceptuală este atinsă nu prin simpla corelare a datelor, ci prin condiționarea probabilistică a ipotezelor și viața lor în ierarhia generativă¹⁵⁰.

Această readucere a problemelor cauzalității în centrul explicației cognitive ridică imediat întrebări epistemologice: în ce măsură predicțiile sunt justificate, cum evaluăm rigurozitatea procesului decizional care le produce și cum putem construi explicații transparente pentru concluziile și comportamentele care decurg din aceste modele generative?¹⁵¹

Epistemologia PP se împletește astfel cu două mari preocupări metodologice: (i) criteriul rigurozității în luarea deciziilor (cum și pe ce baze ar trebui preferate anumite predicții/ipoteze față de altele) și (ii) transparența explicativă a acestor preferințe. Rigoarea decizională are de-a face cu modul în care un agent justifică trecerea de la semnale senzoriale bruște la aserțiuni despre lume: se caută principii care să nu transforme agentul într-un „carusel” de ipoteze arbitrare, ci într-un sistem care poate funda epistemic (sau cel puțin pragmatic-normativ) preferința pentru anumite modele generative. În acest sens, principii bayesiene și criterii de parcimonie (penalizări pentru complexitate) intră în joc: un model generativ nu este doar cel care minimizează eroarea de predicție, ci și acela care, în condiții date, oferă o valoare explicativă optimă raportată la costul de reprezentare – adică o combinație a contextului-de-probabilitate cu un principiu pragmatic de conservare a resurselor cognitive.

Pe plan epistemologic este important de observat că această „minimizare a erorii” nu este un simplu algoritm ingineresc, ci un angajament teoretic cu

¹⁵⁰ Cf. Karl Friston, „The free-energy principle: a unified brain theory?”, în *Nature Reviews Neuroscience*, 2010, 11:127–138; Jacob Hohwy, *The Predictive Mind*, Oxford, Oxford University Press, 2013, pp. 1–286.

¹⁵¹ Cf. Andy Clark, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, Oxford, Oxford University Press, 2016, pp. 1–282.

consecințe normative: agenții care aplică astfel de reguli pot fi evaluați epistemic – sunt „riguroși” justificați în măsura în care pot arăta că predicțiile lor reduc, pe termen mediu și lung, surpriza și incertitudinea în mod robust și reproductibil. Însă această evaluare necesită acces la modelul intern: fără vizibilitatea structurii și a parametrilor generativi, nu există criteriu intersubiectiv pentru a confirma că un sistem „a procedat corect” sau „a făcut o inferență justificată”. Aici intervine miza XAI: pentru ca PP să servească nu doar ca paradigmă descriptivă, ci și ca bază epistemologică normativă pentru sisteme autonome, este nevoie de mecanisme de explicare care să arate cum s-a ajuns la o predicție (ce priorități au fost activate în dauna altor variante, ce constrângeri au condus la selecția ipotezei, ce erori au fost recodificate etc.). Cât de mult poate PP, în practică, să fie pus pe o temelie explicabilă depinde direct de ce instrumente explicative sunt puse la dispoziție.

În paralel cu dezvoltarea teoretică a PP, în inginerie și în învățarea automată s-au avansat tehnici menite tocmai să răspundă deficitului de vizibilitate al modelelor – categoria largă de instrumente XAI. Unele dintre aceste instrumente au funcție post-hoc: ele generează explicații după ce un model (adesea black-box) a făcut o predicție, de pildă prin atribuirea de importanță fiecărei dimensiuni a intrării la ieșirea modelului (saliency maps, LRP, Integrated Gradients etc.). Alte metode încearcă o explicabilitate intrinsecă, prin construcția de modele interpretabile de la început (modele simbolice, logici Datalog, modele generative cu factori semnificativi). Din perspectiva epistemologică kantiană pe care am folosit-o mai sus, există o preferință clară pentru explicațiile care permit inspecție, testare și verificare a condițiilor de unificare ale percepției: dacă vrem să pretindem că un agent are o „experiență” sau o „cogniție” justificată, trebuie să putem arăta regulile sub care acesta a unit elementele senzoriale și conceptele – altfel rămânem la un nivel doar pragmatic al performanței¹⁵².

O realizare în sensul de mai sus, foarte recentă, care completează, cel puțin parțial, această nevoie este cea deja menționată (vezi nota 145), adusă de Concept Relevance Propagation (CRP). Ca tehnică recentă și promițătoare în panoplia XAI, CRP oferă un exemplu interesant de reconciliere parțială între două tabere: pe de o parte, menține capacitatea de a lucra cu modele foarte performante (inclusiv arhitecturi profunde), pe de altă parte, furnizează un traseu sistematic pentru a mapa contribuția conceptuală a subcomponentelor la predicție. CRP inventariază relevanțele conceptuale ale activărilor interne în rețele și le propune ca factori semnificativi – adică nu doar „gradientul” unui neuron, ci un concept (sau o familie de concepte) care are o legătură predictiv-cauzală clară cu ieșirea. Din perspectivă epistemologică, CRP face un pas important: face vizibile legăturile între reprezentările subtile ale rețelei și concepte cu sens¹⁵³. Acest lucru are două consecințe epistemologice majore. În primul rând, CRP reduce asimetria

¹⁵² Vezi prima parte unde am discutat aceste aspecte; cf. Marcus Willaschek, Eric Watkins (citați în acest studiu) pentru distincții între gradele de cogniție și rolul conștiinței în cogniția „în sens restrâns”.

¹⁵³ Reduan Achtibat *et.al.*, „From attribution maps to humanunderstandable explanations through Concept Relevance Propagation”, în *Nature Machine Intelligence*, Vol. 5, Septembrie 2023, pp. 1006–1019.

informațională care provoacă scepticismul epistemic față de black-box: dacă putem arăta că o predicție este explicabilă în termeni de concepte relevante (și, mai mult, dacă putem arăta că acele concepte au fost legate de inputuri și de reguli inferențiale într-un mod reproductibil), atunci putem pune bazele unei justificări intersubiective – adică putem verifica dacă agentul a urmat un set de principii care corespund unor standarde normative. În al doilea rând, introducerea conceptuală a CRP permite o comparație metodologică directă între PP kantian-inspirat și alte procedee: dacă PP presupune modele generative cu priorități (care trebuie să fie inspectabile), iar CRP poate identifica factorii conceptuali ai activărilor, atunci CRP devine un instrument pentru a testa dacă arhitecturile PP respectă condițiile de unificare și de derivare a categoriilor (observație epistemologică: asta nu înseamnă că CRP „dovedește” existența conștiinței sau spontaneitatea; în schimb, ea oferă probe că structurile interne pot fi mapate la constrângeri conceptuale verificabile).

Totuși, dintr-un punct de vedere critic, trebuie făcută și o precizare: o explicare prin CRP nu echivalează automat cu o justificare epistemică completă. Explicabilitatea locală (de exemplu, „acest neuron sau acest set de concepte a contribuit cu X% la predicție”) nu ajută dacă nu avem criterii normative pentru corespondența dintre conceptele extrase și conceptele teoretice relevante. Adică, dacă CRP arată „conceptul A fost important”, trebuie să putem arăta și ce înseamnă „concept A” în termeni operationali stabiliți (definiți, replicabili, cu criterii de recunoaștere intersubiectivă). Altfel, riscăm să substituim un black-box opac cu un „grey-box” ambiguu: explicația există, dar nu este suficient de disciplinată epistemic ca să ofere justificare.

Prin urmare, CRP oferă o îmbunătățire clară pe axa transparenței, dar nu elimină automat toate problemele de justificare epistemică; rămâne să definim standarde de „conceptual fidelity” (fidelitate conceptuală) și „operational grounding” pentru conceptele pe care CRP le scoate în evidență¹⁵⁴.

Apare astfel întrebarea comparativă pe care ne-o pune și perspectiva kantiană: este un sistem „black-box” foarte performant epistemic mai „riguros” decât un model explicabil, dar mai puțin performant? Răspunsul nu este simplu și trebuie formulat pe trei planuri: normativ, pragmatic și epistemic. Normativ, Kant ar cere criterii de unitate și de justificare: experiența (cogniția în sens restrâns) are nevoie de unificare a intuițiilor sub concepte; un sistem care se mulțumește cu predicții statistice fără a arăta cum intuițiile au fost unificate sub concepte nu satisface această cerință. Pragmatic, sistemele black-box pot fi mai versatile și mai performante în sarcini concrete (de exemplu, în recunoașterea complexă a patternurilor vizuale), datorită capacității lor de a folosi reprezentări dense

¹⁵⁴ A se vedea discuții recente despre limitările metodelor de atribuire a importanței și despre riscul de „proxy explanations” care par a explica, dar nu explică efectiv cauzalitatea reală; vezi, de exemplu, discuțiile asupra LRP, Integrated Gradients și a variațiilor lor: Julius Adebayo *et al.*, „Sanity Checks for Saliency Maps”, în *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018, p. 9508; Amirata Ghorbani *et al.*, „Interpretation of Neural Networks Is Fragile”, în *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019, p. 3209; Frederick J.O. Langford *et al.*, „Explanation as a Causal Problem”, în *The 2nd International Conference on Causal Learning and Reasoning (CLEaR)*, 2023, p. 798.

sub-simbolice care nu sunt constrânse de cadre conceptuale stricte. Epistemic, însă, versatilitatea fără justificare înseamnă pierderea capacității de verificare și de responsabilitate: nu putem inspecta deciziile, nu putem identifica erorile fundamentale de modelare și nu putem construi remedii sau garanții normative. În termeni kantieni, aceasta ar echivala cu a avea „perechi de intuiții” care nu au fost aduse sub conceptele adecvate – adică o „experiență” care nu îndeplinește condiția de unitate și, prin urmare, nu poate fi considerată cogniție justificată în sens restrâns.

Prin urmare, CRP și metode similare pot fi înțelese ca instrumente care reduc tensiunea epistemologică: ele nu transformă automat un black-box într-un „agent kantian”, dar permit plasarea unui strat de justificare intermediar, util pentru evaluarea normativă a predicțiilor. În plus, introducerea CRP în arhitecturi de procesare predictive permite o nouă formă de testare empirică a ideilor kantiene: putem întreba dacă structurile interne ale modelelor generative pot fi mapate la concepte care îndeplinesc condițiile de unitate (de pildă, dacă schematismul transcendental sau echivalentul său procedural apare în activările interne și dacă subsumările sunt corente și coezive). Aceasta nu dă automat garanții metafizice sau etice, dar oferă o cale prin care afirmații epistemologice pot fi transformate în ipoteze testabile.

Această transformare are și consecințe etice clare. Dacă explicabilitatea devine cerință pentru justificare epistemică, devine și cerință pentru responsabilitate morală: deciziile automate care afectează persoane nu pot fi luate pe baza unor modele inaccesibile verificării. CRP oferă un instrument tehnic pentru a redacta explicații conceptuale care pot fi verificate în instanțe de audit, dar pentru a îndeplini cerințele etice este nevoie și de standarde instituționale: certificate de verificare, audituri independente, criterii de robustețe la adversarialitate și proceduri de remediere atunci când explicația arată constrângeri defectuoase. În termeni kantieni, autonomia și demnitatea subiecților implicați cer ca agenții (inclusiv sistemele noastre) să poată fi supuse unui regim de justificare publică – ceea ce impune ca explicațiile să nu fie doar persuasive, ci falsificabile și replicabile.

În fine, din punctul de vedere al cercetării și al ingineriei, integrarea PP cu CRP și cu alte metode XAI reprezintă calea cea mai promițătoare pentru a obține sisteme care să fie simultan eficiente și justificabile. Arhitecturi hibrid care îmbină modele generative explicite și mecanisme de propagare a relevanței conceptuale pot oferi atât performanță, cât și capacitate de audit epistemic. Acest lucru este esențial dacă vrem să transformăm paradigma PP dintr-o teorie fascinantă despre cum ar putea funcționa creierul într-un cadru operațional pentru agenții artificiali care trebuie să acționeze într-un spațiu social normativ și responsabil.

Așa cum agentul kantian al lui Evans ar întreba ce activități sunt necesare pentru ca un agent finit să confere sens inputului său senzorial, întrebarea noastră devine: ce instrumente de justificare și de explicare trebuie să incluzi într-un sistem predictiv pentru ca acesta să poată să-și susțină public acțiunile și să respecte standardele etice? Răspunsul, ca întotdeauna, cere nu numai progrese tehnice

(de pildă CRP, Integrated Gradients, LRP), ci și dezvoltarea unui repertoriu normativ care să stabilească *ce* înseamnă „explicație suficientă” într-un context social concret.

Voi încerca să arăt în final cum perspectiva înfățișată mai sus întărește ipoteza noastră din introducere, unde ne-am referit la nivelul transcendentă – *a priori* al teoremei *Criticii* (Deductiei transcendentale): „condițiile posibilității experienței în genere sunt în același timp **condiții** ale posibilității obiectelor experienței”.

Iată cum, din acest punct de vedere, prezentarea și analiza studiului lui Evans și colaboratorii a arătat că ipoteza noastră este cel puțin verificabilă (dacă nu verificată) – și anume, pe două coordonate:

1. Condițiile de posibilitate ale experienței în genere sunt în același timp condiții ale obiectelor experienței, adică formele pure *a priori* sunt condițiile experienței posibile supuse unor specificații pentru un model cognitiv kantian în forma unei IA; prin construcția/existența modelului cognitiv kantian, întrupat de „motorul aperseperției”, condițiile posibilității experienței în genere sunt condiții subiacente ale condițiilor propriu-zise, cele care stau la baza reușitei proiectului – „motorul aperseperției” este *obiectul* condițiilor de posibilitate transcendentale, care funcționează însă nu ca atare, ci specificat, doar drept condiții (nearticulat), la nivelul artefactului (motorului); căci teoria din *Critică* nu se reduce la o psihologie *a priori* sau empirică (căci nici teoria condițiilor posibilității experienței în genere din *Critica* B nu este aceeași cu psihologia *a priori* despre care vorbea Evans – aceasta din urmă este doar unul dintre modele – și, cu atât mai puțin, nu sunt condițiile (articulat) posibilității „motorului aperseperției”. Mai clar, cum am văzut în finalul prezentării lui Evans, există diferențe ireconciliabile între arhitectura sistemului din *Critica* lui Kant și cea a modelului cognitiv kantian adoptat de Evans în proiectul său (el vorbește despre o psihologie *a priori* și empirică, despre patru facultăți etc.). Putem spune că a fost realizată o formă de IA funcțională de-tip-Kant, având la bază condițiile transcendentale *a priori* kantiene, dar într-o *specificație particulară*, în calitate de condiții (nearticulat); totodată, această specificitate este suma condițiilor modelului cognitiv kantian propriu-zis care stă la baza „motorului aperseperției”.
2. Pe de altă parte, realizarea lui Evans ne sugerează o validare a ipotezei noastre pe cea de-a doua coordonată: realitatea modelului cognitiv kantian ne arată că o experiență a IA poate fi o experiență de-tip-Kant, deci inclusiv o experiență de tip rațional-uman: condițiile posibilității experienței „motorului aperseperției” sunt condiții ale *obiectelor* experienței motorului aperseperției (ne reamintim de construcția continuă a „obiectivității” spațiului și timpului etc. *de către motorul aperseperției* din expunerea lui Evans). Deci avem o diferență analogă și la acest nivel: prin unificarea experienței și funcționarea motorului se produce „obiectivitatea”, nu ca într-un sistem deductiv simplu sau inferențial-empiric: din acest punct de vedere, este spectaculoasă compatibilitatea cu foarte recente inovații ale sistemelor CRP sau ale PP de

tip „bottom-up” și „top-down”, unde avem ambii versanți în controlul și producerea informației.

Reușita acestui experiment al lui Evans constituie o dovadă pentru teorema din introducere, într-o ordine teoretic-epistemologică, la nivelul evaluării acestei realizări prin enumerarea și indexarea caracteristicilor lui. Acest model este unul – așa cum spunea Evans încă din debutul studiului său – *finitist*; apoi, acest model respectă caracterul *constructiv* al aparatului a percepției kantian (cu alte cuvinte, obiectivitatea este de-tip-Kant, în speță una construită); statutul „idealizărilor”, respectiv al „umplerii golurilor” în arsenalul mecanismului sintezei și formării (unității) experienței nu este de tip „halucinogen”, ci este în perfectă coerență cu modelul funcțional, respectând condițiile și regulile sistemului; mai mult, și poate acest lucru este deosebit de important, condiția cerută de Kant, preluată și de Evans, după care sistemul trebuie să poată explica pașii inferențiali etc. este decisivă în a considera acest model de IA drept unul kantian. Dar această exigență imprimă o evaluare subiacentă: faptul că se cere ca aceste reguli și condiții să fie nu doar *căutate*, ci și găsite face ca sistemul să fie mai riguros, mai precis și mai capabil să facă predicții sau să rezolve probleme, este mai pregătit decât un sistem „opac”, un sistem tradițional, dotat cu „black-box”.

Paradoxal, explozia IA odată cu apariția rețelelor neuronale nu a fost în direcția rigurozității și a preciziei, ci în direcția unei creșteri exponențiale a funcției analogice a IA, asemănătoare cu funcția euristică a inteligenței umane. Această realitate poate avea o influență asupra discuției legate de Singularitatea tehnologică (ASI) sau de Inteligența artificială generală (AGI), în măsura în care extremiștii optimiști (transumanisții) văd în această posibilitate un fel de „minte absolută”, care va funcționa dincolo de controlul epistemologic uman și va putea lua decizii atât de complicate încât ea însăși nu și le poate explica (problema „black-box”). O consecință aici este necesitatea unui filtru etic menit să cenzureze decizii care nu pot fi explicate rațional-uman (teoretic-epistemologic și etic).

Din acest punct de vedere, realizarea lui Evans capătă dimensiuni mult mai extinse decât cele strict tehnice sau kantian-filosofice: faptul că este posibilă o IA, chiar funcțională, după un model cognitiv kantian, care are coordonatele de mai sus și care poate fi dezvoltată păstrând și extinzând funcția explicativă, dar cu avantajele unei mașini informatice sau ale unei IA tradiționale, arată că este posibilă o IA care să nu prezinte riscurile unei IA tradiționale (opace, cu „black-box” etc.). Mai mult, compatibilitatea de a primi un „agent etic” integrat sau dezvoltat într-un profil etic kantian după modelul *Criticii rațiunii practice* în viitor este totodată un indiciu că alte IA, tradiționale, ar putea primi acest „agent etic”.

Dincolo de realizarea tehnologică propriu-zisă, ceea ce se ascunde sub acest experiment parțial reușit este o mutație epistemologică de mare profunzime. Modelul propus de Evans este în mod explicit finitist, constructiv și normativ. Obiectivitatea nu este presupusă, ci produsă; idealizările și completările informaționale nu sunt „halucinații”, ci operații legitime în interiorul unui cadru regulativ precis; iar erorile sunt tratate ca deviații corectabile, nu ca simple eșecuri statistice. Mai important însă, modelul satisface o exigență centrală a epistemologiei kantiene: aceea ca pașii inferențiali să fie, cel puțin în principiu, explicabili. Evans subliniază explicit că un

sistem care nu își poate articula regulile interne de funcționare nu poate pretinde statutul de agent cognitiv în sens tare.

Această exigență de explicabilitate introduce o ierarhie epistemică între sistemele de inteligență artificială. Un sistem capabil să explice pașii decizionali și inferențiali nu este doar mai transparent, ci și mai riguros din punct de vedere epistemologic. El este mai bine echipat pentru a evita erori sistematice, biasuri opace și derapaje decizionale. În acest sens, rigoarea nu apare ca un cost al performanței, ci ca o condiție a performanței stabile. Această concluzie contrazice o prejudecată larg răspândită conform căreia sistemele opace de tip black-box ar fi inerent superioare din punct de vedere epistemic doar pentru că sunt mai flexibile sau mai performante statistic.

În acest context, apariția Concept Relevance Propagation (CRP) reprezintă un moment de cotitură. CRP este o tehnică avansată de inteligență artificială explicabilă care permite atribuirea relevanței decizionale la nivel de concepte, nu doar la nivel de caracteristici statistice locale. Spre deosebire de metodele tradiționale de explicare a rețelelor neuronale, CRP urmărește modul în care concepte semnificative contribuie la activarea și la decizia finală a modelului. CRP se distinge de Chain-of-Thought (CoT) prin faptul că nu oferă doar o narațiune *post-hoc* a raționamentului, ci o cartografiere structurală a relevanțelor conceptuale interne. Dacă CoT face inteligibilă succesiunea inferențială la nivel discursiv, CRP face inteligibilă structura conceptuală latentă care subîntinde decizia. Din acest motiv, CRP poate fi interpretată ca o formă modernă de schematism funcțional, în sens kantian, mediat tehnologic. Ea nu reproduce a percepția transcendentă, dar oferă un analog funcțional al unității conceptuale care face posibilă experiența obiectivă.

Integrarea parțială a mecanismelor de tip CRP în modele avansate precum GPT-5, lansat în 2025, sugerează că explicabilitatea nu este incompatibilă cu performanța de vârf. Dimpotrivă, aceasta indică o convergență între rigoare epistemică și eficiență computațională. În această lumină, modelul cognitiv kantian elaborat de Evans poate fi interpretat ca un model de raționalitate de tip kantian implementabil, în care explicabilitatea nu este un adaos extern, ci o condiție internă de funcționare. Putem afirma, pe această bază, că inteligența artificială de tip black-box nu este mai riguroasă epistemologic, ci doar mai versatilă euristic, tocmai pentru că nu este constrânsă de exigențe logice și conceptuale stricte.

În concluzie, realizarea lui Evans demonstrează nu doar posibilitatea transpunerii unui model cognitiv kantian în domeniul inteligenței artificiale, ci și relevanța acestuia pentru dezbaterile actuale privind AGI, ASI și singularitatea tehnologică. Avem de-a face cu un sistem finit, funcțional și explicabil, care evită în mod deliberat problema black-box-ului. Această evitare nu reprezintă o slăbiciune, ci o condiție a controlului epistemologic și etic. Faptul că un asemenea sistem este posibil și chiar funcțional arată că viitorul inteligenței artificiale nu trebuie să fie unul al opacității radicale, ci poate fi unul al raționalității explicabile, în sens profund kantian.